Position: Open-Endedness is Essential for Artificial Superhuman Intelligence

Edward Hughes [*1] **Michael Dennis** [*1] **Jack Parker-Holder** [1] **Feryal Behbahani** [1] **Aditi Mavalankar** [1] **Yuge Shi** [1] **Tom Schaul** [1] **Tim Rocktäschel** [1]

Abstract

In recent years there has been a tremendous surge in the general capabilities of AI systems, mainly fuelled by training foundation models on internet-scale data. Nevertheless, the creation of open-ended, ever self-improving AI remains elusive. **In this position paper, we argue that the ingredients are now in place to achieve *open-endedness* in AI systems with respect to a human observer. Furthermore, we claim that such open-endedness is an essential property of any artificial superhuman intelligence (ASI).** We begin by providing a concrete formal definition of open-endedness through the lens of novelty and learnability. We then illustrate a path towards ASI via open-ended systems built on top of foundation models, capable of making novel, human-relevant discoveries. We conclude by examining the safety implications of generally-capable open-ended AI. We expect that open-ended foundation models will prove to be an increasingly fertile and safety-critical area of research in the near future.

1. Introduction

Recent years have seen impressive progress in AI, mainly driven by foundation models (Bommasani et al., 2021). These models are increasingly used as agents in various applications (e.g., Wang et al., 2023a; Wu et al., 2023; Lifshitz et al., 2023; Wang et al., 2023c; Liu et al., 2023b; Zheng et al., 2024; Ahn et al., 2022). This represents significant progress towards artificial general intelligence (AGI), in the sense of reaching human-level performance on a wide range of tasks (Legg and Hutter, 2007). However, we are still missing a formal description of what it would take for an autonomous system to self-improve towards increasingly creative and diverse discoveries *without end*—a Cambrian explosion of emergent capabilities, behaviors, and artifacts. This kind of *open-ended* invention is the mechanism by which human individuals and society at large accumulates new knowledge and technology. Therefore, open-endedness must be a property of an artificial superhuman intelligence (ASI, Morris et al., 2023) that can, by definition, accomplish a wide range of tasks at a level which no human can match. By the very nature of superhuman intelligence, open-ended discovery of innovative solutions is essential to empower humanity to manage its risks, just as society evolves norms and institutions to govern increasingly capable humans across generations (Richerson et al., 2001).

Foundation models such as large language models (LLMs) have scaled learning to large, static datasets scraped from the internet. Extrapolating, we may soon be running out of high-quality textual and visual data for training such models (Villalobos et al., 2022). Thus, open-endedness is unlikely to arise for free by training on ever-larger datasets. Rather, a system endowed with the open-endedness necessary for ASI will eventually have to create, refute and refine its own explanatory knowledge, in interaction with a source of evidence (Deutsch, 2011), as well as learning what data to learn from (Jiang et al., 2022). Moreover, for ASI to be useful and safe, it is important that open-endedness be guided towards knowledge that is understandable by and beneficial for humanity. Foundation models and open-endedness are orthogonal dimensions, whose combination is particularly powerful (cf. Lehman et al., 2022; Huang et al., 2022; Chen et al., 2023a; Meyerson et al., 2023; Zhang et al., 2023; Wu et al., 2023; Wang et al., 2023a). Open-ended algorithms endow foundation models with the ability to uncover new knowledge, while foundation models guide the search space for open-ended AI towards discovering human-relevant artifacts efficiently (Liu et al., 2023a; Ma et al., 2023; Romera-Paredes et al., 2024). A formal definition of open-endedness can catalyze progress in this direction, offering clarity and focus to galvanize the research community.

We provide a new and precise definition of open-endedness in Section 2, inspired by the open-ended systems in nature that have created life, the human brain, culture, and technology, as well as open-ended systems in silico that, for instance, have achieved superhuman level at the game of Go (Silver et al., 2016), generated human-level adaptation

[*]Equal contribution [1]Google DeepMind, London, UK. Correspondence to: Edward Hughes <edwardhughes@google.com>, Michael Dennis <dennismi@google.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

to novel 3D tasks (Bauer et al., 2023), self-improved language models (Fernando et al., 2023; Yang et al., 2023a), unlocked the tech tree in Minecraft (Wang et al., 2023a), and discovered new results in pure mathematics (Romera-Paredes et al., 2024). Open-endedness has been understood in a wide variety of ways (Earle et al., 2021) ever since it gained prominence as a term in the study of artificial life (Bedau, 1992; Bedau et al., 1998) and biological evolution (Holland, 1992; McShea, 1996; Waddington, 2008). Contrary to Stepney and Hickinbotham (2023), we believe quantifying open-endedness is both possible and important going forward, and, akin to Sigaud et al. (2023), we believe it can be achieved via the help of an observer external to the system. Our definition makes formal the aphorism of Lisa B. Soros that, as observers of an open-ended system, "we'll be surprised but we'll be surprised in a way that makes sense in retrospect". Concretely, open-ended systems produce increasingly novel and surprising artifacts that are hard to predict, even for an observer who has learned to better predict by examining past artifacts. Once a system exhibits these characteristics, i.e. producing learnable but novel artifacts, we call it an open-ended system. This allows us to pinpoint the sense in which open-endedness is essential for ASI, to provide examples illustrating how existing open-ended AI systems lack generality, and to argue that present-day foundation models are not yet open-ended.

Historically, the field of open-endedness has faced numerous challenges. Principal among these has been the problem of structuring the search space so as to regularly produce artifacts which are both novel and interesting to humans (Ma et al., 2023). When humans make discoveries, they do so by "standing on the shoulders of giant human datasets" (Clune, 2022); that is to say, utilising prior world, domain and commonsense knowledge, which they have acquired biologically or culturally. Since foundation models have been trained on vast amounts of human data, they capture human notions of interestingness (Zhang et al., 2023). Furthermore, they are general sequence modellers (Mirchandani et al., 2023) and can generate variations from existing examples (Meyerson et al., 2023), thus serving as general mutation operators. This is compelling since with more advanced foundation models, practical implementations of open-ended systems become increasingly feasible. Taken together, *open-ended foundation models* can both vary (i.e., mutate) data and assess novelty and interestingness of real and generated data to decide what data to further explore (i.e., select) (Jiang et al., 2022).

In Section 3 we provide some concrete research directions for this marriage between open-endedness and foundation models, for example leveraging evolutionary algorithms and reinforcement learning. Generally capable open-ended systems may be both extremely powerful and increasingly prevalent, prompting pressing safety considerations (Ecoffet et al., 2020). In Section 4, we argue that research into open-ended systems will be essential to safely and beneficially deploy any increasingly general and autonomous AI.

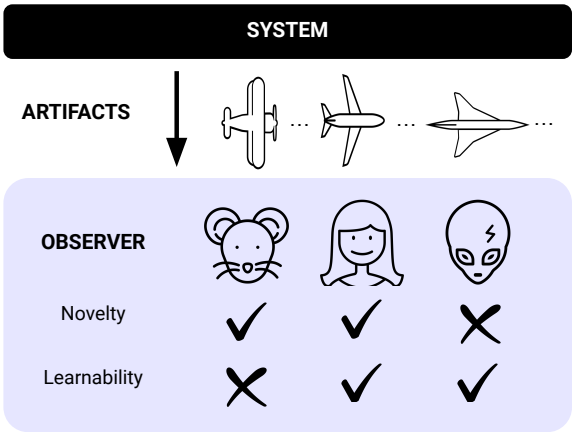

Figure 1. **Illustration of open-endedness definition.** The definition of open-endedness hinges on a system's ability to continuously generate artifacts that are both novel and learnable to an observer. Consider a system that designs various aircraft: a mouse (left) might find these designs novel but lack the capacity to comprehend the principles behind them; for a human studying aerospace engineering (middle), the system offers both novelty and the potential for learning, making it open-ended. However, a superintelligent alien (right) with vast aerospace knowledge might not find the design novel, but would still be able to analyze and understand them. This highlights that open-endedness is *observer-dependent* and that novelty or learnability alone is not enough.

2. Defining Open-Endedness

2.1. Formal Definition

The notion of an open-ended system has received many colloquial definitions (Soros and Stanley, 2014; Stanley and Lehman, 2015; Stanley et al., 2017; Stanley, 2019). More formal approaches have often focused on the case of evolutionary systems, quantifying the increasing complexity (McShea, 1996; Waddington, 2008) and perpetual novelty (Holland, 1992) of biological evolution. Intuitively, an open-ended system endlessly produces novel and interesting artifacts. But novelty and interestingness have generally been characterised without sufficient precision, or in an overly narrow way. We provide a general-purpose, formal definition of open-endedness, as follows.

Definition: From the perspective of an observer, a system is *open-ended* if and only if the sequence of artifacts it produces is both novel and learnable.

More formally, a **system** S produces a sequence of **artifacts** X_t, indexed by time t. An **observer** O processes a new artifact X_T to determine its predictability given a history $X_{1:t}$

of past ones. O possesses a **statistical model** $\hat{X}_t$ which predicts an arbitrary future artifact based on its observations of the artifacts it has seen up to time t. The observer judges the quality of their prediction based on a **loss metric** $\ell(\hat{X}_t, X_T)$, or $\ell(t, T)$ for short. A natural implementation of $\hat{X}_t$ is as a learning algorithm.

A system displays **novelty** if artifacts become increasingly unpredictable with respect to the observer's model at any fixed time t, namely:

$$\forall t, \forall T > t, \exists T' > T : \mathbb{E}\left[\ell(t, T')\right] > \mathbb{E}\left[\ell(t, T)\right] \ .$$

In other words, there is always a less predictable artifact coming further in the future.[1]

The system is **learnable** whenever conditioning on a longer history makes artifacts more predictable, namely:

$$\forall T, \forall t < T, \forall T > t' > t : \mathbb{E}\left[\ell(t', T)\right] < \mathbb{E}\left[\ell(t, T)\right] \ .$$

Finally, a system is **open-ended** from the perspective of the observer O if and only if it generates sequences of artifacts that are both novel and learnable (see Figure 1). The novelty aspect ensures the presence of information gain within the system, while learnability guarantees that this information gain holds meaning and is "interesting" to the observer.

For example, imagine that the system is a noisy TV producing uniform random noise (Burda et al., 2018). A noisy TV is learnable, allowing the observer to learn a statistical model that approximates the uniform distribution increasingly well; however, once the observer's model converges to uniform the system loses its novelty: all that is left is aleatoric uncertainty, which is collapsed by the expectation. Now imagine that the system is a noisy TV switched periodically by a remote control to a random, arbitrary distribution. Every time the channel is changed, the observer may experience novelty; however, the system is now not learnable, because the history of artifacts (previous TV channels) are not correlated with the distribution of the next channel, so the model loss will not decrease in general. We provide an informal positive example in Appendix A.1.

Our definition makes no explicit mention of "interestingness". More precisely, interestingness is represented in our definition by the observer's choice of loss function ℓ. Thus, for us, the interesting parts of artifacts are precisely those features which the observer decides are useful to learn about. Different observers can, and do, find different artifacts interesting, by virtue of the different parts of the feature space they choose to learn with their statistical model.

We hope that our definition will serve as a useful grounding for future work. On the theoretical side, it provides a basis for proving whether a system is open-ended. On a practical note, it raises the prospect of searching for open-ended systems. In this paper, we shall use it to underpin the argument that open-endedness lies on the critical path towards ASI, and in particular that the combination of open-ended algorithms and foundation models is ripe to yield significant progress towards that aim. We examine some subtleties of our definition in Appendix A.2.

2.2. Related Definitions

In the interests of space, we review the definitions of open-endedness most closely related to ours, covering more distantly related work in Appendix C. Soros and Stanley (2014) provided four necessary conditions for an evolutionary process to be open-ended, namely (1) that individuals must meet a minimal criterion in order to reproduce, (2) that evolution of individuals should create novel opportunities to meet the minimal criterion, (3) that individuals themselves should make decisions about how to interact with the world, and (4) that the potential complexity of the phenotype should not be limited by its representation. Our definition overlaps with these necessary conditions, but relaxes the constraint that the open-ended system is evolutionary. Our requirement that learnability is increasing can be seen as a generalisation of the minimal criterion in condition (1). Our requirement that the observer cannot intervene on the system is analogous to condition (3). Our requirement that novelty is increasing is analogous to conditions (2) and (4). Indeed, conditions (2) and (4) suggest that an open-ended system cannot be learned from a fixed data distribution.

To our knowledge, the most recent paper offering a definition of open-endedness is Sigaud et al. (2023). The authors write: "an observer considers a process as open-ended if, for any time t, there exists a time $t' > t$ at which the process generates a token that is new according to this observer's perspective". This definition has considerable overlap with ours. Like us, Sigaud et al. define open-endedness with respect to an observer. They consider the observer examining a sequence of tokens from a process, while we equivalently have the observer consider a sequence of artifacts from a system. Our requirement of novelty and learnability is compatible with their statement that the process should generate a token that is "new according to the observer's perspective". Our definition differs by being more precise about what this phrase means. In particular, we specify that what an observer considers "new" should be artifacts that are unpredictable according to their current statistical model of the system under consideration. Moreover, we specify that the observer's "perspective" is generated by learning that statistical model on the history of artifacts thus far presented by the system. In particular, our definition can rule out systems that display continual "novelty" but are otherwise uninteresting, like white noise on a TV screen, for instance.

[1] We take the expectation over any stochasticity in the artefacts; practically speaking, were the observer to make observations from identical copies of the system S, the expectation of ℓ would be approximated by the empirical mean.

2.3. Types of Observer

The choice of observer is a free parameter of great importance for our definition. From the perspective of AI research, there is a pre-eminent class of observers, namely humans. In other words, we wish to generate artifacts that are valuable to individual humans and to society. This provides a level of grounding for the open-ended system which narrows the search space considerably, as we shall argue in Section 3. Nevertheless, our definition deliberately admits arbitrary observers, for several reasons. Firstly, it allows our definition to encompass open-ended systems which are not anthropocentric, such as biological evolution. Secondly, it allows us to reason about open-ended systems which might exceed human capabilities, so-called ASI. Thirdly, it allows us to determine whether systems can be open-ended with respect to any observer, as we did with the noisy TV.[2]

Practically speaking, any given observer will have some *time horizon* τ which bounds their observations of a system, i.e. $t, T < \tau$. This concept allows us to distinguish between systems which are open-ended on different timescales. We say that a system is *infinitely* open-ended with respect to an observer O if it remains open-ended on any timescale $\tau \to \infty$. We say that a system is *finitely* open-ended with time horizon τ with respect to an observer O if it is open-ended for $t, T < \tau$. Consider, for example, an agent trained in simulation with an automatic curriculum over tasks. In principle, a human observer might find observations of the agent behaviour to be infinitely open-ended, for the agent may accrue the ability to solve ever more diverse and surprising tasks. In practice (cf. AdA, Bauer et al., 2023), novelty starts to plateau after about 1 month of training, due to limitations in the richness of the task space and in the size of the agent's neural network. Thus AdA is finitely open-ended with time horizon ≈ 1 month.

Similarly, an observer's judgement will be influenced by the limitations of their cognitive abilities relative to the breadth of the domain. For example, a human observer who reads a curriculum of ever more complex articles from a current snapshot of Wikipedia may find such a system open-ended, but only until they reach the limit of their memory. A suitable ordering of Wikipedia articles will present novel information, in the sense that every now and then an article will be more unpredictable than we have hitherto seen. We might also expect that this information will be learnable, because human knowledge is interlinked, in the sense that knowing more about one topic makes it easier to understand other topics that may crop up later. However, once human memory capacity is saturated, the human observer will start to forget previous articles. This violates learnability: in calculus, for instance, once one has forgotten the definition of a derivative, one will find it harder to understand an article about the chain rule. Therefore, conditioning on a history longer than an observer's recall doesn't necessarily make the current artifact more predictable.

This example brings to light three interesting threads. Firstly, the open-endedness of human technology, as observed by humans, relies on our ability to compress knowledge into a form that can be maintained within our collective memory: indeed, we present an alternative definition of open-endedness in the language of compression in Appendix B. Secondly, an artificial superhuman intelligence may have less stringent memory constraints than humans, and therefore may judge itself to be open-ended beyond the point at which humans assess it to be so, re-emphasising that human observers must be considered pre-eminent for the purposes of safety, as we explore further in Section 4. Thirdly, the open-endedness in this example is a function of the breadth of the domain. In a narrower domain, elliptic curve cryptography say, the set of relevant Wikipedia articles would be much smaller, so a human observer would find this open-ended only until they had understood every article, at which point novelty would be violated. Nevertheless, humans can, and frequently do, make new discoveries in narrow domains via experimentation and reasoning; amassing a vast, static trove of data is not the be all and end all of open-endedness.

2.4. Examples

In this section, we discuss some popular systems that are open-ended but not general, or that are general but not open-ended, with respect to a human observer. This serves two purposes. Firstly, it demonstrates that our definition is not so restrictive as to rule out systems that are intuitively open-ended, and is not so loose as to include systems that intuitively lack open-endedness. Secondly, it motivates the benefits that foundation models can provide in addressing the limitations of current open-ended systems and vice versa.

Our first archetypal open-ended system is *AlphaGo* (Silver et al., 2016). Consider as artifacts the sequence of policies produced across training by AlphaGo. After sufficient training, AlphaGo produces policies which are novel to human expert players, in the sense that they play moves which would be low probability for human professionals but which nevertheless are winning against the best humans. Furthermore, humans can improve their win rate against AlphaGo by learning from AlphaGo's behavior (Shin et al., 2023). Yet, AlphaGo keeps discovering new policies that can beat even a human who has learned from previous AlphaGo artifacts. Thus, so far as a human is concerned, AlphaGo is

[2]There is one constraint on an observer which must be adhered to for our definition to make sense. The loss function must treat artifacts X and predictions $\hat{X}$ on an equal footing. In particular it must be fixed in advance without any knowledge of the system S. Otherwise, an observer O could find a system S to be open-ended purely by discarding the artifacts from S and constructing its own artifacts that it finds to be both novel and learnable.

both novel and learnable. AlphaGo is just one representative from a class of open-ended algorithms that augment reinforcement learning with *self-play* (Samuel, 1959), achieving or exceeding human-level play in Go, Chess, Shogi (Silver et al., 2017), StarCraft II (Vinyals et al., 2019) Stratego (Perolat et al., 2022), DotA (Berner et al., 2019), and Diplomacy (Bakhtin et al., 2022).

AlphaGo is an example of an open-ended system that achieves narrow superhuman intelligence (Morris et al., 2023). This limits its utility: self-play of this kind cannot by itself help us to discover new science or technology that requires combining insight from disparate fields, or taking actions across a range of modalities, timescales and contexts. The constraints of the game rules make the search for novel and learnable artifacts tractable, and these artifacts are found to be novel and learnable by human observers largely because it was humans who invented the game.

Our second archetypal open-ended system is *AdA* (Bauer et al., 2023; OEL Team et al., 2021). AdA is a large-scale agent that learns to solve tasks in an 3D-environment called XLand2. In XLand2 there are 25B possible task variants, corresponding to different world topologies and a variety of possible games within each world, that are prioritized for learning potential (Jiang et al., 2021). Checkpoints of the AdA agent across training are open-ended with respect to a human observer who attempts to predict what capabilities the agent might show. Across training, the agent gradually accumulates zero-shot and few-shot capabilities over an ever wider set of held-out environments, requiring ever more complex skills. Thus the human continually observes novel capabilities in the agent. Furthermore, the prioritization of task variants provides an interpretable ordering to the accumulation of skills in the agent, rendering this learnable by a human. AdA represents a wider class of open-ended algorithms driven by *unsupervised environment design* (UED, Dennis et al., 2020; Justesen et al., 2018), which establish an *automatic curriculum* (Leibo et al., 2019; Baker et al., 2020) of environments in the zone of proximal development for agent learning (Vygotsky and Cole, 1978).

It is natural to ask whether AdA would continue to be judged as open-ended by a human observer should training be continued indefinitely. Results in Bauer et al. (2023) suggest that novelty starts to plateau, implying that with an order of magnitude more compute AdA would almost certainly not be open-ended. Indeed, the authors show that both increasing the size of the agent and increasing the number of tasks allow the agent to generalize to a wider range of environments. Thus, in order for this system to be open-ended on longer timescales, one would need an even richer environment and an even more capable agent to sustain the agent-environment co-evolution inherent in UED.

Our third archetypal open-ended system is *POET* (Wang et al., 2019; 2020). POET trains a population of agents, each of which is paired with an environment that is evolving over the course of training. These paired agent-environment artifacts are open-ended with respect to a human observer seeking to model the features of the environments that arise, or equivalently the skills the paired agents possess. A *Quality Diversity* algorithm (QD, Pugh et al., 2016; Mouret and Clune, 2015) is deployed with respect to the environments, hunting for challenging problems that lead to diverging performance across the population. QD is an example of a wider class of open-ended algorithms, namely evolutionary algorithms, which we encounter again in Section 3.4.

Crucially, POET periodically transfers agents from one environment to another, which results in an empirical example of the stepping stone phenomenon (Stanley and Lehman, 2015): agents can eventually solve incredibly challenging environments that are not possible to solve with direct optimization. As a result of training for billions of environment steps, POET produces a diverse population of highly capable specialist agents, which can solve novel environments that are created through coevolution with the population (Brant and Stanley, 2017). Novelty arises because of the mutation operator in the QD algorithm, which yields new and unpredictable environments. Learnability arises because each mutation is small, so the past lineage of an environment is a good guide to its current features. Just as for AdA, the key limitation on open-endedness is the environment parameterization itself: eventually POET will plateau once the agent can solve all possible terrains.

Our final example is contemporary *foundation models*. These are a negative example; they are not open-ended by our definition with respect to any observer who can model their training dataset. The justification for this follows immediately from our consideration of the noisy TV in Section 2.1. Contemporary foundation models are typically trained on fixed datasets. If the distribution of this data is learnable, which it must be, for the foundation model learned it in the first place, then it cannot be endlessly novel, because eventually the observer will have modelled the epistemic uncertainty. As we saw in Section 2.3, foundation models may appear open-ended to human observers if the domain of enquiry is sufficiently broad, by virtue of the memory limitations of the human brain. However, if the focus is narrowed, for instance to tasks that require planning (Momennejad et al., 2024; Pallagani et al., 2023; Valmeekam et al., 2023), the limitations of the foundation model in generating novel, correct solutions are exposed.

Since foundation models are periodically retrained on new data, including data generated by their own interactions with humans and the real world, one could argue that the data distribution is not really fixed. In some quarters, this kind of distributional shift is seen as an annoyance, even

one which threatens "model collapse" (Shumailov et al., 2023). We flip this argument on its head, and contend that augmenting foundation models with open-endedness offers a path towards ASI. Similarly, the fact that foundation models are typically conditional on context breaks the logic that they cannot be open-ended. In principle, the context of a foundation model can be recruited to recombine concepts in an open-ended way by leveraging some external measure of validity. This brings us neatly to some concrete suggestions for how to build open-ended foundation models.

3. Open-Ended Foundation Models

We have defined open-endedness and discussed why the current foundation model training paradigm is *not* open-ended. We believe that the trend of improving foundation models trained on passive data by scaling alone will soon plateau, and it will not be enough to reach ASI. Our position is that open-endedness is a property of any ASI, and that foundation models provide the missing ingredient required for domain-general open-endedness. Further, we believe that there may be only a few remaining steps required to achieve open-endedness with foundation models. In the following subsections, we sketch four overlapping paths towards open-ended foundation models that lend credence to this belief. The paths are neither intended to be prescriptive nor exhaustive. Indeed, recent publications such as (Wong et al., 2023b; Sharma et al., 2023) point to other paths.

Before proceeding, we must justify our claim that a future foundation model trained passively on some large corpus of human data is unlikely to spontaneously acquire open-endedness. In principle, should we reach ASI, there will be some sum total of data which the model has consumed during its training, possibly via several intermediate stages. Therefore, our claim is not about the impossibility of assembling such a dataset. Rather, we suggest that it is unlikely that this dataset can be pre-collected offline in an efficient way. The reason is that open-endedness is fundamentally an *experiential* process: producing novelty and learnability in the eyes of an observer requires continual online adaptation on the basis of the artifacts already produced, in the context of that observer's evolving prior beliefs.

What would it take to collect offline a static dataset from which such an experiential skill could be learned? Such a dataset must contain a treasure trove of artifacts which themselves crisply show novelty and learnability. Yet the process by which culture evolves, ideas develop, inventions arise and technologies proliferate is seldom recorded neatly and comprehensively. The alternative paradigm, in which experience is "built in" to the open-ended system, is well illustrated by the scientific method. Since the Enlightenment, the simple process of making hypotheses on the basis of current knowledge, falsifying them with experiments based on a source of evidence, and codifying the results into new knowledge has yielded unprecedented progress in science and technology (Deutsch, 2011). In our view, the fastest path to ASI will take inspiration from the scientific method, compiling a dataset online by the explicit combination of foundation models and open-ended algorithms.

3.1. Reinforcement Learning

The framework of Reinforcement Learning (RL) has been at the forefront of achieving superhuman performance in narrow domains, such as AlphaGo's groundbreaking strategies that have enriched the human understanding of the game of Go. RL agents act deliberately so as to shape their stream of experience for both accumulating reward (exploitation) and learning about how to increase expected reward in the future (exploration). A nuanced extension are agents that set their own goals to (learn to) pursue; and generating the sequence of these goals can itself be an open-ended process, which drives open-ended experience generation (Colas et al., 2022). Voyager (Wang et al., 2023a) provides an early example of how RL-like self-improvement can be built on top of foundation models, without the need for explicit parameter updates or established RL algorithms. Instead, Voyager assembles an LLM-powered curriculum, uses iterative prompting as an improvement operator, and assembles verified skills into a library for hierarchical reuse.

A key problem in RL is how to shape exploration towards novel and learnable behaviors in high-dimensional domains, as discussed in Jiang et al. (2022). Exploration can be guided, for instance, by pseudo-rewards (Bellemare et al., 2016; Burda et al., 2018; Du et al., 2023b), modulation (Schaul et al., 2019) or an automated curriculum that selects relevant tasks (Jiang et al., 2021; Parker-Holder et al., 2022; Samvelyan et al., 2023). To generalize this, a useful abstraction may be the notion of a *proxy observer*, which sits within the system and proactively guides it to generate novel and learnable content for the true external observer. In the past this guidance was provided on the basis of simple metrics such as TD-error, but now we can leverage foundation models to guide exploration towards artifacts that more closely align with what a human observer deems to be novel and interesting (Jiang et al., 2022). There is already evidence that this approach may be effective, with LLMs providing agent rewards from text in an environment (Klissarov et al., 2023) and compiling a curriculum of tasks based on their interestingness (Zhang et al., 2023; Faldor et al., 2024).

While RL considers the first-person perspective of an agent interacting with an environment, a different perspective centers on multi-agent dynamics, and the additional richness arising from all the ways that different (possibly heterogeneous) agents can interact with each other, adapt to each other, or learn from each other. The presence of multiple

learning agents provides a source of non-stationarity, such that the optimal strategy for each individual will change over time, potentially in an open-ended manner. Non-stationary dynamics been used to achieve or exceed human-level performance in games like StarCraft, DotA and Stratego. There is early evidence that multi-agent systems may help to improve factuality and reasoning in LLMs via debate (Du et al., 2023c; Tang et al., 2023), although there is much more research needed before superhuman capability is reached.

3.2. Self-Improvement

To achieve open-endedness, a model must not only consume knowledge from pre-collected feedback as in, for example, RLHF (Ziegler et al., 2019), but also generate new knowledge, in form of hypotheses, insights or creative outputs beyond the human curated training data. A self-improvement loop should allow the agent to actively engage in tasks that push the boundary of its knowledge and capabilities, for example via leveraging tools such as search engines, simulated environments, calculators or interpreters and interacting with other agents (Jiang et al., 2022; Schick et al., 2024). This requires the model to have a scalable mechanism to evaluate its own performance, identify areas for improvement, and adapt its learning process accordingly.

There is growing evidence that foundation models can be leveraged for feedback in place of humans, and can significantly amplify data generated by humans. Examples include self-critique and revision for training harmless assistants (Bai et al., 2022) and guiding human evaluators (Saunders et al., 2022), self-correction for tool-use (Gou et al., 2023), self-instruction for instruction following (Wang et al., 2022), self-debugging for code generation (Chen et al., 2023b), self-rewarding for instruction following (Yuan et al., 2024), and leveraging VLMs as reward functions for control (Baumli et al., 2023). These works hint at the possibility of foundation models generating their own samples and refining them in an open-ended way.

3.3. Task Generation

Closely related to both RL and self-improvement is the problem of task generation, also known as the "problem problem" (Leibo et al., 2019). One great candidate approach for open-endedness is to keep adapting the difficulty of tasks to an agent's capability so that they remain forever challenging yet learnable. Past examples of this type of system include setter-solvers (Schmidhuber, 1991b) and unsupervised environment design (Dennis et al., 2020; Justesen et al., 2018; Wang et al., 2019). With the advent of foundation models, it has become feasible to use the Internet itself as an environment (Jiang et al., 2022; Gur et al., 2021) via web-based APIs, affording agents with an incredibly rich, ever-growing and human-relevant task domain (Zhou et al., 2023).

Another possibility is to instead learn world models—predictive simulators that can generate future outputs conditioned on text or actions. A promising approach is to consider a foundation model to be a world model itself, since it is capable of predicting the future (Wong et al., 2023a; Gurnee and Tegmark, 2023; Park et al., 2023). Learned world models like Genie (Bruce et al., 2024), and text-to-video generation models like Sora (Brooks et al., 2024) demonstrate that foundation video models can be used as learned simulators, including in real-world settings like robotics (Yang et al., 2023b) and autonomous driving (Hu et al., 2023). If these works combine with learned multi-modal reward models (Chan et al., 2023; Du et al., 2023a), they could be used to generate an open-ended curriculum of tasks, scaling to task spaces far larger and more photorealistic than can currently be achieved. At sufficient scale, this may provide a path to generating AI agents with superhuman adaptability across a wide range of previously unseen tasks, which can be deployed in the real world across the rapidly closing Sim-to-Real gap (Huang et al., 2023).

3.4. Evolutionary Algorithms

Evolutionary methods offer a promising path to generate open-ended systems with foundation models (Wu et al., 2024). LLMs are well-placed to act as selection and mutation operators, as they have been trained on vast datasets of human knowledge, culture and preferences. For example, LLMs offer a mechanism through which to make semantically meaningful mutations via text (Lehman et al., 2022; Meyerson et al., 2023; Chen et al., 2023a). The simplest such approach may be via *prompts*, which already allow foundation models to further improve their performance. Recent works have shown it is possible to far surpass human designed prompts, leading to stronger models (Fernando et al., 2023; Yang et al., 2023a; Guo et al., 2023). More recently, Bradley et al. (2023) and Samvelyan et al. (2024) went further, using an evolutionary algorithm and LLMs to both generate variation and evaluate the quality and diversity of candidate text, making it possible to guide the search for creative and novel outputs. In the future it may be possible to further refine a model on these outputs, or use them for planning (Gandhi et al., 2023), to achieve self-improvement.

Another angle of attack for evolutionary methods is in the space of code (also known as genetic programming). Foundation models have proven to be competent at producing diverse and novel programs, providing a means of iterating upon an archive of candidate solutions. For example, Eureka (Ma et al., 2023) evolves code-based reward functions to learn complex control behaviors. Similarly, FunSearch (Romera-Paredes et al., 2024) evolves programs that represent new mathematical knowledge. These examples are focused on specific domains, and it remains an open problem to scale code evolution to a more general setting.

4. Achieving ASI Responsibly

Now that we have foundation models, designing a truly general open-ended learning system may be within our grasp. However, the power of open-endedness comes with a swathe of notable safety risks—beyond existing safety considerations facing foundation models (Ecoffet et al., 2020). Finding solutions to these challenges are interesting and important core problems in open-endedness research. Because the solutions to these problems may well depend on the design of the open-ended system, it is critical that safety and open-endedness are pursued in tandem. We cover them here not to hold them separate from other directions in open-endedness—in fact many of these problems are current practical limitations of artificial open-ended systems. Rather, this section is intended to draw specific attention to these problems as some of the most fundamental and exciting directions for research in the field. Of course, this short section cannot do justice to the breadth of concerns. Hence, where possible, we provide references to the wealth of knowledge in the ASI safety community.

We organize our understanding of these risks similar to (Critch and Krueger, 2020) by focusing on the ways knowledge is created and transmitted through the joint human-AI open-ended process in Figure 2. A powerful open-ended system which has the problems listed in this section is not a beneficial open-ended system, and we believe it is not one we should be striving to build. Solving these problems is not just making open-ended systems safer, but also making them usable by humans. As such, addressing these problems should be thought of as minimum specifications of an open-ended system that we would want to build.

4.1. AI Creation and Agency

AI systems powering the open-ended creation of new knowledge could lead to powerful new affordances. Without direction, these creations could be the source of dual-use dangers (Urbina et al., 2022). The danger is magnified when the open-ended systems take immediate action in an environment. Current state-of-the-art systems operate in narrow, simulated environments (Wang et al., 2023a; OEL Team et al., 2021; Bauer et al., 2023). However, as AI is trained in broader, more diverse simulations or is even deployed (and continues to learn) in the real world, it becomes critical to understand the dangers. The agency of open-ended AI poses several safety risks, such as goal misgeneralization (di Langosco et al., 2022; Shah et al., 2022) and specification gaming (Clark and Amodei, 2016). Open-ended search can be seen as an ambitiously aggressive form of exploration; thus one could hope to use similar approaches to mitigate the dangers of exploration as in RL, like safe exploration (Garcıa and Fernández, 2015) and impact regularization (Krakovna et al., 2018; Turner et al., 2020).

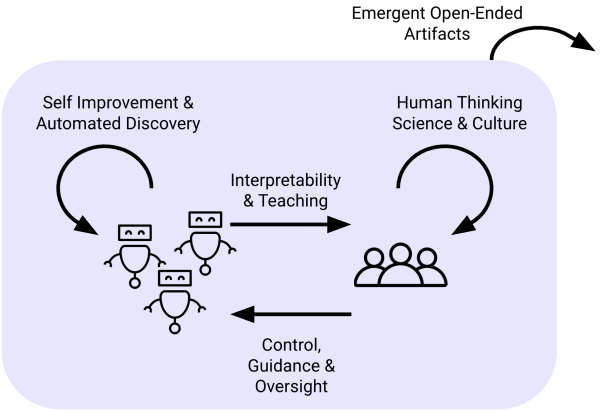

Figure 2. **Knowledge accumulation and transfer in a human-AI open-ended system.** We depict AI building on AI knowledge, humans understanding AI knowledge, AI understanding human knowledge, humans building on human knowledge, and emergent knowledge created by the process as a whole. Every process in this diagram offers an opportunity to embed safety methods that guide the system towards achieving ASI responsibly.

4.2. Humans Understanding AI Creations

In order to provide informed oversight and direction when guiding an open-ended system, human observers need to at least partially understand the significance of the new artifacts that the system produces. This becomes increasingly challenging as the complexity of these artifacts grows, leading to the inability to give informed oversight and guidance. Such a system may not only be unsafe, but would no longer be open-ended for human observers, since it would no longer be learnable. As such, any open-ended system we want to build should have the ability to bring human observers along with it—understanding and interpreting these systems is not only a core problem to make them safe, it is also a core problem to make them useful.

One approach would be to try to understand the policy generated by open-ended systems through interpretability. With current approaches this would require a formidable interpretability effort for each domain of interest. However, with the advent of automated interpretability (Bills et al., 2023), one may hope to build increasingly good explanations of the systems' behaviors which match the increasing complexity of the open-ended system. This presents an sizeable challenge, as such a system would be a universal explainer (Deutsch, 2011), by definition.

An alternative approach is to prefer designs for open-ended systems which promote interpretability and explainability, or whose goal is to teach human observers. Already, there are efforts to train systems which directly inform the user of implicit knowledge (Christiano et al., 2021). One might aim to design systems that at least maintain informed oversight (Amodei et al., 2016; Bowman et al., 2022). This approach

may be especially effective if the design of the open-ended system automatically facilitates understanding and control by human users (Irving et al., 2018).

4.3. Humans Guiding AI Creation

Even if we assume that human observers can understand enough of the behavior of an open-ended system to be in a position to give informed feedback, we arrive at the question of how a human designer could meaningfully guide an open-ended system. This challenge goes beyond the difficulties of directing individual RL agents, as not only do open-ended systems often lack well-defined objectives that could be modified, but they are increasingly unpredictable by design. One possibility would be to use humans in the loop to drive open-endedness (Secretan et al., 2008), a kind of open-endedness from human feedback (Zhang et al., 2023). A complete solution to this problem not only needs to be directable, but must actively raise unexpected and possibly important artifacts to the user's attention.

If open-ended systems could be made as directable as individual RL agents, then work defining objectives which preserve controllability (Hadfield-Menell et al., 2016; 2017; Carey and Everitt, 2023) might be a promising path towards more controllable open-ended systems. However, directing an open-ended system towards any objective effectively while maintaining the open-endedness is an open problem. This problem is not only important for safety, but is important for open-ended systems to be useful. In sufficiently broad domains—such as all of mathematics, all proteins, or all behaviors on a computer—an open-ended system may rabbit-hole into the obscure theorems, useless proteins, or only certain computer applications. Thus, building mechanisms that allow us to direct open-ended systems to not just the safe artifacts, but the interesting and useful artifacts, is a fruitful avenue for collaboration between safety and open-endedness researchers.

4.4. Human Society Adapting

There are significant non-technical concerns in ensuring that society can understand, prepare for, and appropriately react to new technological capabilities emerging from open-ended foundation models. Indeed, the impact of AI systems is not just felt at the individual level, but also at the level of the collectives that structure our society—communities, organisations, markets and nation states, to name a few. Since the artifacts arising from open-ended foundation models will by definition appear novel, we must devote prospective attention to the ways in which these could harm or benefit the cooperative infrastructure of society (Dafoe et al., 2020). Likewise, we must develop mechanisms to avoid tipping points driven by feedback loops, like flash crashes (Aldrich et al., 2017). Decision-makers should be prepared to adapt governance rapidly and retrospectively in response to open-ended artifacts, finding a good balance between collecting information and avoiding entrenchment of undesirable artifacts (Collingridge, 1980).

4.5. Emergent Risks of Open-Ended Systems

Even if each subcomponent of Figure 2 can be made safe, it may still be the case that the aggregate joint human-AI open-ended system leads to unforeseen problems. For instance, two systems that are open-ended in isolation could negatively interact to cause neither to be open-ended. This would mean a cessation of progress and an inability to collectively respond to new challenges. While such emergent effects have been studied in multi-agent systems (Johanson et al., 2022) and ASI safety (Critch and Krueger, 2020) solutions are still elusive, and an understanding of these effects is critical to the safe deployment of open-ended systems.

If such problems are inevitable and unpredictable, we would need our human-AI open-ended systems to adapt to solve novel ASI safety failures as they arise. Due to the inherent unpredictability of knowledge creation, these problems may be both unavoidable and solvable once as they arise (Deutsch, 2011). We should be building an open-ended system whose safety is anti-fragile (Taleb, 2014), adapting to emerging safety risks and getting stronger for it. This entails designing techniques for understanding, monitoring, and rapidly coordinating responses to emerging risks.

5. Conclusion and Outlook

Foundation models have led to a rapid increase in the generality of current AI systems. However, current foundation models are limited in their capability to discover new knowledge. In this paper, our position is that to further advance in levels of AGI towards ASI, we require systems that are *open-ended*—endowed with the ability to generate novel and learnable artifacts for a human observer. There has never been a more exciting time to build such systems, with foundation models already exhibiting general human-like knowledge that both accelerates further learning and guides this learning towards human-relevant artifacts.

As we develop and deploy more generally-capable open-ended systems, novel safety concerns arise that will be critical to address. In order to realise the benefits of such systems, it is important that the human observer remains able to learn from the novel artifacts, bringing fields such as explainability to the forefront of open-endedness research. If these endeavors are successful, then we believe open-ended foundation models could lead to advances that drastically enhance modern society.

Impact Statement

Our work provides a formal definition of open-endedness, and provides a discussion on its significance for the pursuit of ASI. We explore current research directions in the field, emphasising the potential of combining open-endedness with foundation models as a pre-eminent path towards achieving ASI. Developed responsibly, we believe that such open-ended foundation models can have tremendous positive impact on the society, accelerating scientific and technological breakthroughs, enhancing human creativity through a collaborative feedback loop, and acting as an engine for general knowledge expansion across many fields. Recognising the profound implications of this concept, we dedicate the entirety of Section 4 to an initial analysis of potential risks and societal impacts, offering frameworks for the responsible and ethical development of ASI. We hope that highlighting these issues early will help to promote safety, responsibility and accountability as the field grows.

Acknowledgements

We gratefully acknowledge Dave Abel for providing valuable feedback on an early draft of this paper. We are thankful to the designers at the Noun Project, from which we sourced graphics under the CC BY 3.0 licence as follows: "tick" icon by kareemovic, "Delete" icon by kareemovic, "alien" icon by Artem Yurov, "girl" icon by Teewara soontorn, "year of rat" icon by DailyPM, "aircraft" icon by mikicon, "concorde" icon by mikicon, "Plane" icon by CAMB, "humans" icon by Ifanicon, and "Robot" icon by Deemak Daksina.

References

D. Abel, A. Barreto, B. Van Roy, D. Precup, H. van Hasselt, and S. Singh. A definition of continual reinforcement learning. *ArXiv preprint*, abs/2307.11046, 2023. URL https://arxiv.org/abs/2307.11046.

M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, A. Herzog, D. Ho, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, E. Jang, R. J. Ruano, K. Jeffrey, S. Jesmonth, N. J. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, K.-H. Lee, S. Levine, Y. Lu, L. Luu, C. Parada, P. Pastor, J. Quiambao, K. Rao, J. Rettinghouse, D. Reyes, P. Sermanet, N. Sievers, C. Tan, A. Toshev, V. Vanhoucke, F. Xia, T. Xiao, P. Xu, S. Xu, M. Yan, and A. Zeng. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances, Aug. 2022.

E. M. Aldrich, J. Grundfest, and G. Laughlin. The flash crash: A new deconstruction. *Available at SSRN 2721922*, 2017.

D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete problems in ai safety. *ArXiv preprint*, abs/1606.06565, 2016. URL https://arxiv.org/abs/1606.06565.

Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. E. Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, and J. Kaplan. Constitutional AI: Harmlessness from AI Feedback, Dec. 2022.

B. Baker, I. Kanitscheider, T. M. Markov, Y. Wu, G. Powell, B. McGrew, and I. Mordatch. Emergent tool use from multi-agent autocurricula. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=SkxpxJBKwS.

A. Bakhtin, N. Brown, E. Dinan, G. Farina, C. Flaherty, D. Fried, A. Goff, J. Gray, H. Hu, et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022.

J. Bauer, K. Baumli, F. Behbahani, A. Bhoopchand, N. Bradley-Schmieg, M. Chang, N. Clay, A. Collister, V. Dasagi, L. Gonzalez, K. Gregor, E. Hughes, S. Kashem, M. Loks-Thompson, H. Openshaw, J. Parker-Holder, S. Pathak, N. Perez-Nieves, N. Rakicevic, T. Rocktäschel, Y. Schroecker, S. Singh, J. Sygnowski, K. Tuyls, S. York, A. Zacherl, and L. M. Zhang. Human-timescale adaptation in an open-ended task space. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 1887–1935. PMLR, 2023.

K. Baumli, S. Baveja, F. Behbahani, H. Chan, G. Comanici, S. Flennerhag, M. Gazeau, K. Holsheimer, D. Horgan, M. Laskin, et al. Vision-language models as a source of rewards. *ArXiv preprint*, abs/2312.09187, 2023. URL https://arxiv.org/abs/2312.09187.

M. Bedau. Measurement of evolutionary activity, teleology, and life. 1992.

M. A. Bedau, E. Snyder, C. T. Brown, N. H. Packard, et al. A comparison of evolutionary activity in artificial evolving

systems and in the biosphere. In *Proceedings of the fourth European conference on artificial life*, pages 125–134. MIt Press Cambridge, 1997.

M. A. Bedau, E. Snyder, and N. H. Packard. A classification of long-term evolutionary dynamics. *Artificial Life: The Proceedings...*, page 228, 1998.

M. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos. Unifying count-based exploration and intrinsic motivation. *Advances in neural information processing systems*, 29, 2016.

C. Berner, G. Brockman, B. Chan, V. Cheung, P. Debiak, C. Dennison, D. Farhi, Q. Fischer, S. Hashme, C. Hesse, et al. Dota 2 with large scale deep reinforcement learning. *ArXiv preprint*, abs/1912.06680, 2019. URL https://arxiv.org/abs/1912.06680.

S. Bills, N. Cammarata, D. Mossing, H. Tillman, L. Gao, G. Goh, I. Sutskever, J. Leike, J. Wu, and W. Saunders. Language models can explain neurons in language models. *URL https://openaipublic. blob. core. windows. net/neuron-explainer/paper/index. html.(Date accessed: 14.05. 2023)*, 2023.

R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

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

H. Bradley, A. Dai, H. Teufel, J. Zhang, K. Oostermeijer, M. Bellagente, J. Clune, K. Stanley, G. Schott, and J. Lehman. Quality-Diversity through AI Feedback, Oct. 2023.

J. C. Brant and K. O. Stanley. Minimal criterion coevolution: a new approach to open-ended search. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 67–74, 2017.

T. Brooks, B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing, D. Schnurr, J. Taylor, T. Luhman, E. Luhman, C. Ng, R. Wang, and A. Ramesh. Video generation models as world simulators. 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

J. Bruce, M. Dennis, A. Edwards, J. Parker-Holder, Y. Shi, E. Hughes, M. Lai, A. Mavalankar, R. Steigerwald, C. Apps, Y. Aytar, S. Bechtle, F. Behbahani, S. Chan, N. Heess, L. Gonzalez, S. Osindero, S. Ozair, S. Reed, J. Zhang, K. Zolna, J. Clune, N. de Freitas, S. Singh, and T. Rocktäschel. Genie: Generative Interactive Environments, Feb. 2024.

Y. Burda, H. Edwards, A. Storkey, and O. Klimov. Exploration by Random Network Distillation, Oct. 2018.

M. C. Campi and S. Garatti. Compression, generalization and learning. *ArXiv preprint*, abs/2301.12767, 2023. URL https://arxiv.org/abs/2301.12767.

R. Carey and T. Everitt. Human control: Definitions and algorithms. *ArXiv preprint*, abs/2305.19861, 2023. URL https://arxiv.org/abs/2305.19861.

H. Chan, V. Mnih, F. Behbahani, M. Laskin, L. Wang, F. Pardo, M. Gazeau, H. Sahni, D. Horgan, K. Baumli, Y. Schroecker, S. Spencer, R. Steigerwald, J. Quan, G. Comanici, S. Flennerhag, A. Neitz, L. M. Zhang, T. Schaul, S. Singh, C. Lyle, T. Rocktäschel, J. Parker-Holder, and K. Holsheimer. Vision-language models as a source of rewards. In *Second Agent Learning in Open-Endedness Workshop*, 2023.

A. Chen, D. M. Dohan, and D. R. So. EvoPrompting: Language Models for Code-Level Neural Architecture Search, Feb. 2023a.

X. Chen, M. Lin, N. Schärli, and D. Zhou. Teaching Large Language Models to Self-Debug, Apr. 2023b.

P. Christiano, A. Cotra, and M. Xu. Eliciting latent knowledge: How to tell if your eyes deceive you, 2021.

J. Clark and D. Amodei. Faulty reward functions in the wild. *Internet: https://blog. openai. com/faulty-reward-functions*, 2016.

J. Clune. AI-GAs: AI-generating algorithms, an alternate paradigm for producing general artificial intelligence, Jan. 2020.

J. Clune. Ai will go farterh if it stands on the shoulders of giant human data sets. Dec. 2022.

C. Colas, T. Karch, O. Sigaud, and P.-Y. Oudeyer. Autotelic agents with intrinsically motivated goal-conditioned reinforcement learning: a short survey. *Journal of Artificial Intelligence Research*, 74:1159–1199, 2022.

D. Collingridge. *The Social Control of Technology*. St. Martin's Press, 1980. ISBN 9780312731687. URL https://books.google.co.uk/books?id=hCSdAQAACAAJ.

A. Critch and D. Krueger. Ai research considerations for human existential safety (arches). *ArXiv preprint*, abs/2006.04948, 2020. URL https://arxiv.org/abs/2006.04948.

A. Dafoe, E. Hughes, Y. Bachrach, T. Collins, K. R. McKee, J. Z. Leibo, K. Larson, and T. Graepel. Open Problems in Cooperative AI, Dec. 2020.

O. David, S. Moran, and A. Yehudayoff. On statistical learning via the lens of compression. *ArXiv preprint*, abs/1610.03592, 2016. URL https://arxiv.org/abs/1610.03592.

G. Delétang, A. Ruoss, P.-A. Duquenne, E. Catt, T. Genewein, C. Mattern, J. Grau-Moya, L. K. Wenliang, M. Aitchison, L. Orseau, M. Hutter, and J. Veness. Language Modeling Is Compression, Sept. 2023.

M. Dennis, N. Jaques, E. Vinitsky, A. M. Bayen, S. Russell, A. Critch, and S. Levine. Emergent complexity and zero-shot transfer via unsupervised environment design. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

J. Derbyshire. Potential surprise theory as a theoretical foundation for scenario planning. *Technological Forecasting and Social Change*, 124:77–87, 2017.

D. Deutsch. *The beginning of infinity: Explanations that transform the world*. Penguin UK, 2011.

L. L. di Langosco, J. Koch, L. D. Sharkey, J. Pfau, and D. Krueger. Goal misgeneralization in deep reinforcement learning. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 12004–12019. PMLR, 2022. URL https://proceedings.mlr.press/v162/langosco22a.html.

E. L. Dolson, A. E. Vostinar, M. J. Wiser, and C. Ofria. The modes toolbox: Measurements of open-ended dynamics in evolving systems. *Artificial life*, 25(1):50–73, 2019.

Y. Du, K. Konyushkova, M. Denil, A. Raju, J. Landon, F. Hill, N. de Freitas, and S. Cabi. Vision-language models as success detectors. In *Proceedings of The 2nd Conference on Lifelong Learning Agents*, pages 120–136, 2023a.

Y. Du, E. Kosoy, A. Dayan, M. Rufova, P. Abbeel, and A. Gopnik. What can ai learn from human exploration? intrinsically-motivated humans and agents in open-world exploration. In *NeurIPS 2023 workshop: Information-Theoretic Principles in Cognitive Systems*, 2023b.

Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023c.

S. Earle, J. Togelius, and L. B. Soros. Video games as a testbed for open-ended phenomena. In *2021 IEEE Conference on Games (CoG)*, pages 1–9. IEEE, 2021.

A. Ecoffet, J. Clune, and J. Lehman. Open Questions in Creating Safe Open-ended AI: Tensions Between Control and Creativity, June 2020.

M. Faldor, J. Zhang, A. Cully, and J. Clune. Omni-epic: Open-endedness via models of human notions of interestingness with environments programmed in code. *arXiv preprint arXiv:2405.15568*, 2024.

C. Fernando, D. Banarse, H. Michalewski, S. Osindero, and T. Rocktäschel. Promptbreeder: Self-Referential Self-Improvement Via Prompt Evolution, Sept. 2023.

K. Gandhi, D. Sadigh, and N. D. Goodman. Strategic Reasoning with Language Models, May 2023.

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

Z. Gou, Z. Shao, Y. Gong, Y. Shen, Y. Yang, N. Duan, and W. Chen. Critic: Large language models can self-correct with tool-interactive critiquing. *ArXiv preprint*, abs/2305.11738, 2023. URL https://arxiv.org/abs/2305.11738.

Q. Guo, R. Wang, J. Guo, B. Li, K. Song, X. Tan, G. Liu, J. Bian, and Y. Yang. Connecting Large Language Models with Evolutionary Algorithms Yields Powerful Prompt Optimizers, Sept. 2023.

I. Gur, N. Jaques, Y. Miao, J. Choi, M. Tiwari, H. Lee, and A. Faust. Environment generation for zero-shot compositional reinforcement learning. *Advances in Neural Information Processing Systems*, 34:4157–4169, 2021.

W. Gurnee and M. Tegmark. Language Models Represent Space and Time, Oct. 2023.

D. Hadfield-Menell, S. J. Russell, P. Abbeel, and A. D. Dragan. Cooperative inverse reinforcement learning. In D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3909–3917, 2016.

D. Hadfield-Menell, A. D. Dragan, P. Abbeel, and S. J. Russell. The off-switch game. In C. Sierra, editor, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 220–227. ijcai.org, 2017. doi: 10.24963/ijcai.2017/32. URL https://doi.org/10.24963/ijcai.2017/32.

E. Hazan and S. Kale. Extracting certainty from uncertainty: Regret bounded by variation in costs. *Machine learning*, 80:165–188, 2010.

M. Henaff, R. Raileanu, M. Jiang, and T. Rocktäschel. Exploration via Elliptical Episodic Bonuses, Jan. 2023.

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

A. Hu, L. Russell, H. Yeo, Z. Murez, G. Fedoseev, A. Kendall, J. Shotton, and G. Corrado. GAIA-1: A Generative World Model for Autonomous Driving, Sept. 2023.

J. Huang, S. S. Gu, L. Hou, Y. Wu, X. Wang, H. Yu, and J. Han. Large Language Models Can Self-Improve, Oct. 2022.

P. Huang, X. Zhang, Z. Cao, S. Liu, M. Xu, W. Ding, J. Francis, B. Chen, and D. Zhao. What went wrong? closing the sim-to-real gap via differentiable causal discovery. In *Conference on Robot Learning*, pages 734–760. PMLR, 2023.

M. Hutter. *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. Springer Science & Business Media, 2004.

G. Irving, P. Christiano, and D. Amodei. AI safety via debate, Oct. 2018.

M. Jiang, E. Grefenstette, and T. Rocktäschel. Prioritized Level Replay, June 2021.

M. Jiang, T. Rocktäschel, and E. Grefenstette. General Intelligence Requires Rethinking Exploration, Nov. 2022.

M. B. Johanson, E. Hughes, F. Timbers, and J. Z. Leibo. Emergent bartering behaviour in multi-agent reinforcement learning. *ArXiv preprint*, abs/2205.06760, 2022. URL https://arxiv.org/abs/2205.06760.

N. Justesen, R. R. Torrado, P. Bontrager, A. Khalifa, J. Togelius, and S. Risi. Illuminating generalization in deep reinforcement learning through procedural level generation. *arXiv preprint arXiv:1806.10729*, 2018.

M. Klissarov, P. D'Oro, S. Sodhani, R. Raileanu, P.-L. Bacon, P. Vincent, A. Zhang, and M. Henaff. Motif: Intrinsic Motivation from Artificial Intelligence Feedback, Sept. 2023.

F. H. Knight. *Risk, uncertainty and profit*, volume 31. Houghton Mifflin, 1921.

V. Krakovna, L. Orseau, R. Kumar, M. Martic, and S. Legg. Penalizing side effects using stepwise relative reachability. *ArXiv preprint*, abs/1806.01186, 2018. URL https://arxiv.org/abs/1806.01186.

S. Legg and M. Hutter. Universal Intelligence: A Definition of Machine Intelligence, Dec. 2007.

J. Lehman and K. O. Stanley. Abandoning Objectives: Evolution Through the Search for Novelty Alone. *Evolutionary Computation*, 19(2):189–223, June 2011. ISSN 1063-6560. doi: 10.1162/EVCO_a_00025.

J. Lehman, J. Gordon, S. Jain, K. Ndousse, C. Yeh, and K. O. Stanley. Evolution through Large Models, June 2022.

J. Z. Leibo, E. Hughes, M. Lanctot, and T. Graepel. Autocurricula and the Emergence of Innovation from Social Interaction: A Manifesto for Multi-Agent Intelligence Research, Mar. 2019.

S. Lifshitz, K. Paster, H. Chan, J. Ba, and S. McIlraith. STEVE-1: A Generative Model for Text-to-Behavior in Minecraft, June 2023.

S. Liu, C. Chen, X. Qu, K. Tang, and Y.-S. Ong. Large language models as evolutionary optimizers. *arXiv preprint arXiv:2310.19046*, 2023a.

X. Liu, H. Yu, H. Zhang, Y. Xu, X. Lei, H. Lai, Y. Gu, H. Ding, K. Men, K. Yang, S. Zhang, X. Deng, A. Zeng, Z. Du, C. Zhang, S. Shen, T. Zhang, Y. Su, H. Sun, M. Huang, Y. Dong, and J. Tang. AgentBench: Evaluating LLMs as Agents, Aug. 2023b.

Y. J. Ma, W. Liang, G. Wang, D.-A. Huang, O. Bastani, D. Jayaraman, Y. Zhu, L. Fan, and A. Anandkumar. Eureka: Human-Level Reward Design via Coding Large Language Models, Oct. 2023.

A. N. Mavor-Parker, K. A. Young, C. Barry, and L. D. Griffin. How to stay curious while avoiding noisy tvs using aleatoric uncertainty estimation. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 15220–15240. PMLR, 2022. URL https://proceedings.mlr.press/v162/mavor-parker22a.html.

D. W. McShea. Perspective metazoan complexity and evolution: is there a trend? *Evolution*, 50(2):477–492, 1996.

E. Meyerson, M. J. Nelson, H. Bradley, A. Moradi, A. K. Hoover, and J. Lehman. Language Model Crossover: Variation through Few-Shot Prompting, Feb. 2023.

S. Mirchandani, F. Xia, P. Florence, B. Ichter, D. Driess, M. G. Arenas, K. Rao, D. Sadigh, and A. Zeng. Large Language Models as General Pattern Machines, July 2023.

I. Momennejad, H. Hasanbeig, F. Vieira Frujeri, H. Sharma, N. Jojic, H. Palangi, R. Ness, and J. Larson. Evaluating cognitive maps and planning in large language models with cogeval. *Advances in Neural Information Processing Systems*, 36, 2024.

M. R. Morris, J. Sohl-dickstein, N. Fiedel, T. Warkentin, A. Dafoe, A. Faust, C. Farabet, and S. Legg. Levels of AGI: Operationalizing Progress on the Path to AGI, Nov. 2023.

J.-B. Mouret and J. Clune. Illuminating search spaces by mapping elites. *ArXiv preprint*, abs/1504.04909, 2015. URL https://arxiv.org/abs/1504.04909.

OEL Team, A. Stooke, A. Mahajan, C. Barros, C. Deck, J. Bauer, J. Sygnowski, M. Trebacz, M. Jaderberg, M. Mathieu, N. McAleese, N. Bradley-Schmieg, N. Wong, N. Porcel, R. Raileanu, S. Hughes-Fitt, V. Dalibard, and W. M. Czarnecki. Open-ended learning leads to generally capable agents. *ArXiv preprint*, abs/2107.12808, 2021. URL https://arxiv.org/abs/2107.12808.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback, Mar. 2022.

V. Pallagani, B. Muppasani, K. Murugesan, F. Rossi, B. Srivastava, L. Horesh, F. Fabiano, and A. Loreggia. Understanding the capabilities of large language models for automated planning. *arXiv preprint arXiv:2305.16151*, 2023.

J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. Generative Agents: Interactive Simulacra of Human Behavior, Apr. 2023.

J. Parker-Holder, M. Jiang, M. Dennis, M. Samvelyan, J. N. Foerster, E. Grefenstette, and T. Rocktäschel. Evolving curricula with regret-based environment design. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 17473–17498. PMLR, 2022. URL https://proceedings.mlr.press/v162/parker-holder22a.html.

D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell. Curiosity-driven exploration by self-supervised prediction. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 2778–2787. PMLR, 2017. URL http://proceedings.mlr.press/v70/pathak17a.html.

J. Perolat, B. De Vylder, D. Hennes, E. Tarassov, F. Strub, V. de Boer, P. Muller, J. T. Connor, N. Burch, T. Anthony, et al. Mastering the game of stratego with model-free multiagent reinforcement learning. *Science*, 378(6623): 990–996, 2022.

J. K. Pugh, L. B. Soros, and K. O. Stanley. Quality diversity: A new frontier for evolutionary computation. *Frontiers in Robotics and AI*, 3:40, 2016. ISSN 2296-9144. doi: 10.3389/frobt.2016.00040.

R. Raileanu and T. Rocktäschel. RIDE: Rewarding Impact-Driven Exploration for Procedurally-Generated Environments, Feb. 2020.

P. J. Richerson, R. Boyd, et al. Institutional evolution in the holocene: the rise of complex societies. In *Proceedings-British Academy*, volume 110, pages 197–234. Oxford University Press Inc., 2001.

B. Romera-Paredes, M. Barekatain, A. Novikov, M. Balog, M. P. Kumar, E. Dupont, F. J. R. Ruiz, J. S. Ellenberg, P. Wang, O. Fawzi, P. Kohli, and A. Fawzi. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, Jan. 2024. ISSN 1476-4687. doi: 10.1038/s41586-023-06924-6.

A. L. Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229, 1959.

M. Samvelyan, A. Khan, M. Dennis, M. Jiang, J. Parker-Holder, J. Foerster, R. Raileanu, and T. Rocktäschel. MAESTRO: Open-Ended Environment Design for Multi-Agent Reinforcement Learning, Mar. 2023.

M. Samvelyan, S. C. Raparthy, A. Lupu, E. Hambro, A. H. Markosyan, M. Bhatt, Y. Mao, M. Jiang, J. Parker-Holder, J. Foerster, T. Rocktäschel, and R. Raileanu. Rainbow teaming: Open-ended generation of diverse adversarial prompts, 2024.

W. Saunders, C. Yeh, J. Wu, S. Bills, L. Ouyang, J. Ward, and J. Leike. Self-critiquing models for assisting human evaluators. *ArXiv preprint*, abs/2206.05802, 2022. URL https://arxiv.org/abs/2206.05802.

T. Schaul, D. Borsa, D. Ding, D. Szepesvari, G. Ostrovski, W. Dabney, and S. Osindero. Adapting behaviour for learning progress. *arXiv preprint arXiv:1912.06910*, 2019.

T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, and T. Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36, 2024.

J. Schmidhuber. *Adaptive confidence and adaptive curiosity*. Inst. für Informatik, 1991a.

J. Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proc. of the international conference on simulation of adaptive behavior: From animals to animats*, pages 222–227, 1991b.

J. Secretan, N. Beato, D. B. D Ambrosio, A. Rodriguez, A. Campbell, and K. O. Stanley. Picbreeder: Evolving pictures collaboratively online. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 1759–1768, New York, NY, USA, Apr. 2008. Association for Computing Machinery. ISBN 978-1-60558-011-1. doi: 10.1145/1357054.1357328.

G. Shackle. *Expectation in Economics*. Cambridge University Press, 1949. ISBN 9781107629141. URL https://books.google.co.uk/books?id=zEb47udAsOcC.

R. Shah, V. Varma, R. Kumar, M. Phuong, V. Krakovna, J. Uesato, and Z. Kenton. Goal misgeneralization: Why correct specifications aren't enough for correct goals. *ArXiv preprint*, abs/2210.01790, 2022. URL https://arxiv.org/abs/2210.01790.

A. Sharma, D. Czégel, M. Lachmann, C. P. Kempes, S. I. Walker, and L. Cronin. Assembly theory explains and quantifies selection and evolution. *Nature*, 622(7982): 321–328, Oct. 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06600-9.

M. Shin, J. Kim, B. van Opheusden, and T. L. Griffiths. Superhuman Artificial Intelligence Can Improve Human Decision Making by Increasing Novelty. *Proceedings of the National Academy of Sciences*, 120(12):e2214840120, Mar. 2023. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2214840120.

I. Shumailov, Z. Shumaylov, Y. Zhao, Y. Gal, N. Papernot, and R. Anderson. Model dementia: Generated data makes models forget. *arXiv e-prints*, pages arXiv–2305, 2023.

P. Shyam, W. Jaskowski, and F. Gomez. Model-based active exploration. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5779–5788. PMLR, 2019. URL http://proceedings.mlr.press/v97/shyam19a.html.

O. Sigaud, G. Baldassarre, C. Colas, S. Doncieux, R. Duro, N. Perrin-Gilbert, and V.-G. Santucci. A definition of open-ended learning problems for goal-conditioned agents. *ArXiv preprint*, abs/2311.00344, 2023. URL https://arxiv.org/abs/2311.00344.

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, Jan. 2016. ISSN 1476-4687. doi: 10.1038/nature16961.

D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis. Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm, Dec. 2017.

R. J. Solomonoff. A preliminary report on a general theory of inductive inference. Citeseer, 1960.

L. Soros and K. Stanley. Identifying necessary conditions for open-ended evolution through the artificial life world of chromaria. In *ALIFE 14: The Fourteenth International Conference on the Synthesis and Simulation of Living Systems*, ALIFE 2023: Ghost in the Machine: Proceedings of the 2023 Artificial Life Conference, pages 793–800, 2014. doi: 10.1162/978-0-262-32621-6-ch128. URL https://doi.org/10.1162/978-0-262-32621-6-ch128.

K. Stanley and J. Lehman. *Why Greatness Cannot Be Planned: The Myth of the Objective*. Springer International Publishing, 2015. ISBN 9783319155241. URL https://books.google.co.uk/books?id=Llb1CAAAQBAJ.

K. O. Stanley. Why open-endedness matters. *Artificial Life*, 25(3):232–235, 2019. ISSN 1064-5462. doi: 10.1162/artl_a_00294. URL https://doi.org/10.1162/artl_a_00294.

K. 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.

K. O. Stanley, J. Lehman, and L. Soros. Open-endedness: The last grand challenge you've never heard of. *While open-endedness could be a force for discovering intelligence, it could also be a component of AI itself*, 2017.

S. Stepney and S. Hickinbotham. On the open-endedness of detecting open-endedness. *Artificial Life*, pages 1–26, 2023.

R. J. Sternberg and J. E. Davidson. *The nature of insight*. The MIT Press, 1995.

N. N. Taleb. *Antifragile: Things that gain from disorder*, volume 3. Random House Trade Paperbacks, 2014.

X. Tang, A. Zou, Z. Zhang, Y. Zhao, X. Zhang, A. Cohan, and M. Gerstein. Medagents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537*, 2023.

T. Taylor. Requirements for open-ended evolution in natural and artificial systems. *arXiv preprint arXiv:1507.07403*, 2015.

T. Taylor. Routes to open-endedness in evolutionary systems. *arXiv preprint arXiv:1806.01883*, 2018.

A. M. Turner, D. Hadfield-Menell, and P. Tadepalli. Conservative agency via attainable utility preservation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 385–391, 2020.

F. Urbina, F. Lentzos, C. Invernizzi, and S. Ekins. Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, 4(3):189–191, 2022.

K. Valmeekam, M. Marquez, S. Sreedharan, and S. Kambhampati. On the planning abilities of large language models-a critical investigation. *Advances in Neural Information Processing Systems*, 36:75993–76005, 2023.

P. Villalobos, J. Sevilla, L. Heim, T. Besiroglu, M. Hobbhahn, and A. Ho. Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning, Oct. 2022.

O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, J. Oh, D. Horgan, M. Kroiss, I. Danihelka, A. Huang, L. Sifre, T. Cai, J. P. Agapiou, M. Jaderberg, A. S. Vezhnevets, R. Leblond, T. Pohlen, V. Dalibard, D. Budden, Y. Sulsky, J. Molloy, T. L. Paine, C. Gulcehre, Z. Wang, T. Pfaff, Y. Wu, R. Ring, D. Yogatama, D. Wünsch, K. McKinney, O. Smith, T. Schaul, T. Lillicrap, K. Kavukcuoglu, D. Hassabis, C. Apps, and D. Silver. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575 (7782):350–354, Nov. 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-1724-z.

L. S. Vygotsky and M. Cole. *Mind in society: Development of higher psychological processes*. Harvard university press, 1978.

C. H. Waddington. Paradigm for an evolutionary process. *Biological Theory*, 3:258–266, 2008.

G. Wang, Y. Xie, Y. Jiang, A. Mandlekar, C. Xiao, Y. Zhu, L. Fan, and A. Anandkumar. Voyager: An Open-Ended Embodied Agent with Large Language Models, May 2023a.

R. Wang, J. Lehman, J. Clune, and K. O. Stanley. Paired open-ended trailblazer (POET): endlessly generating increasingly complex and diverse learning environments and their solutions. *ArXiv preprint*, abs/1901.01753, 2019. URL https://arxiv.org/abs/1901.01753.

R. Wang, J. Lehman, A. Rawal, J. Zhi, Y. Li, J. Clune, and K. O. Stanley. Enhanced POET: open-ended reinforcement learning through unbounded invention of learning challenges and their solutions. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9940–9951. PMLR, 2020. URL http://proceedings.mlr.press/v119/wang20l.html.

T. T. Wang, A. Gleave, T. Tseng, K. Pelrine, N. Belrose, J. Miller, M. D. Dennis, Y. Duan, V. Pogrebniak, S. Levine, et al. Adversarial policies beat superhuman go ais. In *International Conference on Machine Learning*, pages 35655–35739. PMLR, 2023b.

Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *ArXiv preprint*, abs/2212.10560, 2022. URL https://arxiv.org/abs/2212.10560.

Z. Wang, S. Cai, A. Liu, Y. Jin, J. Hou, B. Zhang, H. Lin, Z. He, Z. Zheng, Y. Yang, X. Ma, and Y. Liang. JARVIS-1: Open-World Multi-task Agents with Memory-Augmented Multimodal Language Models, Nov. 2023c.

L. Wong, G. Grand, A. K. Lew, N. D. Goodman, V. K. Mansinghka, J. Andreas, and J. B. Tenenbaum. From Word Models to World Models: Translating from Natural Language to the Probabilistic Language of Thought, June 2023a.

M. L. Wong, C. E. Cleland, D. Arend Jr, S. Bartlett, H. J. Cleaves, H. Demarest, A. Prabhu, J. I. Lunine, and R. M. Hazen. On the roles of function and selection in evolving systems. *Proceedings of the National Academy of Sciences*, 120(43):e2310223120, 2023b.

X. Wu, S.-h. Wu, J. Wu, L. Feng, and K. C. Tan. Evolutionary computation in the era of large language model: Survey and roadmap. *arXiv preprint arXiv:2401.10034*, 2024.

Y. Wu, S. Prabhumoye, S. Y. Min, Y. Bisk, R. Salakhutdinov, A. Azaria, T. Mitchell, and Y. Li. SPRING: GPT-4 Out-performs RL Algorithms by Studying Papers and Reasoning, May 2023.

C. Yang, X. Wang, Y. Lu, H. Liu, Q. V. Le, D. Zhou, and X. Chen. Large Language Models as Optimizers, Sept. 2023a.

M. Yang, Y. Du, K. Ghasemipour, J. Tompson, D. Schuurmans, and P. Abbeel. Learning Interactive Real-World Simulators, Oct. 2023b.

W. Yuan, R. Y. Pang, K. Cho, S. Sukhbaatar, J. Xu, and J. Weston. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 2024.

J. Zhang, J. Lehman, K. Stanley, and J. Clune. OMNI: Open-endedness via Models of human Notions of Interestingness, June 2023.

B. Zheng, B. Gou, J. Kil, H. Sun, and Y. Su. GPT-4V(ision) is a Generalist Web Agent, if Grounded, Jan. 2024.

S. Zhou, F. F. Xu, H. Zhu, X. Zhou, R. Lo, A. Sridhar, X. Cheng, T. Ou, Y. Bisk, D. Fried, U. Alon, and G. Neubig. Webarena: A realistic web environment for building autonomous agents. In *Second Agent Learning in Open-Endedness Workshop*, 2023. URL https://openreview.net/forum?id=rmiwIL98uQ.

D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving. Fine-tuning language models from human preferences. *ArXiv preprint*, abs/1909.08593, 2019. URL https://arxiv.org/abs/1909.08593.

A. Illustrating Open-Endedness

A.1. An Informal Example

To illustrate our definition informally, we provide a relatable real-world example. Let S be a research lab and the x_t be academic papers published by the lab. A natural choice of observer O is a research student in the field at a different lab. Roughly speaking, a research student sees novelty in a line of work if, based on their knowledge of the literature up to time t, given any subsequent paper x_T they can always find a later paper $x_{T'}$ that is more surprising than x_T. This is intuitively sensible, a putative student with knowledge of Newtonian mechanics will find Maxwell's equations hard to predict, quantum mechanics even more surprising, and contemporary particle physics very far outside their current level of comprehension. A research student sees learnability in a line of work if they find that reading the previous papers helps them better to predict the contents of the current paper. Again, this appeals to our intuition: part of the purpose of citations, for instance, is to point new researchers at previous works that will help to deepen their understanding of the current work.

Our interpretation of "interestingness" as learnability also makes sense from the perspective of a research student. A research student may choose to ignore a paper's choice of font, but will likely pay close attention to the details of a novel method that yields state-of-the-art results. Thus the student finds interesting the parts of the paper from which they can learn the most. Similarly, the requirement that the loss metric ℓ be chosen without knowledge of S finds a natural interpretation here. A research student cannot judge the open-endedness of a stack of papers by choosing to never read the papers and instead inventing their own line of research with no reference to previous works.

A.2. Definitional Subtleties

Self-play illustrates some subtleties in our definition. The first subtlety is the dependence of open-endedness on the choice of observer. Suppose that O is an oracle who knows the Nash strategy to play in Go. Assuming that the oracle is modelling the win-rate of AlphaZero's artifacts against its own policy, it will never find any AlphaZero policy to be novel. Therefore the oracle does not find AlphaZero to be open-ended. The second subtlety is the dependence of open-endedness on the learning limitations of the observer. To an average human Go player, as opposed to an expert, AlphaZero becomes novel earlier in training, and at some point ceases to be learnable, because the average player cannot figure out how to improve their own play with reference to very unusual style of a superhuman policy. Thus, open-ended systems only remain open-ended while they can "educate" their observers. We posit that superhuman intelligence will be interesting to humans only as far as

humans can learn to understand it. The third subtlety is that open-ended systems need not explore a problem space fully to qualify as open-ended. Recently, adversarial search was shown to yield policies that beat reimplementations of AlphaZero and which are so simple that even amateur humans can learn them (Wang et al., 2023b). Novelty and learnability give no guarantee of coverage.

Because our definition is based on the perspective of an external observer, one could worry that this makes it impossible to make any sort of objective claims about the open-endedness of any particular system, in harmony with the arguments of Stanley and Soros (2016); Stepney and Hickinbotham (2023). There are two factors which mitigate this concern. Firstly, the definition of open-endedness becomes objective given any fixed observer, and so it becomes a measurable claim, in the sense that theorems can be written and experiments conducted. For instance, if we care about open-endedness with respect to humans, open-endedness can be measured experimentally by how well humans can predict the system. By having observer-dependence explicit in our definition, we make precise the intuition that different observers, with different prior knowledge, different cognitive capabilities and different timescales, are likely to judge the same system in different ways. Thus our definition gracefully encompasses the diversity in perspectives of human individuals and groups (such as companies or governments), as well as the possibility that AI systems themselves could be observers.

Secondly, while our definition of open-endedness depends on an external observer, it is an open question as to whether all "reasonable" observers would judge the same systems to be open-ended. Since our definition rests on a notion of predictability with respect to the observer, our definition will be as subjective as the underlying notion of predictability. One may believe that predictability can be accurately and objectively modeled as Solomonoff induction (Solomonoff, 1960). Thus if reasonable observers are taken to be those whose predictions eventually follow something approximating Solomonoff induction, then any observer in this class would eventually agree on which systems are open-ended.

Practically speaking, there are various existing methods in the literature which can immediately be adapted to assess the open-endedness of a system. First, one might elicit direct human feedback on learnability and novelty of artifacts, in the same spirit as RLHF (Ouyang et al., 2022) or PicBreeder (Secretan et al., 2008). Second, one can use large language models themselves as judges of novelty and learnability, as argued for in OMNI (Zhang et al., 2023). Finally, one could explicitly learn a model of the artifacts with an online learning method like Follow-the-Regularized-Leader (Hazan and Kale, 2010).

Can an open-ended system be its own observer? In principle, there is nothing in our definition that rules out self-observing open-ended systems. For example, an individual self-improving agent could generate a series of artifacts, each one of which is novel (surprising compared to the previous artifacts) and learnable (increasingly predictable given the more history of the past artifacts). When the feedback from self-observation is used to improve the system itself, we call the observer a *proxy observer* for it no longer sits outside the system.

For example, AlphaGo can be seen as an example of a self-observing system, in that the agent trains in self-play i.e. it observes its own policy as an opponent, is challenged by the novel discoveries of search, and learns from them to improve the policy. Likewise, humans can experience "Eureka moments", when an individual suddenly reconceptualizes a problem in a ways that yields a solution (Sternberg and Davidson, 1995). A series of Eureka moments, each building on the last, is a self-observing open-ended system: the human generates discoveries which are novel to themselves, but which are also predictive of the next discovery.

Our notions of learnability is rather strict, in that it requires that the loss be decreasing for all $t' > t$. A weaker and more practical notion of learnability might state that it should be probabilistically unlikely that the loss will increase as a function of t:

$$\forall T, \forall t < T, \forall T > t' > t : \mathbb{P}\left(\ell(t', T) \geq \ell(t, T)\right) < \delta\,.$$

It would be interesting to compare the consequences of δ being a constant with the situation in which δ has some appropriate dependence on the variables (t, t', T). Similarly, one could weaken the notion of novelty to state that it should be probabilistically unlikely that the loss will decrease as a function of T. We believe that there may be several related and differently useful variants on our definition that would be interesting to independently study, in a similar way that there are many notions of convergence which are interesting, related, and differently useful.

B. Alternative Definition

In Section 2.1 we provided a formal definition of open-endedness in the language of statistical learning. Here we give an alternative definition which we conjecture is equivalent under appropriate conditions. The alternative definition is phrased in the language of compression, a topic with known formal connections to statistical learning (Hutter, 2004; David et al., 2016; Campi and Garatti, 2023; Delétang et al., 2023).

A **system** S produces a sequence of **artifacts** $X_t \in \mathcal{X}$, indexed by time t. An **observer** O processes a new artifact X_T to determine its information content given a history $h_t = X_{1:t}$ of past ones. O possesses a history-dependent

compression map $C_{h_t} : \mathcal{X} \to \{0,1\}^*$ which encodes X_T into a binary string of length $|C_{h_t}(X_T)|$.

The system displays **novelty** if the information content increases, namely:

$$\forall t, \forall T > t, \exists T' > T : |C_{h_t}(X_{T'})| > |C_{h_t}(X_T)|.$$

In other words, the complexity of the artifacts grows, according to the observer.

The system is **learnable** if conditioning on a longer history increases compressibility, namely:

$$\forall T, \forall t < T, \forall T > t' > t : |C_{h_{t'}}(X_T)| < |C_{h_t}(X_T)|.$$

In other words, as its history grows, the observer must be able to keep extracting additional patterns that help it compress future artifacts.

Finally, a system is **open-ended** from the perspective of O if and only if it generates sequences of artifacts that are both novel and learnable.

We allow for the compression map C_{h_t} to be **lossy**. Hence, O also possesses a decompression map $D_{h_t} : \{0,1\}^* \to \mathcal{X}$, a symmetric loss function $\ell : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$, and a threshold $\epsilon \in \mathbb{R}^+$ that upper-bounds the error made by by C_{h_t}:

$$\forall T, \forall t < T : \ell(D_{h_t}(C_{h_t}(X_T)), X_T) < \epsilon.$$

We can strengthen the definition to be independent of ϵ by appealing to rate-distortion theory. A **rate-distortion** curve plots the the minimum information content $|C_h(X)|$ such that $\ell(D_h(C_h(X)), X) < \epsilon$ against ϵ, where the minimum is over the maps C_h and D_h. The information content is referred to as the rate and ϵ is referred to as the distortion. Picture a grid of rate-distortion curves G_{tT} indexed by (discretized) t and T, as in Figure 3. Remember that $T > t$, so G_{tT} is strictly upper triangular, with other entries being undefined. Then **broad novelty** is the requirement that the curves get "fatter" as you move across the columns T on the grid, for every row t. Similarly, **broad learnability** is the requirement that the curves get "flatter" as you move down the rows t on the grid, for every column T. **Broad open-endedness** is the requirement that both broad novelty and broad learnability hold. This notion of broad open-endedness is vague in the same way the notion of "convergence" is vague in that it can be made precise in many subtly different but connected ways. For instance, one could say a system is "uniformly" open-ended if distortion increases across the rows and decreases down the columns for every rate ϵ. Alternatively, one could define "average" open-endedness by requiring that the integral of the rate-distortion curve get larger as you move across the columns and smaller as you move down the rows. We hope that future work will elucidate these subtleties in defining broad open-endedness and determine which variants have theoretical or practical merit.

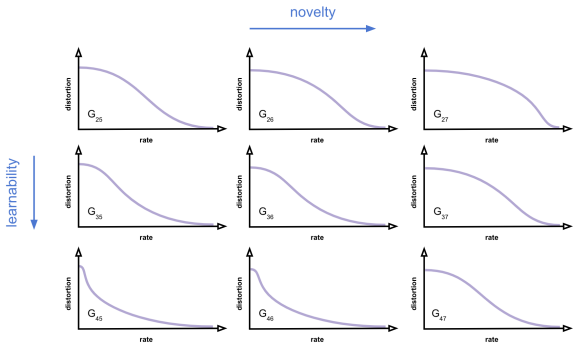

Figure 3. **Open-endedness through the lens of rate-distortion curves.** We depict part of the upper triangular matrix of rate-distortion curves G_{tT} induced an observer after seeing the first t artifacts aiming to lossily compress future artifact T. Here $t = 2, 3, 4$ and $T = 5, 6, 7$. Broad novelty is the property that, as you move from left to right in any fixed row, the rate-distortion curves become fatter. Broad learnability is the property that, as you move from top to bottom in any fixed column, the curves become flatter. For the system to be broadly open-ended, both properties must hold.

C. Further Related Work

Open-endedeness as a term emerged from the AI Life community when trying to quantify and replicate the increasing complexity and perpetual novelty of biological evolution. This is a rich field with a significant degree of disagreement (Earle et al., 2021). As such there are a wide range of metrics proposed within the context of evolutionary systems which aim to quantify it's behavior. For instance persistence filtering, which measures how many generations an organism has persisted for (Dolson et al., 2019), and evolutionary activity statistics (Bedau et al., 1997; 1998). The closely related question around the necessary conditions to produce open-ended evolution has also been deeply studied (Taylor, 2018; 2015). As these definitions are largely specific to biological evolution, we focus the remainder of our discussion on the more recent definitions which aim to define open-ended systems in a way that applies to current ML systems and systems more broadly.

Our definition of open-endedness is closely related to the concept of potential surprise in economics (Shackle, 1949). To measure *potential surprise*, an individual should ask: "how surprised would I be if this outcome actually occurred, if, at the time it occurred, I were still looking at the world in the way I look at it right now?" (Derbyshire, 2017). Interpreting surprise as unpredictability under a statistical model, an open-ended system S is precisely one which produces ever increasing "Shackle surprise" in an observer which is learning. The concept of potential surprise is itself based on the century-old idea of Knightian uncertainty (Knight,

1921). *Knightian uncertainty* is a lack of any quantifiable knowledge about some possible occurrence, as opposed to the presence of quantifiable risk. Thus, somewhat imprecisely, an open-ended system S is one which induces Knightian uncertainty in an observer who is learning.

In Stanley and Lehman (2015), the authors argue that local search for novel and interesting artifacts can be advantageous over optimization for a global objective. This is because stepping stones towards a solution that optimizes the global objective may well not resemble the solution itself. Hence it is hard to translate the global objective into a local improvement operator that reliably accumulates improvements without getting stuck in local optima. To address this deceptiveness, they suggest that novelty search (Lehman and Stanley, 2011), guided by a notion of interestingness, can uncover stepping stones that advance knowledge and capability. We take inspiration from this blueprint and turn it into a definition. In order to clarify the notions of novelty and interestingness, we formalize them with respect to an external observer. Novelty becomes unpredictability according to the observer's history-conditional model, and interestingness becomes learnability of that model across the history of observations.

Our definition naturally relates to the notion of curiosity. Curiosity, implemented as prediction error of a world model, has long been mooted as an intrinsic motivation that can lead to open-ended discovery in RL agents given a sufficiently rich environment space (Schmidhuber, 1991b; Pathak et al., 2017; Raileanu and Rocktäschel, 2020; Henaff et al., 2023). Our definition of novelty is effectively a generalisation of curiosity, without requiring an overarching RL framework. Our requirement of learnability ensures that the observer attempts to capture all the epistemic uncertainty about the artifacts produced by a system. One challenge is that curiosity based on novelty alone leads to "stochastic traps", whereby an agent will seek out sources of random noise with which to sate its curiosity (Schmidhuber, 1991a; Burda et al., 2018; Shyam et al., 2019). In principle, our definition of novelty collapses such aleatoric uncertainty by taking the expectation. In practice, we can only estimate the expectation, so it may be useful to subtract from the loss an estimate of the aleatoric uncertainty as in Mavor-Parker et al. (2022). We hope that future work will examine such subtleties required for an algorithmic implementation of our definition.

The synergies between foundation models and open-endedness have previously been discussed by Jiang et al. (2022). The authors propose a general notion of exploration and detail how open-endedness can be used to solve exploration problems when training foundation models. Our work follows in this line of thinking, providing a formal definition of open-endedness to make the discussion precise, and further developing the connections between open-endedness and ASI. A construction of a particular open-ended learning system is provided in (Jiang et al., 2022), which may or may not fit our proposed definition of an open-ended system depending on how it is instantiated. The system generates Turing machine descriptions of MDPs, explicitly optimizing for an objective containing terms for learning potential, diversity, and grounding. These terms have some high-level relation to our notions of learnability and novelty, but they are quite distinct in the details. For instance, learning potential is divided into three sub-critia, improbability, learnability, and consistency, which are not made entirely formal. More crucially, the learnability discussed by (Jiang et al., 2022) is a property of a single MDP, whereas the learnability we define is a property of a sequence of artifacts. Similarly, in (Jiang et al., 2022) diversity is defined as a distance measure between MDPs, whereas novelty, as we define it, is a property of the learning of the observer with no necessary relationship to distances in the space of artifacts. It would be an interesting direction for future research to understand under what conditions the system described in (Jiang et al., 2022) would be open-ended by our definition, and, more generally, whether one can directly optimize for open-endedness in some circumstances.

Open-endedness is related to, but separate from, the notion of an AI-generating algorithm (AIGA, Clune, 2020). An AIGA automatically learns how to build a general AI, based on meta-learning model architectures, meta-learning learning algorithms, and automatically generating data from which to learn. Adapting the logic of Clune (2020), an AIGA need not be open-ended by our definition; if an AIGA had the objective of passing a Turing test, it need not produce any further novelty once this objective had been achieved. Likewise, an open-ended system need not be an AIGA; as we shall see in Section 2.4, there exist open-ended systems with narrow scope that match or exceed human ability without full domain-generality. Our idea of an Open-Ended Foundation Model in Section 3 lives at the intersection between open-endedness and AIGAs.

Similarly open-endedness is related to, but distinct from, continual RL (Abel et al., 2023). A continual RL problem is one in which the best agents never stop learning. However, as observed by (Sigaud et al., 2023), this does not necessarily imply that the agent policies *accumulate* increasing novelty. Rather, a continual RL agent could cycle among some set of strategies. In the case where continual RL does produce policies which are open-ended according to some observer, this open-endedness will have a scope that is restricted by the environment.