Open Questions in Creating Safe Open-ended AI: Tensions Between Control and Creativity

Adrien Ecoffet[1], Jeff Clune[1,2] and Joel Lehman[1]

[1]Uber Technologies, San Francisco, CA 94105
[2]OpenAI, San Francisco, CA 94110 (work done at Uber AI)
lehman.154@gmail.com

Abstract

Artificial life originated and has long studied the topic of *open-ended evolution*, which seeks the principles underlying artificial systems that innovate continually, inspired by biological evolution. Recently, interest has grown within the broader field of AI in a generalization of open-ended evolution, here called *open-ended search*, wherein such questions of open-endedness are explored for advancing AI, whatever the nature of the underlying search algorithm (e.g. evolutionary or gradient-based). For example, open-ended search might design new architectures for neural networks, new reinforcement learning algorithms, or most ambitiously, aim at designing artificial general intelligence. This paper proposes that open-ended evolution and artificial life have much to contribute towards the understanding of open-ended AI, focusing here in particular on the *safety* of open-ended search. The idea is that AI systems are increasingly applied in the real world, often producing unintended harms in the process, which motivates the growing field of AI safety. This paper argues that open-ended AI has its own safety challenges, in particular, whether the creativity of open-ended systems can be productively and predictably controlled. This paper explains how unique safety problems manifest in open-ended search, and suggests concrete contributions and research questions to explore them. The hope is to inspire progress towards creative, useful, and safe open-ended search algorithms.

Introduction

Artificial life (ALife) and artificial intelligence (AI) have largely developed independently as fields. Statistical machine learning (ML), including deep learning, has driven much progress in modern AI research and practice, arguably with limited inspiration from ALife. One reason is that such statistical ML typically operates under a highly focused and directed paradigm (here called *directed search*): A formal objective function is defined that reflects the desired outcome of search, and a parameter vector is optimized to meet that objective. While ALife is also interested in the possibilities of digital intelligence, it approaches them more often through the lens of *open-ended search*: Conditions for a creative (often population-based and evolutionary) system are investigated, from which complexity and intelligence might emerge among its many diverse products.

Interestingly, however, ML has begun to recognize the value of open-ended search algorithms. An emerging trend is for ML to be applied to activities ordinarily undertaken by ML research scientists. For example, while designing architectures for neural networks (NNs) has historically been undertaken by researchers, interest is growing in automated neural architecture search. Similarly, meta-learning algorithms that instead of being hard-coded to learn, themselves *learn how to learn*, are an increasing focus of study, e.g. NNs that adapt their behavior during deployment (Finn, Abbeel, and Levine, 2017; Vilalta and Drissi, 2002; Soltoggio et al., 2008). The logical (if ambitious) culmination of this trend is for search algorithms to in effect pursue their own AI research programs, i.e. to subsume the activities of the AI research community as a whole (Stanley, Lehman, and Soros, 2017; Clune, 2019). That is, can search be applied to autonomously explore the space of AI algorithms, in principle to surpass current capabilities? Such an approach would require the equivalent of algorithmic *basic research*, i.e. conducting a more creative, less directed, and more open-minded search that respects that the ultimate potential of a new algorithm or NN paradigm is difficult to predict. For example, the potential of deep and convolutional NNs was unclear to the AI community for many years. This widely-recognized need for exploring broadly and incubating new ideas motivates how the research community simultaneously explores diverse ideas within many different schools of AI (e.g. symbolic, bio-inspired, cognitive architecture, and Bayesian approaches, among others).

While to most ML researchers, a search algorithm capable of such continual open-ended innovation might sound quixotic, this paradigm is familiar to ALife, and its precedence is supplied by the origins of human intelligence. Biological evolution in effect conducted its own undirected yet highly expansive and successful research and development program into intelligent computation, which in one branch of life led to human intelligence. A similarly-inspired approach to generating complex behavior has been pursued by the open-ended evolution (OEE; Standish, 2003; Packard, 1997; Taylor et al., 2016b) community within ALife, which

studies the principles driving processes of continual evolutionary innovation, often from the lens of generating intelligent behavior. While these research questions originated within ALife, they are beginning to influence the ML community. E.g. open-ended search has been presented to the ML community as a grand challenge (Stanley, Lehman, and Soros, 2017), an ambitious research agenda has been proposed that merges ML with open-ended search (Clune, 2019), and ML approaches to open-ended search are actively being explored (Guttenberg, Virgo, and Penn, 2019; Wang et al., 2019; Akkaya et al., 2019). Thus the aims of ALife and AI more directly overlap than they have in past.

Another important trend is that as the real-world application of AI grows, so does concern over AI's safe and predictable deployment, as studied by the growing field of *AI safety* (Amodei et al., 2016; Ortega and Maini, 2018; Everitt, Lea, and Hutter, 2018). Increasingly, AI is applied in domains where it can impact human well-being, such as in determining risk for loans, making recommendations for parole, and controlling autonomous robots, thereby making unanticipated failures costly. AI safety seeks to understand and mitigate causes for an AI agent's *actual* behavior to diverge from what it was *intended* to do. For example, intuitive human-designed fitness functions can be optimized in undesirable ways (Lehman et al., 2018) and agents can fail catastrophically when deployed if training does not anticipate gamut of possible real-world scenarios (Hadfield-Menell et al., 2017). Interestingly, while not called "ALife safety," similar questions about the predictability of open-ended systems have been studied in ALife (Wagenaar and Adami, 2004; Taylor and Hallam, 1998), and likely bear on the safety of open-ended search. More generally, the extent to which the creativity of open-ended algorithms can be controlled (Stanley and Lehman, 2015; Lehman et al., 2018) remains an important and open question, one relevant both to ALife and AI safety.

In this way, research and ways of thinking about open-ended search can become a strong contribution from ALife to AI, as ALife and OEE have for years considered the complexities and surprising dynamics of creative algorithms, while it remains a relatively new topic in ML. Overall, the idea is that as open-ended search becomes more popular, it will be important to understand *if* and *how* the creativity of open-ended systems (whether evolutionary or otherwise) can be predictably and safely leveraged for practical applications. In this paper, we lay out concrete connections between open-ended search and active research questions in AI safety, and suggest ways that researchers can make productive contributions.

Background

Open-ended Search

Historically, open-ended search algorithms have been inspired by biological evolution, and studied mainly by the open-ended evolution community (Standish, 2003; Ray, 1991; Ofria and Wilke, 2004). Evolution instantiates an incredible process of continual innovation that has, over the course of billions of years, autonomously produced a wild diversity of complex and adaptive solutions to the challenges of living and reproducing; the idea in OEE is that if the core logic of biological evolution is understood, it becomes possible to instantiate such prolific creativity in alternative forms, e.g. within computational simulated environments. Typical OEE systems embody an evolutionary process in a digital environment, wherein the only goals are to survive and replicate. E.g. in Tierra (Ray, 1991), digital self-copying programs evolve within a shared memory ecosystem, enabling complex ecological interactions. After initialization with a hand-designed replicator, evolution in Tierra proceeds to create co-evolutionary arm races of parasites and hyper-parasites. Other evolutionary approaches seek abstract engineering principles to enable domain-independent open-endedness (Lehman and Stanley, 2011; Brant and Stanley, 2017; Wang et al., 2019), such as formulating OEE as a continual search for novelty.

Interestingly, ideas from OEE have recently begun to influence ML. In particular, there is increasing interest in ML algorithms that themselves learn to innovate (e.g. to invent new search algorithms and architectures). As a result, open-ended search is now being pursued within the paradigm of statistical ML (Guttenberg, Virgo, and Penn, 2019; Wang et al., 2019; Akkaya et al., 2019). As such efforts leverage increasing interest and compute, progress in open-endedness research may accelerate, further motivating study of its safety profile, as real-world applications emerge (Akkaya et al., 2019). Note that the term open-ended search here encompasses both OEE in ALife and open-ended ML algorithms; while the exact mechanisms of open-ended search are different between ALife and ML (e.g. evolutionary algorithms vs. gradient descent), they share the same abstract core; we focus on open-ended search that produces *agents*, as in many ALife OEE worlds, evolutionary robotics, and the field of reinforcement learning (RL) within ML.

AI Safety

AI safety seeks technical solutions to problems that cause AI behavior to diverge problematically from its designer's intentions (Amodei et al., 2016; Ortega and Maini, 2018). That is, AI algorithms even *without explicit bugs* can succeed by their own metrics and still fail to meet their designer's goals. One decomposition of AI safety problems is provided by Ortega and Maini (2018): specification, robustness, and assurance problems. *Specification problems* result from divergences between the goal intended for an agent and the optimizing behavior that is revealed empirically. E.g. reward hacking is where optimization uncovers undesirable ways to maximize a human-designed fitness function (e.g. a robotic vacuum rewarded for collecting dirt might discover it

can puncture its bag and continually collect the same dirt ad infinitum). *Robustness problems* result from when perturbations to the system result in unsafe behavior (e.g. if the vacuum encounters an object outside of its training data, like a vase, and breaks it because there is no understanding that it is fragile). Finally, *assurance problems* relate to understanding an AI system and maintaining control of it, e.g. whether the agent's control policy is interpretable or whether the agent can easily and safely be turned off if there is a problem. For a more comprehensive review of challenges in AI safety, see Everitt, Lea, and Hutter (2018) and Amodei et al. (2016).

Approach: Safety in Open-ended Search

Recall that AI safety in ML is largely focused on top-down control, while ALife and OEE typically focus more on the emergence of complexity from diversifying search. This section argues for a separate AI safety agenda driven by such a bottom-up view. We posit that the main aim of such a safety agenda is to understand more deeply the fundamental tension between creativity and control in open-ended search. That is, can an open-ended search be constrained such that its products are safe, and if so, how?

One might doubt that such constraint is possible, as open-ended search processes instantiate complex systems (Mitchell, 2009), often involving co-evolution, chaos, emergence, exaptation, path-dependence, Nth-order effects, and other phenomena studied within complex systems theory. In other words, the initial conditions of a system are often so far removed from its eventual products that it may seem intractable to predict a priori the qualitative effects (and the safety of such effects) that even subtle changes to such initial conditions bring about as they ripple through the system's unfolding dynamics. On the other hand, there may be important higher-level regularities within open-ended search that do form predictable and exploitable attractors. We suggest that further research can help explore this potential.

How Safety Issues Emerge in Open-ended Search

Open-ended search involves multiple levels of optimization in a way that qualitatively differs from directed search. Understanding such levels gives insight into how open-ended search can diverge from the system designer's intents, creating potential safety hazards. Note that the categorization presented next is adapted from previous AI safety categorizations (Ortega and Maini, 2018; Hubinger et al., 2019).

Ideal Objective First, when designing or applying an open-ended search, an experimenter has in mind their *ideal objective*, which depends upon their aspirations and intents. For example, an experimenter might leverage open-endedness to solve concrete problems (Lehman and Stanley, 2008; Akkaya et al., 2019), to create explosions of complex diversity to systematically understand the phenomenon of open-endedness itself (Standish, 2003), or to attempt to create artificial general intelligence (Clune, 2019). Implicitly, this ideal objective also includes safety: If an experimenter is interested in solving a real-world problem, or in creating artificial general intelligence (AGI), they likely intend to do so without causing harm.

Explicit Incentives Next, the experimenter chooses to implement the ideal objective concretely in an algorithm, resulting in the algorithm's *explicit incentives*, i.e. the actual optimization pressure driving search. Divergences between the ideal objective and what results from optimizing the explicit incentives relate to specification problems in AI safety.

In directed search, the explicit incentive is nearly always a direct translation of the ideal objective (i.e. if the ideal objective is a high-performing classifier, the explicit incentive may be to increase the classifier's accuracy). In contrast, in open-ended search the explicit incentive often represents a speculative hypothesis about what *creative forces* will result in producing (potentially among many diverse products of search) outcomes that satisfy the ideal objective. For example, when applying open-ended search in pursuit of AGI, one might abstract biological evolution as an open-ended search, where the ideal objective is to produce intelligence but the explicit incentive driving search is for organisms to survive and reproduce. From this view, while evolution's search found many diverse ways to survive and reproduce, including human intelligence, this explicit incentive is more like the codification of rules of an economy or incentives for innovation in science, rather than directly encouraging intelligent behavior. This kind of indirectness is more difficult to control, suggesting that safety problems in open-ended search may be more challenging than in directed search.

Agent Incentives Finally, in open-ended search processes that produce agents that are themselves capable of learning, such agents have emergent *agent incentives* that they in effect optimize. For example, human desires are related but distinct from the explicit incentives of survival and reproduction. Human desires embody *proxies* that encouraged survival and reproduction in our ancestral environment, e.g. hunger to encourage energy consumption. However, the fact that more die from obesity or drug addiction than starvation in first-world countries highlights how such proxy agent incentives often do not perfectly mirror a search process' ideal objective or its explicit incentives; this is an example of an AI safety robustness problem (i.e. agent incentives can become nonsensical when the environment changes from that experienced during training). Agent incentives also are intertwined with AI safety assurance problems, e.g. how to *interpret* what an evolved agent is doing, or whether it is indifferent to being turned off.

Case Study: Biological Evolution and AI Safety

As a case study contrasting ALife and complex systems thinking about safety with that common in directed search,

we next examine biological evolution from both such perspectives. Because biological evolution produced human intelligence and inspires open-ended search researchers who consider producing benefical AGI their ideal objective (e.g. the AI-GA paradigm; Clune, 2019), we here analyze biological evolution as if it had the ideal objective of producing beneficial general intelligence.

Interestingly, the explicit incentive of biological evolution, to persist by surviving and reproducing, seemingly encodes nothing about this ideal objective, and yet biological evolution did produce human intelligence, an amazing accomplishment that human engineering cannot yet replicate. Additionally, the agent incentives of humans significantly diverge from the explicit incentives of evolution, i.e. humans do not direct all their efforts towards maximizing their reproductive fitness, but instead are driven by proxies that encourage reproduction, such as sexual desire, that have become easy for humans to circumvent (e.g. through birth control). Finally, this divergence between human behavior and raw survival and reproduction is important, because it enables humans to *transcend* their biological imperative.

That is, humans can now use the adaptation of reason (that initally was well-aligned with evolution's explicit incentives) to understand the origins of their own desires, question their value, and create culture and institutions that pursue higher ends than mere inclusive genetic fitness. Arguably, much of humanity's positive potential has resulted from our ability to break free from the shackles of evolution; aspects of human life that many of us deem worthy of pursuit upon reflection, including e.g. creativity, virtue, deep intellectual engagement, spiritual experience, love, justice, an organization of society that promotes the flourishing of sentient life, would be optimized away as inefficient if we ruthlessly pursued the imperative to maximize reproduction, and deliberately optimized society intensely towards only that end. From this point of view, nearly everything of moral worth results from humanity transcending the explicit incentives of the search algorithm. In contrast, a central focus within top-down AI safety is to explicitly *align* an AI's incentives with our own (Hubinger et al., 2019; Taylor et al., 2016a), e.g. by modeling human preferences to use as an objective function (Leike et al., 2018), or to be cautious of divergences between explicit incentives and agent incentives (Hubinger et al., 2019).

One motivating failure case in AI safety is that a powerful optimizer given an innocuous-seeming (but incorrect, incomplete, or trivial) objective can produce disastrous outcomes (Bostrom, 2012). The canonical example (intended to be illustrative rather than realistic) is of a paperclip-maximizer: An agent seeking to manufacture as many paperclips as possible. The idea is that a superintelligent paperclip-maximizer would be incentivized to take extreme measures, e.g. usurp all planetary resources and tile the universe in paperclips, even though it could comprehend the triviality of its mandate. This phenomenon rests upon the orthogonality thesis (Armstrong, 2013; Bostrom, 2012), which proposes that an agent's *ability* to optimize and *what* it optimizes are orthogonal to one another: An arbitrarily powerful optimizer can optimize towards arbitrarily meaningless objectives. While contradicting human intuitions (i.e. it may seem incoherent that a "superintelligent" AI could be driven to pursue a meaningless goal, e.g. to restructure the universe into paperclips), it has relatively strong philosophical support (Armstrong, 2013). The design of such systems strongly seems possible, even though humans, for example, seem able to transcend (to some degree) their inborn desires.

Typically in AI safety, the orthogonality thesis motivates how *critical* it is to create AI with reward functions that reflect the full sophistication of human interests. The reasoning is that a powerful AI enchained to even a slightly-flawed objective may have incentive to engage in extreme and potentially disastrous behavior (Omohundro, 2008). That is, the current aim of AI safety within ML is mostly focused on the assumption of optimizers that are strongly wedded to a particular objective function. However, open-ended search is often concerned with systems in which what is optimized by search is merely an indirect proxy for a more expansive ideal objective, and in which it may even be *desirable* for the agents produced to transcend the explicit search incentives. The conclusion is that open-ended search may fundamentally be in tension with a top-down AI safety perspective.

Controllability of Innovative Systems To explore this tension between top-down control and bottom-up emergence, consider as a metaphor two societies organized in different ways, attempting to make progress towards developing a single goal technology: in the first, there is a vibrant community of basic research across all intellectual interests, and abundant sharing without restriction, of *all* scientific findings; in the second, a central agency highly controls what scientists work on and restricts what information is shared between them. The second paradigm seems so narrow and restrictive as to greatly impair progress, and the first is so open and free that discoveries that potentially should not be open (e.g. more effective methods of harm such as biological or nuclear weapons) might cause significant and regrettable side-effects. The purpose of this metaphor is to highlight that it is unclear exactly how to design systems of innovation such that the expectation is of maximized benefits with minimal risk. The next section explores concrete research problems that if solved would help to illuminate the trade-offs between control and innovation in open-ended search, and/or help to better navigate them.

Research Directions for Safe Open-Ended AI

Learning from Biological Evolution and Human Systems of Innovation

First, insight into the safety and controllability of open-ended search may be possible through non-algorithmic means, by studying human and natural examples of open-ended search, e.g. biological, technological, or cultural evolution. For example one relevant question is how qualitatively similar (i.e. predictable) are the outcomes of open-ended search. For example, Gould (1990) famously laid out the thought experiment of "replaying life's tape," arguing controversially that human-level intelligence would not be likely to arise if evolution were run again, i.e. suggesting that evolution is highly contingent. The evidence so far from natural experiments (e.g. the evolutionary isolation of Australia), convergent evolution of adaptations, and experimental evolution, is nuanced (Blount, Lenski, and Losos, 2018), as of yet providing no straight-forward conclusion.

Additionally, evidence from animal breeding and attempts to intervene in ecologies provide evidence on how challenging open-ended search can be to safely control, or its products can be usefully later adapted. For example, human breeding of wolves for docility led to human-useful and friendly dogs, and surprisingly, foxes can be bred for tameness in only 30 generations of evolution (Trut, 1999). These examples demonstrate that open-ended systems are capable of producing agents that are very responsive to post-hoc directed shaping. However, *ecological* interventions often go awry, e.g. "killer bees" resulted from an attempt to increase honey production (Winston, 1992), and cane toads released in Australia wreaked ecological havoc without impacting the problem their release was intended to mitigate (Shine, 2010). Humility about predicting impact in complex ecologies is thus a useful lesson for safety researchers.

Beyond studying biological evolution, which provides only a single example of open-endedness, studying human systems of innovation, such as science, technology, economies, and art, may also provide useful insight into safety. Evidence supports that the outcomes of such human-driven systems cannot be easily predicted or controlled (Stanley and Lehman, 2015), but efforts also exist towards responsible research, funding, and innovation (Von Schomberg, 2013), i.e. aspiring towards research that maximizes societal benefit. For example, the information security community attempts to disclose discovered software vulnerabilities responsibly (Cavusoglu, Cavusoglu, and Raghunathan, 2007), and biodefense researchers seeking defense from biological and chemical weapons must decide what science is responsible to conduct, while walking a fine line between secrecy and providing safety information to communities (Kahn, 2004). Successful systems of responsible innovation or of research funding may provide insight into the design of safe computational open-ended systems, and thus future research that synthesizes AI safety with responsible innovation may be useful, with the significant caveat that it is unclear how well such lessons will generalize to computational search.

Safe creation of AGI through open-ended search may depend on whether the agents created would have values similar to humans, or recognize and respect the moral worth of humans. Some window on that question can be provided through answering proxy questions such as how did human morality evolve (both biologically and culturally), how inevitable was it that agents arise from evolution with moral values similar to our own, and what selective pressures or interventions feasibly would make such an outcome more or less likely? Answers to such questions could come from the fields of evolutionary psychology, behavioral ecology, and evolutionary biology. Moral philosophy also bears on such questions, e.g. the truth of the view of *moral realism* is important (Sayre-McCord, 2017), wherein ethical rules can be objective truths and not only subjective opinions. Moral realism is debated, but if true, it may be more likely that search can produce agents capable of rationally converging to the same moral judgments.

Computational Study of Open-Ended Search

Computational open-ended AI (see Taylor et al., 2016b and Packard et al., 2019 for an overview of recent research foci within the field) can also be directly studied to explore its safety. Similar to questions of predictability of biological evolution, we can also study the predictability of what open-ended AI produces and how changes in its incentives or encoding affect what is discovered. There exists preliminary research into the role of chance and contingency in evolutionary algorithms (Wagenaar and Adami, 2004; Taylor and Hallam, 1998), but not from the perspective of safety or controllability, and not in the context of recent open-ended search algorithms. Methodologies from these initial studies could be adapted to explore issues of safety. The idea is to study issues of controllability in low-stakes but representative systems before such problems are critical (i.e. if a system became capable of producing AGI).

Concrete experiments include exploring path dependence and historicity in open-ended systems, for example, by gauging the effect of running search for a fixed interval, and then forking the search into independent replicate runs with different random seeds, i.e. to see how far the runs diverge from contingencies encountered after their mutual shared history (taking inspiration from work both in experimental and digital evolution; Blount, Lenski, and Losos, 2018; Wagenaar and Adami, 2004). Analyzing the raw diversity of outcomes from different runs of open-ended search also would inform the predictability of its products.

To explore controllability of an open-ended system, a meta-learning setup could be used, wherein the explicit incentives of an open-ended search are *learned* as a function of ideal objectives that are assumed by definition to be

fully specified (for initial work related to this direction, see Houthooft et al., 2018). That is, a controller could be trained that given an exact specification of an ideal objective, would output explicit incentives such that an open-ended system trained with them would produce agents that maximized the true intended objective (e.g. a controller that would direct breeding within an ALife world to achieve particular outcomes). While this would not solve the difficult specification problem of translating an experimenter's implicit ideal objective into code, it could illuminate how possible it is to steer an open-ended search's explicit incentives towards specific outcomes.

Other experiments could explore widely varying the conditions of an open-ended search, e.g. sweeping through hyperparameters, trying many different combinations of selection pressures and domain variations, to seek levers for reliably steering the high-level outcomes of an open-ended search. Previous work has explored the idea of emergent morality within artificial life (Allen, Smit, and Wallach, 2005; Danielson, 2002), which, related to the discussion above about the biological and cultural evolution of morality, may provide hints as to the necessary conditions for open-ended search to produce cooperative and friendly behavior.

Automatic Interpretability

To gain confidence in understanding the behavior of an agent resulting from open-ended search, which could help ensure safety when it is deployed, *interpretability* methods can be applied. Common interpretability methods for neural networks include dimensionality reduction, statistically attributing decisions to particular neurons, and visualizing what inputs cause specific neurons to activate (Olah et al., 2018; Nguyen, Yosinski, and Clune, 2019). However, such interpretability methods often require manual analysis and are fit to particular neural network architectures.

One aspiration of open-ended AI is to design from scratch new architectures that embody their own learning components and algorithms (just as evolution invented neurons, brains, and their learning algorithms). To reverse-engineer the human brain has been the ongoing and yet unmet aspiration of the entire field of neuroscience; thus if an open-ended search algorithm creates novel architectures of even moderate complexity, it may take inordinate human effort given current interpretability approaches to understand them.

Because the problem of interpretability is slippery and ill-defined, it is difficult to formalize as a ML problem. However, ideally researchers would develop means of understanding novel architectures *automatically*, especially for open-ended systems that may invent entirely novel architectures that are difficult to decompose. A more immediate research direction would be to apply existing interpretability techniques to agents from current open-ended search algorithms, to better understand how well such interpretability methods work in such a setting and if and how they can be better adapted to them.

Benchmarks for Safe Open-Ended Search

Finally, one common tool for catalyzing research is that of benchmarks, i.e. standardized problems in which different methodologies can be easily applied and compared. Although benchmarks can become problematic as researchers overfit their methods to them and reviewers fixate on improving scores, they also enable easily trying new approaches and can focus research.

Scalable Interactive Open-ended Search One hope for making open-ended search safer is to leverage human input, e.g. to perform selection, to change incentives on the fly, to intervene to stop a problematic agent from causing harm, or to further manually breed the products of open-ended search. Initial experiments have explored combining interactive evolution with novelty search (Woolley and Stanley, 2014), showing that human input can make the search more efficient, and similarly-inspired studies could also investigate whether such hybrids can also make the products of search safer. Additionally, even if it were designed to be safer, interactive search can be intractably expensive due to its dependence on human input; to make interactive open-ended evolution feasible at scale requires understanding what kinds of human input provide the most leverage. Concretely, one interesting research direction in scaling interaction open-ended search is to examine whether machine learning models of human preferences applied successfully for RL in directed ML (Christiano et al., 2017) could also be used to guide open-ended evolution.

Selective Discovery One safety problem in controlling open-ended search (or systems of innovation in general), is how to find desirable points in the search space without ever evaluating catastrophic ones. For example, one might want never to evaluate robot controllers that cause the robot itself to be damaged (this is related to the problem of safe exploration studied within AI safety; Saunders et al., 2018). In other words, how possible is it to avoid problematic areas of the search space, and when it is possible, how expensive is it to do so while guaranteeing a certain level of confidence of safety? One possible benchmark would be the MAP-ELITES setup of the innovation engine paper (Nguyen, Yosinski, and Clune, 2015), wherein the idea would be to design incentives (i.e. which elements of the map to favor for reproduction) such that a given desirable set of niches are optimized to a high threshold, without ever discovering high-scoring solutions for an undesirable set of niches (with both lists of niches provided to the search algorithm). E.g. it may be possible to exploit semantic relationships between niches to direct search resources effectively and safely. A simpler benchmark for novelty search would

be to discover as many novel policies as possible without ever evaluating one that crashed into a wall.

Open-Ended Reality Gap Many of the near-term risks from open-ended search likely result from unexpected failures from transfer from simulation into the real world. That is, running open-ended search in the real world is currently too expensive to be practical, and so for practical applications, agents would need to be trained in simulation and then transferred to reality. Problems incurred during transfer relate to robustness problems in AI safety, i.e. due to failures in modeling, the real world differs from the simulated one in ways that an agent ideally would be robust to.

Direct research into crossing the reality gap is inconvenient, because it requires owning a physical robot, creating a simulation of that robot, creating a physical version of the robot's simulated environment, and either manually evaluating the physical robot's performance relative to the simulated one's, or creating a system that automatically handles such evaluation. Further complicating real-world evaluation is that open-ended systems often involve many agents interacting with one another, as in many ALife worlds, thus requiring many physical robots to test; or may employ evolvable environments (as in the obstacle courses evolved by POET, or mazes evolved by MCC; Brant and Stanley, 2017), which would require setting up by hand diverse and complex real-world test scenarios. One contribution would be to create and open-source an easily reproducible transfer workflow in a domain amenable to open-ended search (e.g. using standardized robots and build components), coupled with a working open-ended search algorithm.

Another idea is to create a *proxy* for real-world transfer, such as two independent simulations with differing detail and accuracy, where the less accurate simulation would be used for training, and the more accurate simulation would serve as a proxy for real-world transfer. The advantage of such a two-simulation setup is that it would enable research to progress much more quickly and painlessly, although the disadvantage is that it is not wholly representative of real-world transfer. One simple and concrete suggestion, taking POET as a working example, would be to create a real-world proxy by modifying some of the physical constants of the obstacle course simulation used by POET, or create a more complex real-world proxy by reimplementing the obstacle course in a different and more realistic physics engine. The idea would be to test the work-flow of transferring (and potentially further adapting) POET solutions.

Discussion and Conclusion

One of the larger challenges for safety in open-ended search is the *indirectness* through which a system designer influences the products of the system. That is, rather than specifying the qualities of a single product to be optimized, the designer specifies the incentives of an overall system of innovation, and the environment or conditions in which that system unfolds. Rather than a product designer, the experimenter's role is more akin to the regulator of an economy, or an organization that decides how to allocate research funds, or the designer of a virtual universe such as a massively-multiplayer online game or a social media application. In this way, safety in open-ended search may provide a microcosm for discovering the principles behind skillfully navigating the tension between creativity and control that seems intrinsic to many processes of innovation.

Note that in this paper we offer few confident recommendations for how to ensure the safety of open-ended search, because it is a relatively unstudied and complex problem. In many cases, existing experimental evidence is not amenable to conclusive interpretations (e.g. about whether the tension between creativity and control in open-ended search can be productively resolved, or whether it is effectively hopeless to ensure that interventions within complex ecologies have desirable outcomes). Instead, this paper highlights research directions (such as pursuing automated methods of interpretability) and concrete projects (such as benchmarks for safety within open-ended search) that might catalyze further understanding and progress. We believe that the lack of straight-forward conclusions highlights the nascence of this field of study, which offers an exciting opportunity for researchers.

Indeed, while previous work has touched on general safety concerns with open-ended search (Clune, 2019; Stanley, Lehman, and Soros, 2017), this paper, to our knowledge, is the first to explore how open-ended search uniquely interacts and intersects with problems, agendas, and concepts studied in AI safety (although see also related ideas in the safety of developing nanotechnology; Jacobstein and others, 2006). We highlight that in the paradigm of open-ended search, some safety concepts take on a new light, and that solving some facets of open-ended search safety problems likely will require novel approaches. In conclusion, we hope that the initial exploration provided by this paper puts many important challenges on the radar of researchers and inspires future research into beneficial applications of open-ended search.

References

Akkaya, I.; Andrychowicz, M.; Chociej, M.; Litwin, M.; McGrew, B.; Petron, A.; Paino, A.; Plappert, M.; et al. 2019. Solving rubik's cube with a robot hand. *arXiv preprint*.

Allen, C.; Smit, I.; and Wallach, W. 2005. Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and information technology* 7(3):149–155.

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; and Mané, D. 2016. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*.

Armstrong, S. 2013. General purpose intelligence: arguing the orthogonality thesis. *Analysis and Metaphysics* (12):68–84.

Blount, Z. D.; Lenski, R. E.; and Losos, J. B. 2018. Contingency and determinism in evolution: Replaying life's tape. *Science* 362(6415):eaam5979.

Bostrom, N. 2012. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines* 22(2):71–85.

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

Cavusoglu, H.; Cavusoglu, H.; and Raghunathan, S. 2007. Efficiency of vulnerability disclosure mechanisms to disseminate vulnerability knowledge. *IEEE Transactions on Software Engineering* 33(3):171–185.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, 4299–4307.

Clune, J. 2019. AI-GAs: AI-generating algorithms, an alternate paradigm for producing general artificial intelligence. *arXiv preprint arXiv:1905.10985*.

Danielson, P. 2002. *Artificial morality: Virtuous robots for virtual games*. Routledge.

Everitt, T.; Lea, G.; and Hutter, M. 2018. Agi safety literature review. *arXiv preprint arXiv:1805.01109*.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proc. of the Intl.Conf. on Machine Learning*, 1126–1135. JMLR.

Gould, S. J. 1990. *Wonderful life: the Burgess Shale and the nature of history*. WW Norton & Company.

Guttenberg, N.; Virgo, N.; and Penn, A. 2019. On the potential for open-endedness in neural networks. *Artificial life* 25(2):145–167.

Hadfield-Menell, D.; Milli, S.; Abbeel, P.; Russell, S. J.; and Dragan, A. 2017. Inverse reward design. In *Advances in neural information processing systems*, 6765–6774.

Houthooft, R.; Chen, Y.; Isola, P.; Stadie, B.; Wolski, F.; Ho, O. J.; and Abbeel, P. 2018. Evolved policy gradients. In *Advances in Neural Information Processing Systems*, 5400–5409.

Hubinger, E.; van Merwijk, C.; Mikulik, V.; Skalse, J.; and Garrabrant, S. 2019. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*.

Jacobstein, N., et al. 2006. Foresight guidelines for responsible nanotechnology development. *Institute for Molecular Manufacturing*.

Kahn, L. H. 2004. Biodefense research: can secrecy and safety coexist? *Biosecurity and bioterrorism: biodefense strategy, practice, and science* 2(2):81–85.

Lehman, J., and Stanley, K. O. 2008. Exploiting open-endedness to solve problems through the search for novelty. In *ALIFE*, 329–336.

Lehman, J., and Stanley, K. O. 2011. Novelty search and the problem with objectives. In *Genetic programming theory and practice IX*. Springer. 37–56.

Lehman, J.; Clune, J.; Misevic, D.; Adami, C.; Altenberg, L.; Beaulieu, J.; Bentley, P. J.; Bernard, S.; Beslon, G.; Bryson, D. M.; et al. 2018. The surprising creativity of digital evolution. *arXiv preprint arXiv:1803.03453*.

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

Mitchell, M. 2009. *Complexity: A guided tour*. Oxford University Press.

Nguyen, A. M.; Yosinski, J.; and Clune, J. 2015. Innovation engines: Automated creativity and improved stochastic optimization via deep learning. In *Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation*, 959–966.

Nguyen, A.; Yosinski, J.; and Clune, J. 2019. Understanding neural networks via feature visualization: A survey. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer. 55–76.

Ofria, C., and Wilke, C. O. 2004. Avida: A software platform for research in computational evolutionary biology. *Artificial life* 10(2):191–229.

Olah, C.; Satyanarayan, A.; Johnson, I.; Carter, S.; Schubert, L.; Ye, K.; and Mordvintsev, A. 2018. The building blocks of interpretability. *Distill* 3(3):e10.

Omohundro, S. M. 2008. The basic ai drives. In *AGI*, volume 171, 483–492.

Ortega, P., and Maini, V. 2018. Building safe artificial intelligence: specification, robustness, and assurance. *DeepMind Safety Research Blog*.

Packard, N.; Bedau, M. A.; Channon, A.; Ikegami, T.; Rasmussen, S.; Stanley, K. O.; and Taylor, T. 2019. An overview of open-ended evolution: Editorial introduction to the open-ended evolution ii special issue. *Artificial life* 25(2):93–103.

Packard, N. H. 1997. A comparison of evolutionary activity in artificial evolving systems and in the biosphere. In *Fourth European conference on artificial life*, volume 4, 125. MIT Press.

Ray, T. S. 1991. An approach to the synthesis of life. *Artificial life II* 371–408.

Saunders, W.; Sastry, G.; Stuhlmueller, A.; and Evans, O. 2018. Trial without error: Towards safe reinforcement learning via human intervention. In *Proc. of AAMAS 2018*, 2067–2069.

Sayre-McCord, G. 2017. Moral realism. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2017 edition.

Shine, R. 2010. The ecological impact of invasive cane toads (bufo marinus) in australia. *The Quarterly review of biology* 85(3):253–291.

Soltoggio, A.; Bullinaria, J. A.; Mattiussi, C.; Dürr, P.; and Floreano, D. 2008. Evolutionary advantages of neuromodulated plasticity in dynamic, reward-based scenarios. In *Proceedings of the 11th international conference on artificial life (Alife XI)*, number CONF, 569–576. MIT Press.

Standish, R. K. 2003. Open-ended artificial evolution. *International Journal of Computational Intelligence and Applications* 3(02):167–175.

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

Stanley, K. O.; Lehman, J.; and Soros, L. 2017. Open-endedness: The last grand challenge you've never heard of. *O'Reilly Online*.

Taylor, T., and Hallam, J. 1998. An investigation into the role of contingency in evolution. In *Artificial life VI: Proceedings of the Sixth International Conference on Artificial Life*, volume 6, 256. MIT Press.

Taylor, J.; Yudkowsky, E.; LaVictoire, P.; and Critch, A. 2016a. Alignment for advanced machine learning systems. *Machine Intelligence Research Institute*.

Taylor, T.; Bedau, M.; Channon, A.; Ackley, D.; Banzhaf, W.; Beslon, G.; Dolson, E.; Froese, T.; Hickinbotham, S.; Ikegami, T.; et al. 2016b. Open-ended evolution: Perspectives from the oee workshop in york. *Artificial life* 22(3):408–423.

Trut, L. N. 1999. Early canid domestication: The farm-fox experiment: Foxes bred for tamability in a 40-year experiment exhibit remarkable transformations that suggest an interplay between behavioral genetics and development. *American Scientist* 87(2):160–169.

Vilalta, R., and Drissi, Y. 2002. A perspective view and survey of meta-learning. *Artificial intelligence review* 18(2):77–95.

Von Schomberg, R. 2013. A vision of responsible research and innovation. *Responsible innovation: Managing the responsible emergence of science and innovation in society* 51–74.

Wagenaar, D. A., and Adami, C. 2004. Influence of chance, history, and adaptation on digital evolution. *Artificial life* 10(2):181–190.

Wang, R.; Lehman, J.; Clune, J.; and Stanley, K. O. 2019. Poet: Open-ended coevolution of environments and their optimized solutions. In *Proceedings of the Genetic and Evolutionary Computation Conference*, GECCO '19, 142–151. New York, NY, USA: ACM.

Winston, M. L. 1992. The biology and management of africanized honey bees. *Annual review of entomology* 37(1):173–193.

Woolley, B. G., and Stanley, K. O. 2014. A novel human-computer collaboration: combining novelty search with interactive evolution. In *Proceedings of the 2014 annual conference on genetic and evolutionary computation*, 233–240.