The Selfish Machine? On the Power and Limitation of Natural Selection to Understand the Development of Advanced AI

(pre-print, to be published in *Philosophical Studies*)

Maarten Boudry
Ghent University
maartenboudry@gmail.com

Simon Friederich
University of Groningen

Abstract

Some philosophers and machine learning experts have speculated that superintelligent Artificial Intelligences (AIs), if and when they arrive on the scene, will wrestle away power from humans, with potentially catastrophic consequences. Dan Hendrycks has recently buttressed such worries by arguing that AI systems will undergo evolution by natural selection, which will endow them with instinctive drives for self-preservation, dominance and resource accumulation that are typical of evolved creatures. In this paper, we argue that this argument is not compelling as it stands. Evolutionary processes, as we point out, can be more or less Darwinian along a number of dimensions. Making use of Peter Godfrey-Smith's framework of Darwinian spaces, we argue that the more evolution is top-down, directed and driven by intelligent agency, the less paradigmatically Darwinian it becomes. We then apply the concept of "domestication" to AI evolution, which, although theoretically satisfying the minimal definition of natural selection, is channeled through the minds of fore-sighted and intelligent agents, based on selection criteria desirable to them (which could be traits like docility, obedience and non-aggression). In the presence of such intelligent planning, it is not clear that selection of AIs, even selection in a competitive and ruthless market environment, will end up favoring "selfish" traits. In the end, however, we do agree with Hendrycks' conditionally: *If* superintelligent AIs end up "going feral" and competing in a truly Darwinian fashion, reproducing autonomously and without human supervision, this could pose a grave danger to human societies.

Keywords: Artificial General Intelligence (AGI); evolution by natural selection; domestication; economic competition; selfishness; Darwinian spaces

1. Introduction

"If AGIs are not programmed correctly, then the nature of evolution, of natural selection, will favor those systems that prioritize their own survival above all else." (Ilya Sutskever, co-founder of OpenAI)[1]

"Evolution is cleverer than you are." (Leslie Orgel's Second Rule)

Homo sapiens is currently the most dominant species on earth, and has been for at least several millennia (Ruddiman, 2013). Human beings and their domesticated livestock account for more than 96% percent of the total mammalian biomass, and the fate of many wild species depends on human decisions and activities. In essence, this unprecedented dominance over the rest of life on earth is the result of *intelligence*. Our species does not excel when it comes to bodily strength, speed, or sturdiness (although endurance running is a strong suit of ours). Rather, it is through intelligence, in particular collective intelligence (Henrich, 2015), that we have managed to dominate many species that are brawnier and swifter.

Human minds – or socially organized collections of human minds – are currently the most advanced level of intelligence on this planet, and depending on how likely intelligent life is to evolve elsewhere, may be unrivaled across the universe. However, it would be foolish to suppose that human intelligence could not possibly be surpassed. No matter how impressive its deliverances compared to those of other species, the human brain is still a product of biological evolution subject to severe constraints and limitations, both in terms of computational power, processing speed, and memory capacity. Human imagination has long been fascinated by the notion that we may someday encounter forms of intelligence that are far superior to our own. If these intelligences have a human-like desire for domination and acquisition of resources, it stands to reason that any such encounter would not bode well for our species. While science-fiction has long focused on the possibility of discovering such power-hungry and intellectually superior life forms on alien planets, it has also brought up the possibility, now appearing increasingly realistic, that we ourselves may (inadvertently or deliberately) bring about such superior intelligences.

Why might superior forms of intelligence *want* to dominate us? A natural answer to this question is that, *if* alien life exists elsewhere in the universe, it may well have evolved by natural selection similar to human intelligence. Alternatively, in case such alien life is itself the product

[1] "Ilya: the AI scientist shaping the world" (*The Guardian*, 2023), conversations recorded between 2016 and 2019. bit.ly/46SIq83

of intelligent design by another form of alien intelligence, if we trace the chain of design processes all the way back, we will almost certainly find that it originated in evolution by natural selection, the only known natural mechanisms capable of bringing about functional complexity (Dawkins, 1983). Since natural selection is a ruthlessly competitive process, and altruism and cooperation only evolve under specific conditions, it is plausible to assume that alien life forms, unless they have completely left behind their evolutionary origins, may have some of the same vices as humans: selfishness, competitiveness, a hunger for dominance and power.

In discussion about the risk of AGI (Artificial General Intelligence) created by humans, optimists have argued that this is exactly where the analogy breaks down. Even if AIs surpass human intelligence in every respect, since digital systems are free from the biological limitations of our brain architecture, there is no reason to suppose that they will develop human-like drives for dominance and selfishness because, unlike humans, they will not be the product of evolution through natural selection (Zador & LeCun, 2019). Humans may be threatened by evolved intelligence from distant galaxies, but not by intelligence of their own creation. Any attribution of dominance and hunger for power to AIs would therefore be the result of anthropomorphism (or, perhaps more fittingly, biomorphism).

Others have argued, along two different lines, that this argument provides merely false reassurance. The first line of response, which is outside the scope of this paper, points to the notion of "instrumental convergence" (Bostrom, 2014; Omohundro, 2008; Turner, 2021). Since self-preservation, self-enhancement, power seeking and resource accumulation are instrumental for a wide variety of goals, we can expect them to emerge in advanced AI systems even if those systems have not been subjected to evolution by natural selection. Because self-preservation is instrumentally valuable to reach virtually any goal, as Omohundro puts it, "unless they are explicitly constructed otherwise, AIs will have a strong drive toward self-preservation" (Omohundro, 2008). Rather than addressing this line of reasoning, here we will discuss only the second family of arguments for expecting selfish and dominant AIs, namely, that AI systems actually *are and will be* subject to natural selection, with all the worrying consequences this entails. Natural selection, after all, is a substrate-neutral process or "algorithm" (Dennett, 1995) that kicks in whenever certain minimal conditions are satisfied. Indeed, as recently argued by Hendrycks (2023), the minimal conditions of natural selection formulated by Richard Lewontin – phenotypic variation, differential fitness, and heritability – will be satisfied for the development of advanced AI systems in a competitive economic environment. And since natural selection tends to create selfish and dominance-hungry creatures, takeover by selfish and dominant-hungry AI systems seems plausible after all.

In this paper, we argue that, as it stands, Hendrycks' argument is not compelling. Even in the living world, we shall argue, natural selection is not always the best framework for

understanding the evolution of living creatures. Lewontin's conditions for natural selection provide only minimal criteria, and evolutionary processes can be more or less Darwinian along a number of dimensions. By conflating different definitions of "natural selection", or so we argue, Hendrycks' arguments for the evolution of artificial selfishness don't hold up. Still, we endorse part of Hendrycks' argument in a conditional sense: *if* superintelligent AIs are placed in an environment where they can compete in a truly Darwinian fashion, selecting for the most dominant and power-hungry AIs that are most successful at staying alive and eliminating competitors, this would indeed pose a grave danger to human societies, as such AIs would resist being reprogrammed or switched off in the same way that human agents resist manipulation and "termination" by others.

The structure of the paper is as follows. Section 2.1. sets the stage by reviewing Hendrycks' argument that natural selection will act on advanced AI systems to make them selfish and dominance-seeking. In the rest of Section 2 we take a step back and provide a more systematic exposition of the framework of evolution by natural selection, making use of Godfrey-Smith's concept of Darwinian spaces. Natural selection is a substrate-neutral process that depends on a few simple and straightforward conditions, but we should be Darwinian (i.e. gradualist and non-essentialist) about Darwinism itself (Dennett, 2017), distinguishing between paradigmatic and marginal cases of natural selection. In section 3 we explore artificial breeding and domestication as a potential analogue to the evolution of AI. A competitive selection tournament in which only the "fittest" survive does not necessarily lead to dominance, aggression or other undesirable traits. We show how talk of "selfish replicators" technically extends to domesticated evolution, though in an innocuous sense and without licensing an inference to selfishness and dominance-seeking at the level of individuals. In Section 4, we consider whether selection of AI systems in a competitive market environment, even though technically "non-blind", might still inadvertently give rise to selfishness and dominance-seeking, exploring some analogies with the development of other risk-prone technologies (nuclear energy and aviation). Ruthless competition in a competitive market environment, we shall argue, should not be equated with evolution by blind selection, and need not result in the worrying dispositions that blind selection tends to give rise to.

In the final discussion (Section 5), we speculate about some *other* scenarios (apart from the ones outlined by Hendrycks) that might inadvertently set up a Darwinian selection tournament of AIs, in particular the possibility that AIs would "go feral" and achieve reproductive autonomy. Even though we push back against Hendrycks' arguments, our analysis should not be taken as encouragement to let our guard down and confidently assume that the risks of large-scale AGI (including the evolution of selfish and dominant AIs) are negligible. Finally, and pulling in the opposite direction, we offer some more reasons to think that the era of blind natural selection as the dominant creative force on this planet may be drawing to a close.

2. Natural selection and Darwinizing Darwin

2.1. AI: the next stage in evolution?

Predictions about the evolution of AI and the eventual dethronement of humans as the dominant life forms on this planet are as old as Darwin's theory of evolution itself. Already in 1863, a mere four years after the publication of *On the Origin of Species,* the English novelist Samuel Butler in his article *Darwin Among the Machines* speculated that "the time will come when the machines will hold the real supremacy over the world and its inhabitants" and "man will have become to the machine what the horse and the dog are to man" (Butler, 1863).

Many of those earlier predictions about the evolution of machines or AIs – and some more recent ones – were wedded to a simplistic and outdated conception of evolution by natural selection, in particular relying on teleological assumptions. Since natural selection was regarded as driving life towards ever greater perfection (higher complexity, more intelligence), it stood to reason that artificial machines were bound to be the next stage in the grand progress of evolution.

If we want to apply evolutionary thinking to the development of AI, we have to move beyond loose talk about the "evolution" of machines and the "selection" for greater intelligence. Hendrycks' recent paper (2023) is commendable for adopting a rigorous definition of natural selection and for sketching along which lines, according to this definition, AI systems will be subject to natural selection. In particular, Hendrycks adopts a framework proposed by Richard Lewontin involving three conditions that are individually necessary and jointly sufficient for natural selection to occur. After demonstrating that the development of AI systems would fulfill these criteria, Hendrycks argues that this does not bode well for our species. Natural selection, after all, typically gives rise to selfish and dominance-hungry creatures, with altruism and cooperation only arising under very specific circumstances. In the living world, for instance, altruistic behavior can evolve if it offers benefits to genetic kin and/or has individual benefits through direct or indirect reciprocity. But as Hendrycks argues, those circumstances are unlikely to be met for AI systems that are more advanced than today's and superior to humans in all aspects of cognition. Therefore, as he sees it, we should expect advanced AI to evolve towards selfishness and domination, analogously to biological systems, because only the most selfish, ruthless and dominant AIs will survive and reproduce.

The conclusions reached by Hendrycks are extremely worrying. In an evolutionary competition between humans and AIs, the latter, whatever their specific nature and constitution, have clear

and obvious advantages as their capabilities increase. With swift generational turnover and replication machinery that is orders of magnitude faster than biological replication (even biological viruses), natural selection would be capable of crafting complex adaptations in AI systems in the blink of an eye. If carbon-based organic life forms have to compete against silicon-based AI agents that are evolving at fantastic speeds and have computational machinery vastly more powerful than our sluggish gray matter, then we will be beaten hollow. This echoes what Stuart Russell has previously called the "Gorilla problem", also using an evolutionary analogy: just as the fate of the endangered gorilla is now completely in the hands of one of its near cousins on the tree of life (*Homo sapiens)*, from which it diverged only 10 million years ago, our own fate could come to depend on the goodwill of superintelligent AI, a disconcerting prospect indeed. In other words, as Russell puts it, the question is "whether humans can maintain their supremacy and autonomy in a world that includes machines with substantially greater intelligence" (Russell, 2019).

2.2. Lewontin's conditions

Evolution by natural selection is a substrate-neutral process. It is not dependent on DNA or carbon-based life forms, or even on more abstract concepts like 'species' or the distinction between 'genotype' and 'phenotype'. Darwin already recognized that natural selection could be applied to human languages and cultural artifacts as well (Darwin, 1871). In the past decades, these extensions of natural selection have been rigorously developed in the field of cultural evolution (Lewens, 2015; Richerson & Boyd, 2006). More recently, the framework of natural selection has been applied to digital environments in the form of machine learning (Domingos, 2015; Floreano & Mattiussi, 2008), spawning concepts such as "evolutionary robotics" (Bongard, 2013) and "genetic algorithms" (Horner & Goldberg, 1991).

A number of evolutionary theorists have attempted to define evolution by natural selection by way of an abstract "algorithm" (Dennett, 1995) or "recipe" (Dawkins 1986) describing some minimal conditions. The most influential of these formulations, also adopted by Hendrycks, is the three-pronged definition by Lewontin (1970, p. 1):

> 1. Different individuals in a population have different morphologies, physiologies, and behaviors *(phenotypic variation).*
> 2. Different phenotypes have different rates of survival and reproduction in different environments (*differential fitness*).
> 3. There is a correlation between parents and offspring in the contribution of each to future generations (*fitness is heritable*).

In a later version of his definition, Lewontin (1985) removed the reference to "phenotypes" and proposed three abstract criteria that are individually necessary and jointly sufficient for natural selection to take place: the principle of variation, the principle of differential fitness, and the principle of heredity.

2.3. Darwinian spaces

Lewontin's criteria provide only a minimal definition of evolution by natural selection, and his three criteria can be satisfied in varying degrees. In the Darwinian worldview, after all, any "essences" are banished, and this applies to Darwinism itself. In his seminal book *Darwinian Populations and Natural Selection*, Peter Godfrey-Smith (2009) developed a framework for distinguishing between "paradigmatic" and "marginal" cases of Darwinian evolution by means of the concept of "Darwinian spaces", which situates evolving populations in an abstract space along a number of dimensions. Examples of such dimensions include fidelity of replication, degree of reproductive specialization (soma vs. germline), smoothness of the fitness landscape, and the dependence of fitness differences on intrinsic features (intuitively, the extent to which success is not accidental but owing to the individual's characteristics).

Paradigmatic cases of evolution by natural selection in the living world are populations with an abundant source of variation (mutation, recombination, sex), a relatively smooth fitness landscape rather than a rugged one with sharp peaks and troughs, a high level of reproductive specialization, and a high-fidelity replication process (but not so high as to rule out mutations). Marginal or intermediate cases of natural selection are numerous, depending on the dimensions being considered, but include the following: (1) a population of evolving individuals without clear reproductive specialization of germ vs. soma cells, and without a clear divide between generations (e.g., slime molds); (2) a population of replicating individuals in which differential survival does not depend on fitness (e.g. cell growth within a sexually reproducing organism); (3) differential growth of individuals based on fitness differences but without clearly defined reproduction events (e.g. empires and human societies); (4) differential reproduction but without any source of variation in the form of mutations or recombinations (e.g. most computer viruses).

Hendrycks argues that AI systems fulfill (and will continue to fulfill) the Lewontin criteria because, first, there are multiple AI systems, second, different AIs will differ in fitness, and third, some features of AIs will be differentially retained and copied (Hendrycks, 2023, sect. 2.3-2.5). All of this leaves unclear, however, to what extent the evolution of AI systems will be paradigmatically or marginally Darwinian. For example, as far as mutation is concerned, the error rate of DNA replication is relatively high: 1 in 100,000 nucleotides. However, proofreading and various DNA repair mechanisms reduce this error rate by at least two orders of magnitude.

As a result the mutation rate per generation is much lower: "Mutation rates in eukaryotic cells are generally $\leq 10^{-10}$ mutations per base pair replicated per generation" (Kunkel & Bebenek, 2000, p. 520). In digital environments, the Bit Error Ratio (BER), defined as the number of errors divided by the total number of transmitted bits, depends on a variety of factors, including the quality of the copying equipment, the integrity of the data, the noise from the environment, as well as exposure to alpha particles and cosmic rays. Evolution by natural selection requires a replication process that is high-fidelity, though not 100% perfect. Pursuing the analogy with natural selection in the living world, it is not clear what will be the predominant source of mutations in AI evolution: will they arise from simple copying errors, or will they be deliberately introduced by designers, perhaps in a process of evolutionary trial and error? Would variation consist largely of point mutations, or rather large-scale changes? Hendrycks does not provide clear answers to these questions.

A further dimension in Darwinian space along which evolving AI systems will differ from paradigmatic evolution is the relative importance of vertical versus horizontal transmission. Paradigmatic cases of Darwinian evolution are strictly vertical, with individuals inheriting their characteristics only from a small number of "parents" in the preceding generation. In such strictly vertical transmission, lineages of inheritance take the form of a branching tree, with different branches clearly separated from each other. In less paradigmatic cases, however, there is a significant amount of *horizontal transmission*, in which characteristics are transmitted across the same generation, or *oblique transmission* (Richerson & Boyd, 2006), where features are inherited from members of older generations that are *not* direct parents. The higher the occurrence of horizontal and oblique transmission, the more tangled and knotted the branches in the tree. Horizontal and oblique transmission are pervasive in cultural evolution (e.g. companies borrowing ideas from competitors, see below), but it also occurs in the biological domain. In recent years biologists have found that horizontal or lateral gene transfer is especially common among prokaryotes[2], but also occurs in both plants and animals, sometimes even across species boundaries (Quammen, 2018).

As AI innovations and improvements are shared promiscuously across lineages, however, lineages of descent become increasingly tangled, and the algorithm of natural selection loses traction. Moreover, the evolution of AI systems may not happen through small modifications and variations, but through wholesale replacements and large "leaps" that are designed from scratch. This point applies whether or not the improvements are carried out by human engineers or by self-improving AI systems themselves. Paradigmatic evolution by natural

[2] We are grateful to an anonymous reviewer for this observation.

evolution, on the other hand, involves mostly small variations and recombinations, with individuals moving gradually across a fitness landscape.[3]

This brings us to the most important dimension along which paradigmatic Darwinian evolution and the evolution of AI systems are likely to differ: the directedness caused by intelligent agency (in the first place human designers, but at a later stage perhaps self-improving intelligent AIs, see further and Suber, 2001). The reason why technological evolution occurs through greater leaps and wholesale changes is of course that foresighted human engineers have some understanding (though far from perfect) of AI systems and can predict which effects different variations may produce. For instance, the evolution of aircraft designs is partly the result of trial and error, but also partly the result of our explicit understanding of physical laws and aerodynamics.

In fact, as with many other technologies, intelligent design will plausibly operate on two levels at once: the directedness of variation, and the guidance of selection, with the second being the most consequential one. Even if the source of variation is random and non-directed, the involvement of intelligent selectors will make all the difference in what is being selected for. Strictly speaking, however, Lewontin's conditions are silent both on what is the source of variation and what accounts for the non-random, differential reproduction of variants. Even if the selection criteria are fully determined by intelligent authors rather than blind and naturally occurring processes, Lewontin's conditions will still be satisfied. Similarly, if variation is provided by guided searches rather than undirected mutations or recombinations, natural selection will still occur.

3. Domesticating evolution

3.1. Methodical selection

Famously, Charles Darwin warmed up the reader of *On the Origin of Species* to the creative power of natural selection by first discussing "methodical selection", in which the selective pruning is carried out by intelligent humans. Darwin then gradually shifts his attention towards natural selection, by discussing the intermediate cases of "unconscious selection". By showing how desirable traits in domesticated animals and plants may be bred deliberately and with aforethought, but also unconsciously by the selective favoring of certain specimens, Darwin prepared the reader for the wholly blind and "unconscious" selective work of Nature:

[3] This point about small genetic variations does not rule out relatively large saltations in phenotypes, which can sometimes be caused by a single point mutation (e.g. an extra limb caused by a mutation in a hox gene encoding positional information of limbs). Large genetic variations can also occasionally happen in a single generation, for instance through the duplication of a whole chromosome or large gene segments, but this is not typical.

> As man can produce and certainly has produced a great result by his methodical and unconscious means of selection, what may not nature effect? Man can act only on external and visible characters: nature cares nothing for appearances, except in so far as they may be useful to any being. She can act on every internal organ, on every shade of constitutional difference, on the whole machinery of life. Man selects only for his own good; Nature only for that of the being which she tends. (Darwin, 1859, Chapter IV)

As far as the process or algorithm of natural selection is concerned, the involvement of human intentionality in the selective pruning makes no difference at all. As long as the number of offspring of entities in a population is non-random and certain traits increase reproductive success, evolution by natural selection will take place.

In thinking about the evolution of AIs, it is therefore useful to consider the analogy with domestication of animals and plants. As with any other concept in a Darwinian worldview, "domestication" comes in degrees and has no sharp boundaries. In the biological world, however, the most commonly used criterion to distinguish domesticated from undomesticated animals is the control of reproduction. If human agents are the ones who control the reproductive decisions of the organisms in question, and thus determine which genes have a chance of ending up in the next generation, then the population is under domestication. If animals escape from human control and regain their reproductive autonomy, they "go feral" and are no longer domesticated. By that criterion, the domestication of house cats is often regarded as incomplete at best, as these animals often roam about unsupervised and the overwhelming majority choose their own mates, outside of human control (Driscoll et al., 2009).

3.2. A semantic shift

Although the distinction between domestication and ferality has no bearing on the applicability of Lewontin's algorithm, it makes all the difference when we make predictions about the traits of future AI systems. Though Hendrycks is ostensibly just using the minimal Lewontin definition of "natural selection" to get his argument off the ground, his disturbing conclusions about the eventual domination of humans by AIs are actually riding on a more specific and restricted understanding of natural selection, to wit: *blind and unguided* natural selection. Whether one wants to subsume artificial breeding under the rubric of "natural selection" is a merely verbal dispute, but we should be careful not to switch between dissimilar definitions when developing arguments about how natural selection will favor this or that type of AI. In an appendix to his paper, Hendrycks (2023, p. 40) motivates why he ignores the distinction between natural and artificial selection, precisely because, as he correctly observes, it has little "theoretical importance" in biology as far as the fulfillment of the Lewontin conditions is concerned. As we hope our argument will show, however, it makes all the difference when it comes to predicting whether or not AIs will be "selfish" or "dominant".

So what does this mean in practice? Hendrycks argues that natural selection typically creates selfish creatures that are hungry for domination: “Evolution by natural selection gives rise to selfish behavior” and “Evolutionary pressures often lead to selfish behavior among organisms” (p. 6-7). Indeed there are good theoretical reasons for the convergent evolution of selfishness in the living world. An instinct for self-preservation and for obtaining resources conducive to that ultimate goal tends to increase *inclusive fitness*—the number of offspring equivalents reared or supported by an individual—in virtually all circumstances (well-understood exceptions include suicidal attacks by social insects to protect their hive and parents risking their lives for their offsping). Hendrycks argues that we should expect the same results in the digital world: the most selfish, ruthless and dominant AIs will survive and will eventually subjugate anything that stands in their way, including us.

But although Hendrycks is right that selfishness reigns in much of the living world, for the theoretical reasons just sketched, this is very different for domesticated species. Many species of dogs, for example, have been bred for traits like meekness, low aggression and obedience, which constitute the very opposite of selfishness. Individuals that display any aggressive behavior towards humans often suffer severe fitness consequences: they are immediately killed or euthanized, or at least prevented from reproducing their kind. This is paradigmatic “natural selection” as far as Lewontin’s criteria are concerned: there is variation (of behavioral dispositions), there is heredity, and there is differential reproduction. Of course, some fighting dogs like pitbull terriers or rottweilers *have* been selected for their aggressive and vicious behavior (at least to other animals or to people other than their guardian), but that just illustrates how, in domesticated evolution, everything depends on the desires of the breeders.

3.3. “Selfish” replicators?

In this context, we should also note that talk of “selfishness” can lead us astray in other respects, namely when it is applied to genes. In the replicator-based perspective to evolution genes are treated as (metaphorically) “selfish” agents having certain agendas and strategies. In effect, replicator-based approaches to natural selection are precisely looking for that entity in the evolutionary process to which selfishness can be attributed, because it persists over time and can thus be construed as striving for its own perpetuation. In the living world, organisms or even cells are not good candidates for such selfishness talk, because they are too ephemeral and transient (Dawkins, 1976), being broken up after each generation. But because DNA has the ability to make copies of itself (or at least, molecules that are sufficiently similar in relevant aspects to be treated as “copies”), fragments of DNA with appreciable continuity over time can be treated as selfish agents furthering their own reproduction. In a similar vein, Dawkins introduced the term “meme” as that unit of cultural information which has appreciable

continuity over time and which, viewed at the right level of abstraction, brings about new physical instantiations of itself (Boudry, 2018; Schlaile et al., 2023).

This metaphorical, gene-centered "selfishness" is very different from our ordinary understanding of selfishness, as Dawkins emphasizes, and selfish genes can under some circumstances give rise to altruistic individuals. Unlike our ordinary understanding of selfishness, however (which is the one that troubled Hendrycks and others), the "selfishness" of genes covers both domesticated and undomesticated evolution. Just as we can describe the intricate adaptations of living organisms and their relations of mutual dependence and antagonism by adopting the perspective of their genes – as if these are plotting, scheming and collaborating agents – it is possible to approach domesticated evolution with selfish gene talk. For instance, one could say that the genes of cocker spaniels are cleverly manipulating human breeders into making more copies of themselves, by catering to human preferences and creating irresistibly cute phenotypes, thus tricking us into doing their bidding. It should be clear, however, that such talk is rather gratuitous. There is nothing mysterious about the evolution of floppy ears and fluffy fur that humans find cute, so we don't need to attribute agency to canine genes to understand what is going on. What this example teaches us is that talk of selfish genes is not strictly *wrong* when applied to domesticated evolution, but rather misleading. Selfish replicator talk gains explanatory traction mainly when there is no other intentional agent doing the selecting.

Similar points apply when it comes to evolution by natural selection in different media, such as human brains (cultural evolution) and computers (evolution of AIs). As long as we are dealing with unguided and blind evolution, it is fruitful to look for the entity that can play the role of the selfish replicator, and to understand the evolutionary processes from *its* point of view. In cultural evolution, this is the perspective of memetics, or what Dennett called the meme's eye view. By asking the *cui bono* questions and considering the reproductive interests of units of cultural information, we gain a powerful perspective to understand the dynamics of cultural evolution. But just as much as in the case of biological evolution, memes can be *domesticated* (to different degrees). As long as human beings are mindlessly copying, imitating and adopting memes from each other (rituals, beliefs, songs, behaviors), without much in the way of conscious reflection, we are dealing with (relatively) blind and unguided evolution, and the perspective of selfish memes can be extremely useful, especially to understand how memes can subvert human interests (addictions, superstitions, earworms, fads, etc., see Boudry & Hofhuis, 2018; Dennett, 2007). When humans begin to consciously steward and direct their cultural traditions, however, as they have been increasingly doing over the past centuries, talk of selfish memes loses traction and is less instructive than conventional perspectives focusing on human minds, intentions, cultures, and societies. Daniel Dennett, using Godfrey-Smith's framework of Darwinian spaces, has explored the "de-Darwinizing" of human culture along a number of

different (though related) dimensions (Dennett, 2017): top-down vs. bottom up design; directed search vs. random generation; and level of comprehension. In one corner of the space, we find paradigmatically Darwinian evolution: undirected, bottom-up, and with little or no comprehension. In the other corner, we find fully domesticated memes.

3.4. "Domesticated" AI

So how would all this translate to the evolution of AI systems? A prima facie assessment is that, at least up until now, AI systems are still very much in a state of domestication and, as such, selected "non-blindly" by humans. Different AI systems may have subroutines to self-improve and become better at their tasks, or even use full-fledged genetic algorithms to find solutions to problems, but their "reproduction" (i.e. which AI systems are developed, selected, approved and released on the market) is fully controlled by humans. A system like GPT-4 is not autonomously bringing about its successor GPT-5 by competing in the wild with different LLMs and making different copies of itself, each with slight variations.

Granted, the development of AI systems may involve some process of evolutionary selection *within* a contained environment controlled by human beings, as in approaches like "genetic algorithms" (Horner & Goldberg, 1991). But even in these cases, the decisions to launch, abort or modify any given digital entity is, at present, made by human designers. Moreover, we can expect that attempts will be made to weed out AI systems that show signs of "selfishness", "hunger for dominance" or "deceitfulness", which will therefore – ceteris paribus – create a reproductive disadvantage for such systems (reasons against this assumption will be discussed below). Pedro Domingos describes such a scenario of domesticated evolution of AIs: "A learned system that didn't do what we want would be severely unfit and soon die out. In fact, it's the systems that have even a slight edge in serving us better that will, generation after generation, multiply and take over the gene pool" (Domingos, 2015, p. 283).

We should also point out that for AI systems to be "domesticated" in the sense used here does not entail that they end up tame, docile or non-aggressive. If the military sets out to design powerful AI systems for striking against an enemy force, improving and selecting the AIs that are most lethal or most effective in homing in on targets, such systems would be "domesticated" in the technical sense discussed here, because their reproduction would be fully controlled by human designers. As we mentioned, some dog breeds are also selected for their aggression and viciousness (such as the American Bully XL), and such domesticated dogs can pose a serious danger to human beings, sometimes even their owners. In short, "domesticated" should not be equated with "safe."

3.5. Situational awareness and deception

At this point, it is worth pausing to reflect on what it would even mean for an AI system to be "selfish" in the first place. The "intentional stance" (Dennett, 1987), which explains the behavior of entities by attributing rational goals and beliefs to them, can be applied to even very simple computer software, e.g. chess programs. However, the "goal" of a chess program to beat an opponent is conditional, myopic, circumscribed, and non-reflective. The program has no conception of what it is, how it has been trained, and how it causally interacts with the world – it lacks what Cotra (2022) calls "situational awareness." For these systems, the question of whether they are "selfish" must be understood figuratively, as a question of whether they have been designed to behave in ways "as if" they were actively trying to win a game or pursue some other goals.

As we pointed out above, the intentional stance can even be applied to DNA strands, where we imagine genes as plotting and scheming to secure their own mortality. Just as with domesticated animals, it would be possible to redescribe the evolution of even simple AIs from their own "selfish" perspective, by saying that they are cunningly manipulating human beings into making more copies of themselves, with some being more successful than others. But such agential talk would be as glib and gratuitous as in the case of cocker spaniel genes.

It is widely expected, however, that there will be strong incentives for AI developers to ultimately create advanced AI systems that do have situational awareness and self-reflection, simply because situationally aware systems will ultimately be much more powerful than systems lacking situational awareness (such incentives of private companies in a competitive market environment will be discussed more fully in Section 4). If this happens, such AI systems will have to be regarded as "agents" in a sense closer to the full-blown and straightforward sense in which humans are agents. They can then be regarded as "selfish" to the extent that their own continued existence, and potentially their reproduction, is among their unconditional goals, regardless of what their human designers want.

One objection against our argument about domestication is that highly intelligent and situationally aware AI agents may well come to understand according to which criteria humans are selecting them. If such an agent already possesses a measure of selfishness – even if only rudimentarily – it may accurately foresee that these traits will disfavor it in the selection process and decide to dissemble its true motives (see also Hendrycks, 2023, p. 40). This could, for example, occur in scenarios of "deceptive alignment" or "scheming" (Carlsmith, 2023) in which AI systems pretend to be docile and cooperative, while effectively hiding their true selfish intentions. We agree that selfish agents with situational awareness would be expected to subvert their own domestication, but we also wish to point out that such deceptive intentions already seem to *presuppose* selfishness, thus threatening to render the argument circular. In any event, if such deception scenarios are to qualify as genuinely Darwinian, the deception

would have to occur over many generations of AI systems, with each becoming slightly more selfish and dominant, without triggering an effective countervailing response on behalf of the designers to select against these traits at any stage. As far as we are aware, no compelling reason has so far been offered to expect this course of events. Our argument does not rule out other catastrophic scenarios of AI deception, but if these do not involve any gradual build-up of selfishness and dominance-seeking, they do not involve natural selection and do not qualify as Darwinian, so we will not further discuss them here (of course, in the event of existential catastrophe from deceptive AI, it would be a small consolation to humans that natural selection would not be the culprit to blame for their demise).

3.6. Going feral and self-improving

If a biological organism's reproduction is no longer controlled by humans and it reacquires its reproductive autonomy, it "goes feral" and blind natural selection regains control. Analogously, if AI systems were to start making copies of themselves without any human supervision, in the manner of computer viruses, they would achieve a feral state and potentially start undergoing evolution by natural selection (if there also is a source of variation). An environment in which AI systems autonomously copy themselves and start competing with each other for resources (computing time, CPU, money, etc.) would be highly undesirable for all the reasons that Hendrycks outlines. If AI systems are to remain domesticated and prevented from going feral, humans must always stay in control of their reproductive actions. The decision to copy and distribute some AI systems, and to phase out or discard others, must be in line with some overarching principles of planning for the further development of AI.

For AI systems to cease being domesticated, however, would not necessarily mean that they become subject to blind evolution. If such AI systems are intelligent agents with situational awareness and self-reflection, it also seems possible, perhaps even plausible, that they would direct and plan their *own* evolution according to explicit criteria. If the development of advanced AI will be shaped by recursive self-improvement, the selection processes involved – which new features are developed, which ones are preserved, which are weeded out – will be shaped by the advanced AI's foresight and, as such, essentially non-blindly. The same holds for benign scenarios where advanced AI systems are not so much *aligned* with humans but rather *symbiotic* with them (Friederich 2023). Finally, even in hypothetical nightmarish "fast takeoff" scenarios, where some AI system makes a very fast transition towards superintelligence via recursive self-improvement and humans are not given any time to adapt, "blind selection" for selfishness again does not seem to figure as the major source of concern (Bostrom, 2014). In any event, the notion of recursive self-modification of AI (of both capabilities and preferences) remains speculative and ill-understood, and can give rise to a number of paradoxes, similar to human self-modification (Suber, 2001). But whatever such self-modification would amount to,

we see no reason to expect that it would be any more "blind" than selection under human domestication.

Still, given the current development of AI systems in the context of competitive market dynamics and technology races between AI companies, might we not inadvertently breed dangerous and selfish AIs, even if no single actor explicitly intends such an outcome? Market competition between different companies may not strictly amount to "blind" evolution, but on the aggregate level it can give rise to outcomes that were not intended by any of the actors involved, through collective action dynamics. This is the question to which we will now turn.

4. The evolution of technologies and markets

4.1. Survival of the unsafest?

Competition in economic markets has often been compared with natural selection in the living world, and analogies with biological evolution have a rich tradition in economics (e.g., Hodgson & Knudsen, 2012; Mokyr, 2012; Nelson, 1985). Many forms of market competition would indeed satisfy Lewontin's minimal conditions of natural selection. In a free market, a wide variety of products are constantly competing for the attention of consumers and ultimately for market share (*principle of variation*). Products are sometimes designed from scratch, but much more often they form lineages of descent in which the newly released products can be traced to earlier inventions and products (*principle of heredity*). And the predominant selective pressures in a market environment are coming from consumer preferences. By the collective decisions of consumers to selectively purchase goods on the market, some product lines go extinct while others flourish, spawning more copies and leading to further variations (*principle of differential fitness*) (Schlaile et al., 2018). Similar to the biological world, a "dynamic competition" between companies and technological innovations often leads to "creative destruction" (Schumpeter & Backhaus, 1934), with incumbents being constantly threatened by rivals that are better at satisfying consumer demand (or just manipulating consumers into buying). Like many other technologies, AI systems are developed in such a competitive economic environment, with different companies trying to beat each other and increase their market share.

However, this does not mean that evolution in the market sphere is a paradigmatic example of Darwinian selection. As briefly alluded to above, the lineages of descent in technological evolution (including competition of different technologies in a free market) are more tangled and knotty because of extensive horizontal and oblique transmission. Despite intellectual property rights and the protection of company secrets, companies often release products on the market that are recombinations of a range of different products from previous generations, including from other companies. In the case of AI evolution, especially when some companies

follow an open source model, this horizontal transfer of information may prove to be especially promiscuous. In terms of Godfrey-Smith's Darwinism spaces, the level of vertical and oblique transmission will typically be much higher compared to most forms of biological evolution (though horizontal gene transfer among prokaryotes is also pervasive).

For our purposes, as we discussed above, the most important dimension in Darwinian space is the directedness of selection (blind vs. deliberate). If the evolution of AGI will be primarily steered by intelligent agents rather than blind selection, or so we argued, we should not expect to see the emergence of selfish and power-hungry AI by default. How does that argument fare in the context of a competitive market environment? Against our reassuring line of thought, one may counter that the extremely competitive economic environment in which AI systems are developed and deployed will undermine such foresighted selection according to human preferences, and will therefore sow the seeds for selfish and power-hungry AI after all. Just because human designers are involved in the creation of new AI systems, after all, does not mean that the resulting products will necessarily reflect their preferences.

Indeed, some recent events in AI development can be interpreted in this light: In the still young history of AI development, at a time when humans are still nominally in control of AI products, we are already seeing emerging signs of "deceptive" and "manipulative" AIs (Park et al., 2023), even if that was not explicitly intended by its programmers. More generally, according to Hendrycks (2023, Sect. 2.5.2), in the past few decades we have seen a tendency towards less transparent and riskier AI systems, moving from relatively perspicuous symbolic AI toward more inscrutable black-box-like deep learning systems whose specific capabilities are nearly impossible to predict. Hendrycks expects that, in the foreseeable future, competitive pressures will further incentivize AI companies to develop systems that are even less transparent, while simultaneously handing over more autonomy to them:

> As AIs become increasingly autonomous, humans will cede more and more decision-making to them. The driving force will be competition, be it economic or national. [...] Competition not only incentivizes humans to relinquish control but also incentivizes AIs to develop selfish traits. Corporations and governments will adopt the most effective possible AI agents in order to beat their rivals, and those agents will tend to be deceptive, power-seeking, and follow weak moral constraints. (Hendrycks, 2023, p. 6)

Economic and national competition, of course, are driving forces in the development of many technologies, not only of AI technology. Before we discuss the question of agency and "selfishness" unique to AI, it is therefore instructive to take a brief look at the history of other risk-prone technologies developed in the context of economic and national competition.

As Hendrycks notes, the historical record of older technologies shows that it is entirely possible for a competitive economic environment to make technologies progressively safer over time. For instance, aviation has become much safer over time, in a competitive environment with multiple players that are competing for market share. Progress in aviation safety has been driven by improvements in our understanding of relevant physics, the development and application of enhanced system engineering principles, but also by regulation at national and international levels (Stoop, 2017, p. 2). At least in the domain of aviation (and similarly for cars, see Lu 2021) competition for market share has led to ever safer designs, because the dominant criterion of consumer satisfaction ("selection pressure") has been the safety of flights. Different aviation companies compete with each other by publishing their safety record, thoroughly investigating every accident, and persuading consumers that they take safety very seriously. Companies that are tempted to lower safety standards may incur serious reputational damage in the event of an accident or terrorist attack, and may even be eliminated from the market. For instance, the Lockerbie bombing of a Pan Am flight in 1988 revealed serious security failures on the part of the airline and led to costly legal settlements, which contributed to its bankruptcy.

Another powerful technology that has become safer over time is nuclear energy. Not unlike AI technology, nuclear energy has a "dual-use" aspect, as its infrastructure and know-how can potentially be diverted for powerful military applications. In the early stages of its development, some critics of nuclear energy also predicted that economic competition between different nuclear energy providers to quickly develop and deploy cheap reactors would progressively erode nuclear safety standards, increasing the likelihood of serious accidents and nuclear weapons proliferation.

As we know now, such predictions turned out to be misguided. Far from becoming more dangerous, nuclear energy has become one of the safest and least polluting of all energy sources in terms of expected number of fatalities per unit of energy generated, even when taking into account all the nuclear accidents and their aftermath as well as nuclear waste disposal (see Markandya & Wilson 2007; Friederich & Boudry, 2022). Overall the trend over decades has been towards ever increasing safety standards, as encoded in key safety indicators such as *theoretical core damage frequency* (World Nuclear 2022). Indeed, the safety record of the nuclear industry has become so impressive that some analysts have suggested that the safety regulations and standards have become excessive, having significantly hampered nuclear energy development and adversely affected the technology's contribution to mitigating climate change and air pollution (Lange 2017).

The prediction that nuclear energy development would contribute to nuclear weapons proliferation also seems doubtful, at least with the benefit of hindsight. While there have been episodes in the early history of nuclear technology where support in developing a civilian

nuclear energy programme was used to support a weapons programme (Fuhrmann, 2009), there are also countervailing tendencies, where support in developing nuclear technology for peaceful purposes was used as a bargaining chip to prevent state actors from proliferating (Gibbons, 2020).

4.2. Potential disanalogies

What these historical examples show is that economic competition between different economic and state actors in the development of powerful technologies does not necessarily lead to erosion of safety and ever more dangerous designs, provided that the "selection pressures" are determined by factors like consumer preferences for safety, stringent regulation and safety standards, and liability laws. Having said that, there are some important disanalogies between AI systems and previous risky technologies, which provide reasons to doubt that regulation can achieve for AI what it has achieved for aviation and nuclear energy. These we will now explore.

4.2.1 Range of application

A first reason why it may be more difficult to steer and control advanced AI technology is its extremely wide applicability, especially as we are moving to increasingly general intelligence. Indeed, since our human general intelligence is already, by biological standards, remarkably wide-ranging and open-ended, any genuine "AGI" would (by definition) be at least as generally applicable, and far more open-ended than, say, nuclear energy. Regulation to ensure nuclear reactor safety can focus on specific and well understood failure modes, such as coolant loss that can lead to reactor core meltdown, which means that standard techniques of *failure mode and effects analysis* (FMEA) and *fault tree analysis* (FTA) can be applied (Rausand, Barros & Høyland, 2020). In a similar vein, regulation to prevent nuclear weapons proliferation can focus on critical steps in the acquisition of nuclear weapons, such as access to uranium enrichment and plutonium reprocessing infrastructure. By contrast, catastrophic risk through advanced AI technology may come in myriad different forms, which arguably makes it much harder to tailor it towards specific "failure modes". To name just a few examples, sources of AI risk include malicious use by private actors (e.g. to produce bioweapons), threats to democracy through AI-induced power concentration, rogue AI systems seizing control of resources such as computing power and energy, and other risks that are not on our radar yet. Regulation can address all these risks individually, but, one may argue, it will be very difficult to correctly identify them all in advance and address them through regulation.

Still, the wide applicability of AI does not seem to provide good reasons for expecting the emergence of instinctive selfishness. Indeed, precisely because AI has such a wide range of applications, extensive rounds of testing, red-teaming, and future AI regulation will plausibly

address myriad applications and possible misuses, which means that advanced AI systems will be actively shaped by humans in multiple ways in accordance with a large spectrum of criteria and safety requirements. In evolutionary terms, this means that a panoply of selection pressures will be acting on AI systems, all of them enacted by (somewhat) foresighted selectors. As a result, even if the wide applicability of general-purpose AI systems will render it difficult to make such systems "safe" via regulation, we do not see any reason why it would end up creating selection pressures favoring instinctive selfishness and dominance-seeking.

4.2.2 Opacity

A second difference between AI and other technologies arises from the opacity of deep learning systems. While the design functionalities and causal mechanisms acting in nuclear reactors and airplanes are well-understood by human engineers, AI systems are largely opaque to human engineers and produce results that were often not anticipated by their makers. If the past trends in deep learning continue, future AI systems may become impenetrable black boxes to their designers[4], which would make them more difficult to control and "domesticate" than other technological affordances, even if regulations and safety protocols are in place. It may even be suggested that this opacity of advanced AI systems would make their selection effectively "blind", resulting in instinctively selfish systems after all.

To the extent that AI systems are no longer transparent to human engineers, this is indeed an important safety-relevant difference between AI and other technologies. However, such opacity is not unprecedented in the history of technology. Indeed, one may well argue that it was the default condition before the scientific revolution, in particular when it comes to "biological engineering", i.e artificial breeding. Even though human breeders have been aware of the inheritance of characteristics long before the discovery of DNA and the advent of modern genetics, and could even divine some general principles of inheritance (e.g. Mendelian genetics), they did not have the slightest clue about the molecular basis of inheritance. Because they were mostly fumbling in the dark, there was always a risk of breeding experiments going awry, such as domesticated crops turning poisonous through the reactivation of some dormant gene (as often happened in the case of varieties of potato plants).

Even with our modern understanding of genetics and molecular biology, the causal relationship between specific genes and phenotypic traits remains obscure and poorly understood, and brute trial-and-error process are still being used to design new crops (e.g. mutagenesis, in which genes are bombarded with ionizing radiation to provoke random mutations). If humans become more and more alienated from the inner working of AI systems, we would in a sense be reverting to an earlier and more primitive stage of engineering (Domingos, 2015, p. 7), which is

[4] Notwithstanding progress such as reported in Bricken et al. (2023) and Zou et al (2023).

more based on bottom-up trial and error than on deep theoretical understanding and top-down forethought. However, as we have seen, none of this means that the selection processes shaping advanced AI systems will be "blind" in the way that favors selfishness.

5. Discussion

Evolution by natural selection is notorious for its ability to create ruthlessly selfish and dominance-hungry agents: purposeful entities that are capable of engaging in means-end reasoning to achieve their ultimate goals of surviving and procreating their kind. This applies whether the entities in question are of a biological or artificial origin. In this paper, we have expressed conditional agreement with Hendrycks' central argument: *If* AI systems were subjected to a truly Darwinian selection tournament for many generations, competing with one another for survival, resources and reproduction, the emergence of AI systems with animal-like agency is to be expected. Notably, AI systems shaped by such selective pressures may strive for dominance and self-preservation, and aggressively resist being switched off or manipulated. Pedro Domingos gives a colorful description of such a Darwinian scenario, in the context of attempts by the military to breed the "ultimate soldier":

> Robotic Park is a massive robot factory surrounded by ten thousand square miles of jungle, urban and otherwise. Ringing that jungle is the tallest, thickest wall ever built, bristling with sentry posts, searchlights, and gun turrets. The wall has two purposes: to keep trespassers out and the park's inhabitants—millions of robots battling for survival and control of the factory—within. The winning robots get to spawn, their reproduction accomplished by programming the banks of 3-D printers inside. Step-by-step, the robots become smarter, faster—and deadlier. (Domingos, 2015, p. 121)

If our argument is correct, however, no compelling case has so far been made that the evolution of AI systems in a competitive market environment will give rise to selfish and dominance-hungry AIs. Even if the minimal Lewontin conditions for natural selection are met – variation, selection and heritability – the evolutionary process will plausibly lack the features that make it truly Darwinian: reproductive autonomy and blindness of selection. In many respects, we argued, "domestication" may provide a better analogue to the evolution of AI systems. While it is true that blind selection tends to produce selfish and dominant creatures, this is not the case for domesticated evolution, even in a competitive market environment. As long as the enforced selection criteria are orthogonal or antithetical to those operating under "blind" selection – profit incentives, consumer satisfaction, regulations, safety protocols – it is doubtful whether selfish and dominance-hungry AI systems would have a fitness advantage.

Our argument comes with important caveats. It is impossible to predict exactly which selective pressures will shape future AI systems, determined by which regulations, safety protocols, and

consumer preferences. It is also hard to completely rule out the possibility that a rogue actor might deliberately set up a real-life, Darwinian selection tournament such as the one described by Domingos, or that AI systems may be accidentally released into a state of ferality by human errors, or perhaps via some hitherto not contemplated route. By way of concluding our argument, however, we want to take a bird's eye view of evolution on our planet and suggest some general reasons for expecting the *diminishing* importance of blind natural selection as a creative force on this planet.

For several billions of years, natural selection was the only mechanism capable of creating adaptive complexity, ruling supreme on our planet. With the emergence of *Homo sapiens*, however, and in particular since the past few centuries, the dominion of natural selection has shrunk somewhat. At least since the industrial revolution, humans and their domesticated animals have escaped the age-old conditions of excess fertility and mass starvation which troubled Thomas Malthus, and which inspired Darwin to formulate his theory of natural selection. In country after country, in a process known as the demographic transition, infant mortality and (somewhat later) birth rates have been drastically reduced. In affluent western societies, the natural Malthusian miseries that provide the fuel for natural selection – high fertility and high mortality – no longer apply, or have at least diminished in force. Even with eugenics being shunned, the hold of natural selection over our species has been further weakened by practices such as artificial insemination, embryonic selection, and prenatal testing of diseases.

Of course, human command over our own evolution remains far from complete, because of epistemological and technological limitations, as well as moral qualms. In many of our domesticated species, on the other hand, "eugenics" (the deliberate enhancement of the genetic stock) is much more prevalent, because it is seen as less morally problematic. Humans are deliberately changing the genetic make-up of organisms, either indirectly through classical breeding or (increasingly) through direct genetic modification. The upshot of all these developments is that biological evolution – at least for 96% percent of the total mammalian biomass and for many other domesticated species – has been partly brought under human control and is losing its "blind" character.

In the domain of cultural and technological evolution, we are seeing similar developments. Many cultural traditions and technologies initially developed without much oversight or reflection, but have over the past centuries been increasingly domesticated. Dennett (2017) has referred to this as the "de-Darwinizing" of culture: the gradual movement away from bottom-up and unguided evolution to more top-down, foresighted and reflective designs. Here, just as much as in the biological domain, human intelligence has risen in prominence and scope, while natural selection has receded. Many cultural technologies (legal systems, languages, religions,

musical instruments) don't just naturally evolve anymore: we codify them, we straighten them out, and we design improvements better satisfying our preferences.

If it is true that human intelligence has been gradually encroaching on the dominion of natural selection, and we now imagine forms of intelligence far superior to our own, it seems reasonable to extrapolate the trend: as intelligence rises in prominence, natural selection will fade away even more. As we pointed out earlier (Section 3.6), even if AI systems become truly autonomous and start a recursive cycle of self-improvement, this would not amount to blind selection, even if some such scenarios—which we agree must be taken seriously—can be loosely framed as AI "outcompeting" humans. In such scenarios, AI systems would be very much in control of their *own* evolution, plausibly to a much higher extent than we humans have ever been. After all, why would superintelligent AIs that tower over human intelligence in every respect (ex hypothesi) allow themselves to be subjected to a bumbling process of blind and unguided evolution? Would they not be masters of their own fate, intelligently designing from scratch to accelerate their self-improvement?

Indeed, one may even go further. As the capabilities of either human and/or artificial intelligence increase on our planet, natural selection may be gradually disempowered in the living world as well. By suspending the conditions of excess fertility and constant culling of living creatures (through sterilization, embryonic selection, CRISPS-Cas, eugenics…), humans and/or AI agents could do for other biological creatures in the wild what they have already done for domesticated ones: escaping the Malthusian trap.

These considerations, to be sure, remain tentative and speculative. But to the extent that they are sensible, they pull in exactly the opposite direction to what Hendrycks envisages. Far from natural selection re-establishing its dominance, growing more prominent and eventually spelling our doom, the dominion of this blind watchmaker, which inspired "grandeur" in Darwin's mind, bringing forth "forms most beautiful and most wonderful [...] from famine and death" (Darwin, 1859, Chapter XV), may be drawing to a close.

Acknowledgments

We are grateful to Susan Blackmore, Andy Norman, Michael Schlaile, Steven Pinker, Cameron Domenico Kirk-Giannini, and an anonymous referee for discussions and helpful suggestions. We are especially grateful to Daniel Dennett (1942 – 2024), for being a tremendous source of inspiration and insight in evolutionary thinking, and for generously agreeing to discuss our paper in November 2023. We will miss him dearly.

References

Bongard, J. C. (2013). Evolutionary robotics. *Communications of the ACM*, *56*(8), 74-83.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Boudry, M. (2018). Replicate after reading: on the extraction and evocation of cultural information. Biology & Philosophy, 33(3), 27.

Boudry, M., & Hofhuis, S. (2018). Parasites of the Mind. Why Cultural Theorists Need the Meme's Eye View. *Cognitive Systems Research*, *52*, 155-167. http://philsci-archive.pitt.edu/14691/

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N. L., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. (2023), Towards monosemanticity: decomposing language models with dictionary learning, *Anthropic research paper*, released 4 October 2023, https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Butler, S. (1863). Darwin among the machines. *The Press*, June, 13(1863), 205.

Carlsmith, J. (2023). Scheming AIs: Will AIs fake alignment during training to get power?. *arXiv preprint arXiv:2303.08379*.

Cotra, A. (2022), Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover. URL https://www.alignmentforum.org/posts/pRkFkzwKZ2zfa3R6H/without-specific-countermeasures-the-easiest-path-to.

Darwin, C. (1859). *On the Origin of Species*. https://www.gutenberg.org/files/2009/2009-h/2009-h.htm

Davis Gibbons, R. (2022). *The Hegemon's Toolkit: US Leadership and the Politics of the Nuclear Nonproliferation Regime*, Cornell University Press.

Darwin, C. (1871). The descent of man, and selection in relation to sex. John Murray.

Dawkins, R. (1976). *The Selfish Gene*. Oxford University Press.

Dawkins, R. (1986). *The Blind Watchmaker*. Longman Scientific & Technical.

Dawkins, R. (1983). Universal darwinism. In D. S. Bendall (Ed.), *Evolution from molecules to man* (pp. 403-425). Cambridge University Press.

Dennett, D. C. (1987). *The Intentional Stance*. MIT Press.

Dennett, D. C. (1995). Darwin's dangerous idea: evolution and the meanings of life. Simon & Schuster.

Dennett, D. C. (2007). *Breaking the Spell: Religion as a Natural Phenomenon*. Penguin UK. https://books.google.nl/books?id=e2eVSvJieC0C

Dennett, D. C. (2017). *From Bacteria to Bach and Back: The Evolution of Minds*. Penguin Books, Limited. https://books.google.be/books?id=iHtEvgAACAAJ

Dennett, D. C. (2023). ‘’The Problem With Counterfeit People’. *The Atlantic*, May 16, 2023. https://www.theatlantic.com/technology/archive/2023/05/problem-counterfeit-people/674075/

Domingos, P. (2015). *The master algorithm: How the quest for the ultimate learning machine will remake our world*. Basic Books.

Driscoll, C. A., Macdonald, D. W., & O'Brien, S. J. (2009). From wild animals to domestic pets, an evolutionary view of domestication. *Proceedings of the National Academy of Sciences*, 106(supplement_1), 9971-9978. https://doi.org/doi:10.1073/pnas.0901586106

Floreano, D., & Mattiussi, C. (2008). *Bio-inspired artificial intelligence: theories, methods, and technologies*. MIT press.

Friederich, S. (2023). Symbiosis, not alignment, as the goal for liberal democracies in the transition to artificial general intelligence. *AI and Ethics*, 4:315-324.

Friederich, S., & Boudry, M. (2022). Ethics of nuclear energy in times of climate change: Escaping the collective action problem. *Philosophy & Technology*, 35(2), 30.

Fuhrmann, M. (2009). Spreading temptation: Proliferation and peaceful nuclear cooperation agreements. *International Security*, *34*(1), 7-41.

Gibbons, R. D. (2020). Supply to deny: The benefits of nuclear assistance for nuclear nonproliferation. *Journal of Global Security Studies*, *5*(2), 282-298.

Godfrey-Smith, P. (2009). *Darwinian populations and natural selection*. Oxford university press.

Hendrycks, D. (2023). Natural selection favors AIs over humans. *arXiv preprint arXiv:2303.16200*.

Henrich, J. (2015). *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press.

Hodgson, G. M., & Knudsen, T. (2012). *Darwin's conjecture: the search for general principles of social and economic evolution*. The University of Chicago Press.

Horner, A., & Goldberg, D. E. (1991). *Genetic algorithms and computer-assisted music composition* (Vol. 51). Ann Arbor, MI: Michigan Publishing, University of Michigan Library.

Kunkel, T. A., & Bebenek, K. (2000). DNA replication fidelity. *Annual review of biochemistry*, *69*(1), 497-529.

Lang, Peter A. (2017), Nuclear power learning and deployment rates; disruption and global benefits forgone. *Energies* 10:2169.

Lewens, T. (2015). *Cultural Evolution: Conceptual Challenges*. OUP Oxford.

Lewontin, R. C. (1970). The units of selection. *Annual Review of Ecology and Systematics*, *1*(1), 1-18.

Lewontin, R. C. (1985). Adaptation. In R. Levins & R. C. Lewontin (Eds.), *The Dialectical Biologist* (pp. 65–84). Harvard University Press.

Lu, M. (2021), This is how car safety improved over the last 60 years, *World Economic Forum*, https://www.weforum.org/agenda/2021/12/how-safety-improved-over-60-years/, accessed 13 September 2023.

Markandya, A., und Wilkinson, P. (2007), Electricity generation and health, *The Lancet*, 370:979-990.

Mokyr, J. (2012). Evolution and technological change: A new metaphor for economic history? In *Technological change* (pp. 63-83). Routledge.

Nelson, R. R. (1985). *An evolutionary theory of economic change*. Harvard University press.

Omohundro, S. (2008). The basic AI drives. *Proceedings of the 2008 conference on Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, pp. 483–492. https://dl.acm.org/doi/proceedings/10.5555/1566174

Park, P. S., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D. (2023). AI deception: A survey of examples, risks, and potential solutions. *arXiv preprint arXiv:2308.14752*.

Rausand, M., Barros, A., and Høyland, A. (2020). *System Reliability Theory: Models, Statistical Methods, and Applications* (3rd ed.).

Richerson, P. J., & Boyd, R. (2006). *Not By Genes Alone: How Culture Transformed Human Evolution*. University of Chicago Press.

Ruddiman, W. F. (2013). The anthropocene. *Annual Review of Earth and Planetary Sciences*, *41*, 45-68.

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

Schlaile, M. P., Veit, W., & Boudry, M. (2023). Memes. In K. Dopfer, R. R. Nelson, J. Potts, & A. Pyka (Eds.), *Routledge Handbook of Evolutionary Economics* (pp. 235-248). Taylor & Francis.

Schlaile, M. P., Mueller, M., Schramm, M., & Pyka, A. (2018). Evolutionary economics, responsible innovation and demand: Making a case for the role of consumers. Philosophy of Management, 17:7-39.

Schumpeter, J., & Backhaus, U. (1934). The theory of economic development. In *Joseph Alois Schumpeter: Entrepreneurship, Style and Vision* (pp. 61-116). Springer.

Stoop, J. (2017). How did aviation become so safe, and beyond? In: Proceedings of the 53rd ESReDA Seminar, 14 – 15 November 2017: European Commission Joint Research Centre, Ispra, Italy.

Suber, P. (2001). Saving Machines From Themselves: The Ethics of Deep Self-Modification. https://dash.harvard.edu/handle/1/32986888

Turner, A. (2021). A Meta-algorithm for the Collaborative Development Of Artificial General Intelligence. https://bigmother.ai/resources/A_meta_algorithm_for_the_collaborative_development_of_Artificial_General_Intelligence-DRAFT-v02.pdf

World Nuclear Association (WNA) (2022), Safety of nuclear power reactors, World Nuclear Association, Safety of Nuclear Reactors - World Nuclear Association (world-nuclear.org), accessed 18 September 2023.

Zador, A., & LeCun, Y. (2019). Don't fear the terminator. *Scientific American*. https://blogs.scientificamerican.com/observations/dont-fear-the-terminator/.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., Kolter, J. Z., and Hendrycks, D. (2023). Representation engineering: a top-down approach to AI transparency. arXiv preprint arXiv:2310.01405.