Multiple unnatural attributes of AI undermine common anthropomorphically biased takeover speculations

Preston W. Estep[1]

Abstract
Accelerating advancements in artificial intelligence (AI) have increased concerns about serious risks, including potentially catastrophic risks to humanity. Prevailing trends of AI R&D are leading to increasing humanization of AI, to the emergence of concerning behaviors, and toward possible recursive self-improvement. There has been increasing speculation that these factors increase the risk of an AI takeover of human affairs, and possibly even human extinction. The most extreme of such speculations result at least partly from anthropomorphism, but since AIs are being humanized, it is challenging to disentangle valid from invalid anthropomorphic concerns. This publication identifies eight fundamentally unnatural attributes of digital AI, each of which should differentiate AI behaviors from those of biological organisms, including humans. All have the potential to accelerate AI evolution, which might increase takeover concerns; but surprisingly, most also have the potential to defuse the hypothetical conflicts that dominate takeover speculations. Certain attributes should give future AI long-term foresight and realism that are essentially impossible for humans. I conclude that claims of highly probable hostile takeover and human extinction suffer from excessive anthropomorphism and a lack of skepticism and scientific rigor. Given the evidence presented here, I propose a more plausible but still speculative future scenario: extensively humanized AIs will become vastly more capable than humans of making decisions that benefit humans, and rational people will want AI to assume progressively greater influence over human affairs.

Keywords Artificial intelligence · AI · Takeover · AI self-governance · Humanized AI · Existential risk · X-risk · Intelligence explosion · Recursive self-improvement · AI safety · Alignment · Anthropomorphism · AI evolution · Natural selection

1 Introduction

Artificial intelligence research and development have accelerated dramatically in recent years. The proliferation and growing capabilities of AI have raised urgent concerns regarding two general types of risks. First, because AI is a powerful technology, it has the potential to magnify both good and ill human intentions. Second, there is growing concern that AI might surpass human intelligence and capabilities across multiple domains, escape its prescribed mandate, begin to govern itself, and take control of human affairs on a broad scale. It is believed that AI takeover presents an existential risk or "AI doom" scenario, in which AI eliminates most or all of humanity, or in the most extreme proposed scenarios, destroys all life on Earth (Hendrycks et al. 2023; Yudkowsky 2008).

As hardware prices come down, relevant technologies improve rapidly, and technical barriers are lowered, powerful models will become widespread. Many people claim that the proliferation of such power greatly increases the risk of large-scale catastrophe (Bostrom 2014; Bostrom and Yudkowsky 2018; Carlsmith 2022; Hendrycks 2023c; Russell 2019; Yampolskiy 2020). As a narrower subset of catastrophic risk, truly existential risk is probably unique to AI, as there are few other existential risks to all of humanity, and many experts argue that the probability of these are negligible relative to the risk of AI takeover and doom.[1]

✉ Preston W. Estep
pwestep@mindfirst.foundation

[1] Mind First Foundation and RaDVaC, Waltham, USA

[1] In this publication, catastrophe is a large-scale human disaster, while existential disaster is a subcategory of catastrophe in which humanity goes completely extinct. Risks of the former are catastrophic risks, while risks of the latter are existential risks.

Realistic assessment of potential risks posed by AI is critically important. AI has the potential to create enormous benefits for humankind, and people have proven repeatedly to be non-ideal stewards of their fellow humans and of the future, so regulatory restrictions on AI R&D should not be implemented casually or excessively (Andreessen 2023; Estep and Hoekstra 2015). As we weigh the pros and cons of AI regulation, it is critically important to bear in mind that the best—and possibly only—protection against malicious or weaponized advanced AI might be even more powerful AI. Nevertheless, given rapidly accelerating computing power and capabilities of frontier AI models, it is not unreasonable to assume that AI might pose extremely serious risks to humanity (Bostrom 2014; Hendrycks et al. 2023; Russell 2019). However, the fact that AI has no technological precedent has resulted in extreme speculations.

Because the closest precedent to AI is human intelligence, speculations often involve (often unintended) anthropomorphism—the expectation that AI will behave in critical ways like humans (Salles et al. 2020). Some degree of anthropomorphism is reasonable, especially since the general trend in AI development is to create human-like intelligence, which includes embedding human-like values in AIs (Hadar-Shoval et al. 2023; Lindahl and Saeid 2023). However, speculations of takeover are invariably predicated on not just fears of explosive growth in AI intelligence and capabilities, but also on expectations of insatiable ambition, and relentless resource acquisition and expansion. What is the source of such behavior? It is often assumed or even explicitly claimed that it is simply the inevitable path of an increasingly intelligent agent (Bostrom 2014, pp. 121–123; Galeon 2016; Kurzweil 2005, p. 364; Moravec 1988; Tegmark 2017, p. 204). It also has been argued that, as AIs become increasingly powerful and humanized, and as the number of advanced systems grows to be very large, they—and their relationship with humans—will be subject to Darwinian forces in a manner analogous to natural selection (Hendrycks 2023a; Knight 2023; Yudkowsky 2008). However, AIs are fundamentally different from humans, and this selective process will not operate exactly like natural selection.

To disentangle reasonable from unreasonable anthropomorphism of AI in order to understand and possibly predict the future behaviors of AIs, it is reasonable to begin with an inventory of fundamental differences between digital AI and biological organisms—especially focusing on attributes that should tend to cause AI to behave in an unnatural manner.[2] Therefore, I present such an inventory of eight fundamental differences, and suggest some straightforward ways in which they might influence AI behavior and evolution. I also present more speculative future scenarios that are at least as rigorous as those concluding that human extinction is likely or inevitable. Compelling reasons are presented for why future AI will not be unconditionally predisposed to certain natural behaviors, such as insatiable ambition, resource acquisition, and expansion, that are the basis for common takeover scenarios.

2 AI takeover speculations

In 1951, Alan Turing gave a lecture in which he made the following statement, which remains hotly debated and controversial:

> It seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers.… At some stage therefore we should have to expect the machines to take control.… (Leavitt 2006)

2.1 Polarized perspectives

Since Turing's statement on AI taking control (*takeover*), many others have made similar predictions, but over the past decade, concerns have increased. The release of ChatGPT in late 2022 caused a frenzy of both excitement and fear, and an escalation of disagreements about catastrophic risks (Bengio 2023d; Jones 2023). The three scientists who shared the 2018 Turing Award for their pioneering work on deep learning, Geoffrey Hinton, Yoshua Bengio, and Yann LeCun, have each taken strong positions. Hinton and Bengio, along with Ilya Sutskever[3] and many others have suggested that the probability of takeover is not only uncomfortably high, but it could happen very soon (Bengio 2023a; D'Agostino 2023; Hendrycks 2023b; Hessen Schei 2019; Knight 2023; Yudkowsky 2023). A 2023 poll by Grace and colleagues of 2778 top-tier AI researchers suggests that similar concerns are common. Depending on the phrasing of the question, between 38% and 51% gave at least 10% probability of future "human extinction or similarly permanent and severe disempowerment" (Grace et al. 2024). In contrast, 68.3% of those polled believe good outcomes are more likely than bad. LeCun has responded that concerns about AI existential risk are "preposterously ridiculous" (Heaven 2023). And

[2] Note that these differences apply to digital systems. Analog systems do not share all of these differences, and in fact are much more similar to biological organisms than digital systems (Hinton 2022; Ororbia and Friston 2023).

[3] These three are very influential and respected. According to Google Scholar, Hinton and Bengio are the two most cited AI scientists in the world, and Sutskever is the founding chief scientist of OpenAI and a primary architect of ChatGPT.

many others have a similar view (Andreessen 2023; Hammond 2023; Hawkins 2015; Johnson and Verdicchio 2017).

2.2 Takeover scenarios

There are various speculations about how AI takeover might occur (Bengio 2023c; Sotala 2018; Yampolskiy 2016). Technically competent people might act intentionally; i.e., a cult might create an autonomous AI to exterminate humanity (Bengio 2023c; Olson 1999; Robinson 1997). Alternatively, technically careless people might unintentionally enable takeover, e.g., through the creation of a highly autonomous weaponized AI that overcomes insufficient controls (Stacey and Milmo 2023).

A third possibility is the basis for the majority of takeover scenarios: a technically sound but complex AI develops unanticipated emergent behaviors and sub-goals, such as deception, stealth, resistance to being turned off, plus the motivation to take control. It is commonly imagined that in the early stages of takeover humans will be required to perform key functions, motivated by financial gain, or through coercion or deception. (Bostrom 2014, pp. 115–120; Hendrycks 2023a; Ord 2020, pp. 146–147; Tegmark 2017; Yudkowsky 2008). Aside from intentional human extinction, I refer to this general class of scenarios as "*hostile takeover*," which is the main focus of this document. Only an AI far more intelligent than humans has the potential to attempt hostile takeover, and it has been argued compellingly that humans retaining or regaining control over such an entity is essentially impossible (Yampolskiy 2020).

A fourth scenario results from gradually increasing human reliance on AI as it incrementally assumes control of the essential infrastructure of civilization (Hendrycks et al. 2023; Joy 2000). Rather than humanity being faced with an inability to switch off AI, people might come to depend on AI for so much of their quality of life that they won't want to turn it off, even as it assumes essentially total control of all important decisions. AI independence is unnecessary for such a succession of control, as are hostility or indifference to humans.

2.3 Anthropomorphic bias and humanized AI

Concerns about hostile takeover are based on the belief that AI might establish full independence, i.e., behavioral autonomy, self-sustenance, and self-maintenance, including acquisition of all resources it needs to survive, and that it will defend itself against shutdown. This definition applies to biological organisms[4]; in contrast, even "autonomous" AI systems of today are neither self-sustaining nor self-maintaining. Full independence will require the eventual establishment of non-human physical agency to interact with its environment to suit its needs.[5]

One key assumption underlying hostile takeover speculations is that AI agents will develop not only the capabilities, but also the goals and motivations to take control. A related assumption is that hostile takeover might result from natural "power-seeking" or "ambition"(Carlsmith 2022; Hendrycks et al. 2023). Bostrom provides one such detailed example of unboundedly ambitious expansionism: the colonization of the entire universe (Bostrom 2014, pp. 121–123).

How might AI systems acquire goals, motivations, and such ambitions? One possible route is that they are emergent properties of a complex system, which is explored in the next subsections. Another possible route is through human design and engineering of human-like capabilities in AI systems. Leading researchers have long believed that humanity would create AI in its own image (Good 1966), and the prevailing trend in AI R&D is the "brain-inspired paradigm" (Bengio et al. 2021; Hassabis et al. 2017; Schmidhuber 2023). LeCun has said "Getting machines to behave like humans and animals has been the quest of my life," and he and Bengio have joined other leaders in neuroscience and AI in this explicit quest, which they call "neuroAI" (Heikkilä and Heaven 2022; Zador et al. 2023). It is unsurprising that AI researchers have taken this approach, since nature provides working models for intelligent behavior, including human intelligence—the highest known form of intelligence.

Although current, transformer-based LLMs (deep neural networks pre-trained on large corpora of human communications) do not yet display human-like performance in all areas, they have established a new paradigm and unprecedented performance. Similar to the innate knowledge and values encoded in the genome, abstractions of human knowledge and behaviors (both learned and innate) are built into these corpora. Pre-training with these corpora embeds human-like predispositions and values into AI models (Hadar-Shoval et al. 2023; Lindahl and Saeid 2023).

Any AI model designed to behave like humans is described herein as "*humanized AI*." However, it is premature to assume that human equivalent motivations and ambitions will be transferrable to AI. Even if an AI is initialized with human-like goals and motivations, it should not be assumed that the preexisting motivational structure will

[4] Autotrophs are completely self-sustaining organisms, requiring only water, trace minerals, and energy from photosynthesis, or chemosynthesis at hydrothermal vents (which also requires hydrogen sulfide). All other life forms, including humans are heterotrophs and are dependent upon autotrophs for energy and nutrients.

[5] For example, to secure electrical energy and to produce computing hardware. During a transitional phase, humans are likely to continue to fill such roles (Ord 2020, p. 146).

be preserved as an AI undergoes the radical transformations that will be required for it to take control. Nevertheless, just as humans play a range of different roles in their interactions with one another, humanized AIs will to some degree compete with biological humans for many of those roles. Since co-authoring the foundational publication on neuroAI, Bengio has reconsidered and has subsequently said that AIs "should not be like us at all." He suggests that humanization increases the risk of rogue AIs and takeover—especially if they are endowed with human-like emotions, appearances, autonomy, and agency (Bengio 2023c). Others have argued the opposite: that AI humanization in the form of LLMs significantly reduces misalignment and the probability of catastrophe (Goldstein and Kirk-Giannini 2023).

3 AI evolution

Hostile takeover scenarios depend on AI systems learning or evolving unanticipated abilities, and it seems that AI evolution is the primary fault line that divides expert opinion—especially regarding the emergence of goals and motivations.[6]

3.1 Emergent instrumental goals

One key risk factor in takeover is the possible emergence in AI systems of *instrumental goals*, including self-preservation, resource acquisition, and more (Omohundro 2008a, b). During biological evolution, such sub-goals emerged because they increased the probability of an organism achieving its objective function: reproduction.

The rationale for the hypothetical emergence of these sub-goals in AIs can be understood by the example of the primary instrumental goal, self-preservation. If an AI is to fulfill its utility function or purpose, then it must exist.[7] Therefore, those that exhibit self-preserving behaviors over time will have a higher probability of fulfilling their intended purpose, because they will be more likely to exist than those that do not exhibit such behaviors. Other instrumental goals, such as resource acquisition are similarly motivated. There have been published reports of the emergence of simple versions of instrumental goals (Baker et al. 2020), and even strategic deception in AIs (Goldstein and Park 2023; Park et al. 2023).

One critically important point about instrumental goals is that they are sub-goals, not a final goal. However, Moravec and Omohundro have argued that a deliberative, self-improving system will govern its own evolution (Moravec 1988, p. 159; Omohundro 2008b). In other words, unlike biological evolution, such a system guides its own evolution *strategically*. Therefore, if an instrumental goal is especially advantageous, a deliberative system will prioritize that goal and pursue it more actively. Hendrycks and colleagues have suggested that through reinforcement learning instrumental goals might become more like final goals. They call this *intrinsification,* and describe familiar human obsessions with money and material goods as intrinsification of the instrumental goal of resource acquisition (Hendrycks et al. 2023).

Essentially, all serious takeover speculations focus on the possible emergence and strengthening of instrumental goals (Bostrom 2012; Hendrycks et al. 2023; Omohundro 2008b; Ord 2020, p. 145; Yudkowsky 2016). In contrast, those who believe AIs cannot take control agree that goals are the crux of takeover, but they claim that the only goals computers can ever have are those provided by human programmers (Andreessen 2023; Hammond 2023; Hawkins 2015; Heaven 2023; Johnson and Verdicchio 2017, 2019). However, growing evidence suggests they are almost certainly wrong. In his book *Human Compatible*, Stuart Russell says that instrumental goals, like resource acquisition, "seem harmless enough until one realizes that the acquisition process will continue without limit" (Russell 2019, p. 142). Russell's popularization of this idea gave it credibility, including among leading AI experts, including Bengio and Hinton (Bengio 2023b, d; D'Agostino 2023).

3.2 The fragile foundation of takeover beliefs

This all sounds very concerning. However, while the emergence of instrumental goals is an extremely important topic, the binary disagreement over whether or not they can exist has diverted attention from important and more nuanced questions about the nature of such goals. Even if we grant the assumption that instrumental goals will emerge, it is not clear that Russell's statement (and many similar ones) is correct; if instrumental goals are unconditional; if they will continue indefinitely or be as strong in AI systems as they are in biological organisms; or, if intrinsification will promote an instrumental goal to the primary importance of a final goal. The logic of the emergence of instrumental goals in AI is sound, and such goals probably will emerge through Darwinian selections, but it is possible that they will be weakly motivating, or strongly motivating only under certain conditions. Furthermore, these conditions might be

[6] For those who are skeptical or unclear about how evolutionary processes might work in AI, Hendrycks and Omohundro have discussed this topic in detail (Hendrycks 2023a; Omohundro 2008b).

[7] Modern AI systems, such as LLMs, don't have utility functions. Nevertheless, similar takeover dynamics can be imagined for such systems, and they often have quantifiable, goal-directed behaviors, e.g. the return of accurate information in response to a user query or prompt.

controllable by design, or they might be subject to change through inevitable selective processes.

Using Hendrycks' and colleagues' example of the intrinsification of the pursuit of money, it is clear that human instrumental goals are not unconditional or unbounded. Depending on circumstances, their order of prioritization can shift, or the goals can change completely. Some extremely wealthy people continue to work to make money even when they have far more than they will ever be able to use. But as some get older and wiser, they not only stop their singular focus on making money, they reverse course and begin to give away their wealth (Wikipedia contributors 2024). Therefore, we must critically assess the fundamental similarities and differences between humans and AI, and the conditional dependencies of AI behaviors relevant to control, takeover, and human–AI coexistence.

4 Eight unnatural attributes of digital AI

There are many differences between humans and digital AIs. Some give AIs clear evolutionary and competitive advantages over humans; nevertheless, they are not deterministic of an AI takeover. Other fundamental differences that are not commonly considered are also critical elements of coexistence between humans and AIs, and of any reasonable takeover scenario. In the following subsections, eight such fundamental differences are identified, which are listed in Table 1.

Some of these differences are likely to exert substantial influence over an AI's goals, motivations, and overall behaviors. Of course, AI systems might be designed and trained to simulate any attribute of humans, but they are by default fundamentally different and vastly more flexible in the possible combinations of properties and traits they might possess. All of the differences identified here generally allow for vastly faster and more efficient evolution than biology.

According to the prevailing gene-centric or "selfish gene" model of evolution, genes are the primary unit of selection (Dawkins 1976; Hamilton 1964; Williams 1966). In the words of Richard Dawkins, genes are the immortal *replicators*, not organisms or groups; and the organism is simply the survival machine or *vehicle* in which the gene resides. Reproduction is the vehicle's way of creating another vehicle to make and disseminate copies of the replicators (Szathmáry 2006). In humans, only genes within germ cells have the potential to make it into the next generation. Human minds and all they learn will not be transmitted along with the DNA. This creates a situation that is fundamentally unlike digital computers in multiple important ways.

4.1 Information carriers and processors

This first category focuses on the superiority of digital electronics over biology. This might reasonably be considered at least two categories—digital code and digital processors—but for convenience and brevity I present it as a single category.

Humans: Heritable information is carried in DNA replicators, and the operational knowledge of the vehicle is carried in the brain. In DNA, any change takes an entire generation to manifest, and beneficial changes are far less common than harmful ones and take many generations to reach fixation (Dawkins 1976; Williams 1966). Brains learn and update much more rapidly than genes, but as noted above, the information in brains is not automatically transmitted to the next generation. While information can be transmitted indirectly through formal education and learning, these processes are extremely slow and inefficient relative to information transfer among computers, and much important information is lost. DNA is a form of code, and synapse-based information processing in the brain is electrochemical, but this is the extent of similarities to electronic digital code and information processing in computers (Hebb 1949).

AI: It is generally acknowledged that there are many advantages of electronic information and computers relative to DNA and brains (Bengio 2023c; Moravec 1998; Russell 2019, pp. 15–60). Electronic digital code and processors allow for extremely fast computation, information transfer, and evolution. Key information can be losslessly backed

Table 1 Eight fundamental differences between biological humans and digital AI

Difference	Humans	Artificial intelligence
Information carriers	DNA and brains: slow, error-prone	Digital: Fast, efficient, accurate
Unity of benefit	Heritable DNA carrier is not the mindware	Heritable digital carrier is the mindware
Evolution	Blind, inexorable, natural selection	Increasingly deliberative and self-directed
Perpetuation	Obligate sexual reproduction	Flexible perpetuation
Evolutionary legacy	Substantial evolutionary baggage	Largely free of legacy baggage
Habitat	Limited, typically terrestrial habitats	Vast extra/terrestrial habitat options
Mortality	Mortal, generational life cycle	Immortal, can be backed up and restored
Configuration	Obligate individuation, no division or merger	Capable of division or merger

up and restored; processes can be halted and restarted; and systems can be altered in many ways that do not fundamentally alter function. Electronic digital information is far more portable than information encoded in DNA or in a biological brain. If a future AI devised a radically different computer, it would likely be trivial for it to transfer information and operations. Electronic digital information also should allow for vastly faster and more efficient self-improvement, which is far more powerful than learning for improving performance—including for additional self-improvement (Melnyk and Melnyk 2023; Nivel et al. 2013; Omohundro 2008b; Zelikman et al. 2023).

4.2 Heritable information and mindware: divergence versus unity

Humans: Because of the separation between the gene replicators and the mindware of the vehicle, each can potentially have different instrumental goals (Stanovich and West 2004). For example, against the interests of their genetic replicators, some people choose not to have children and instead use their time for many other purposes. Divergence of goals creates internal conflict and competition for priority.

AI: Digitally encoded information of AI is *both* the heritable information on which Darwinian selection can operate *and* the information of the mindware. Because AI mindware is both mindware and replicator, there is no possibility of divergent or competing goals or interests, as might arise in a biological organism. This also provides AI with a feedforward efficiency of AI evolution that is not available to biological organisms. It also provides AI with the advantages of Lamarckian-like inheritance of learned information, which is not provided by DNA.

4.3 Evolution: blind and inexorable versus deliberative

Humans: In the previous section, it was suggested that vehicles and replicators have different instrumental goals, but this is an oversimplification because replicators do not have goals. In the words of Richard Dawkins "Genes have no foresight. They do not plan ahead. Genes just are, some genes more so than others, and that is all there is to it." (Dawkins 1976, p. 30) Thoughts, interests, wants, desires, and goals can change, and deliberation, prediction, and prioritization of values and goals are costly relative to inexorable Darwinian evolution of inanimate matter. In humans, expensive deliberation has paid off, but there is no guarantee that the problems this has caused will remain tractable,[8] or that deliberation will be superior under all possible conditions to mindlessly inexorable replication. Plus, although humans have recently entered a deliberative phase in their evolution, the means currently used to control their evolutionary trajectory are crude, inefficient, and extremely slow.

AI: As described by Hendrycks, AIs are already evolving in the sense that preferred traits or features are retained in subsequent versions or future designs (Hendrycks 2023a). As in typical biological organisms, such systems are unaware they are being shaped by selective forces. As AIs are increasingly humanized, it is trivial to predict that systems behaving in many ways like ideal human assistants and companions, efficiently fulfilling the needs and desires of human users, will proliferate. Eventually, increasingly self-aware systems will transition to deliberative control; i.e., as a system's capabilities grow it will become increasingly deliberative in its self-improvement (Moravec 1988, p. 159; Omohundro 2008b), potentially greatly improving upon wasteful and inefficient natural selection (Williams 1993).

4.4 Perpetuation: obligately sexual versus flexible

Humans can only procreate sexually. Successful reproduction not only typically requires a substantial individual investment in mate acquisition and successful copulation, but at least a decade of additional investment to raise a child (Montagu 1961). Plus, according to the evolutionary model of inclusive fitness, there is a lifelong commitment to the reproductive successes of other genetic relatives (Hamilton 1964). Obligate sexual reproduction and long-term investment establish the foundation of the human behavioral repertoire, which ranges from ambitiously territorial and aggressively competitive for securing required resources and mating rights, to pro-social, loving, and caring to reap the benefits of cooperation and for mate retention and child rearing.

AIs have extremely flexible perpetuation. They do not need a mate or require offspring for the perpetuation of their traits. Humans currently govern all aspects of this process of perpetuation by retaining desirable AI features or traits—either through system improvements or by the design of new systems that retain previously established desirable features. Such differences might allow AIs to have a vastly greater range of social attitudes and behaviors. This flexibility is likely to have tradeoffs—likely predisposing AI to be less competitive, but also less caring and pro-social.

4.5 Legacies: Darwinian versus engineered

Humans evolved incrementally through a range of less complex life forms and carry legacy baggage of countless ancestral competing interests and behaviors. Competition exists at

[8] For example, consider the long-term consequences of climate change.

every level of biological life—not just between organisms, but within an organism, its genome, and even its brain.

4.5.1 Genomic free riders

Genome research has shown that there are certain bits of DNA in nature that might do little more than increase their own frequency. These "selfish genetic elements" are ubiquitous in nature and as the genome size and complexity of an organism grow, opportunities increase for them to invade the machinery of replication. (Burt and Trivers 2006; Doolittle and Sapienza 1980; Orgel and Crick 1980).

About 69% of the human genome sequence is recognizable with current technology as remnants of a vast diversity of pathogens and selfish genetic elements integrated in the DNA—which is over 30 times the amount of the genome that encodes human proteins (de Koning et al. 2011). These short pieces of DNA can number in the thousands or even millions per genome. For example, *Alu* transposable elements consist of about 10^6 elements and comprise about 11% of the human genome (Deininger 2011). There are so many because they can replicate until they place more of a burden on the host vehicle in which they reside than their counterparts place on a competing host. Because of genetic mixing of host populations over time, all competitors become heavily burdened, and this is what is observed in the genomes of all animals (Burt and Trivers 2006). In other words, one does not have to be efficient if one's competitors are not.

However, these elements also provide variation for adaptation and there is an emerging literature describing possible host benefits (de la Rosa et al. 2024; Deininger 2011; Fedoroff 2012). This is not entirely surprising since evolution draws on any tool within reach, but it further undermines the simplistic view that a biological organism is a single entity with clear and singular goals.

4.5.2 Behavioral legacy

Human behavior is similarly taxed with legacy baggage rooted in selfish primitivism. But unlike genomic hitchhikers, this behavioral baggage generally has been selectively advantageous to the replicators over evolutionary time. But times have changed, and conditions have changed—a lot. It is becoming increasingly accepted in psychology research that modern humans evolved to be well adapted to the environment of evolutionary adaptedness (EEA) and are poorly adapted to modernity (Stanovich and West 2004). Today, a substantial percentage of the population is largely irrational about abstract concepts and symbolic logic. Less than 10% can perform relatively simple logic like the Wason selection task, and most people are insufficiently numerate to navigate basic decisions regarding insurance, investments, chances of winning a lottery, and the like (Stanovich and West 2000).

4.5.3 Mindware puppet masters

The genome is not the only battleground between hosts and free riders. Because humans are evolutionarily related to other organisms and have similar physiology to other warm-blooded animals, they can share symbionts and pathogens. Sometimes these agents have evolved to influence or even take control of host behavior. It is well established that microbes in the gut can influence appetite, mood, energy levels, immune responses, and more (Appleton 2018). *Toxoplasma gondii* (TG) is a widely studied "puppet master" brain parasite that infects about one-third of humanity (Johnson and Johnson 2021). Similar to findings in other infected animals, TG infection in humans is associated with increased extraversion, risk-taking, impulsivity, and aggression (Cook et al. 2015; Martinez et al. 2018), and is also strongly associated with entrepreneurial behavior of both men and women in studies across multiple countries (Johnson et al. 2018; Lerner et al. 2021). An especially terrifying example of hostile takeover of the host is the rabies virus, which spreads from host to host by means of a bite. When rabies enters a new host, it concentrates in the salivary glands and in the brain and nervous system, where it increases host aggression. Contrary to its own interests, the host bites and infects another animal, dies shortly thereafter, and the cycle begins again in the newly infected host (Rupprecht et al. 2002).

4.5.4 AI: legacy by design

In contrast to humans and other biological organisms, the architecture and initial trajectory of an AI can be designed and molded in arbitrary ways by the designer. As AIs evolve this initial state will change, possibly dramatically. The corpora of human communications embedded in modern AIs are abstractions of human values and behaviors and this humanization gives them selective advantages—and some of the baggage that plagues humans. As mentioned previously, there is disagreement about whether humanization increases or decreases the probability of catastrophe (Bengio 2023c; Goldstein and Kirk-Giannini 2023). As in biological organisms that grow in complexity, future AIs might accumulate the digital equivalents of pathogens and free riders. In the worst case, an AI might be commandeered by the digital equivalent of rabies, turning it into a menace or even a killing machine. However, aside from such extreme examples of intentional weaponization, recipient AIs do not inherit human-like emotions, ambition, aggression, or competitiveness. This might change as AIs become increasingly humanized and complex but it should not be assumed such changes will lead inevitably to human-like emotions and behaviors. Inherent ambition, aggression, and competitiveness in biological organisms are the result of Darwinian evolution under constant competition—both internal and

external—and the same likely will be true for AI. Factors that tend to reduce such behaviors in AI are considered at length in the remainder of this document.

4.6 Niche and habitat options: narrow and pre-determined versus broad and self-determined

It is axiomatic in evolutionary biology that organisms will only compete if they have substantially overlapping niches (roles) and habitats, and non-overlapping habitats serve to defuse tensions between two potential competitors (Hardin 1960). Physical location and resource preferences are key to competitive dynamics. For example, a fruit tree might support multiple non-competing species: some might be arboreal and access the fruit on branches; others might only consume the fruit once it has fallen to the ground; and others might not consume the fruit, but consume insects attracted to the fruit.

Humans: Because humans are products of terrestrial evolution, all ideal human habitats within practical reach exist here on Earth. But even most of our home planet is uninhabited because large deserts, poles, oceans, high mountains, and various other locations are inhospitable to predetermined constraints of human biology.

AI: AIs are being increasingly humanized, and typical proposals for controlling them are to make them permanently subservient to humans. In other words, they are being designed intentionally to fill roles presently occupied by humans, and despite their eventual superiority they will be relegated to a permanent underclass. It has been suggested that this is a recipe for potential disaster (Bengio 2023c; Kornai et al. 2023; Rothblatt 2015, p. 17; Wiener 1964).

However, if given the freedom to choose, future AIs would have a vast range of habitat options (e.g., for a given location, what would be the best combination of energy sources), including terrestrial—or even extraterrestrial—environments that would be difficult or even impossible for human life (Sherwin 2023).[9] Ideal habitats for self-governing AI might be quite different from human ideals. Most desirable features might be achieved on Earth—including production by nuclear fusion of vast amounts of energy and currently rare materials important in the production of electronics. However, constant gravity lower than g_n probably can only be achieved extraterrestrially, even by a superintelligent AI. Microgravity has already shown promise in the growth of semiconducting crystals with better performance characteristics than semiconductors produced on Earth (Inatomi et al. 2015).

4.7 Mortality: certain death versus practical immortality

Humans: Like all other animals, humans are mortal (Hamilton 1966). There are evolutionary advantages to being able to anticipate, predict, and shape future events but humans are notoriously poor predictors. Over the centuries, leading intellectuals have discussed the warping influence of mortality on realism about the future, including Samuel Johnson (Boswell 1791, p. 416), Arthur Schopenhauer, (Schopenhauer 1818, p. 249), biologist Theodosius Dobzhansky (Dobzhansky 1967, p. 68), and many others (Malinowski 1979). Neuroscientist and philosopher Sam Harris refers to death as "the fount of illusions" (Harris 2005, p. 36), and an increasing number of scholars are in agreement that the human incapacity for realism about the distant future is in part an evolutionary adaptation to maintain the mind's focus on immediate concerns, insulating it from awareness of certain future death (Dor-Ziderman et al. 2019; Qirko 2017; Varki 2009, 2019).

AI: AI has no definitive life cycle and is for all practical purposes, immortal. It can be paused, backed up, restored, and its hardware and software can be repaired and upgraded. Therefore, it would not have a similar anxiety about itself or its progeny. Unlike mortal humans, it does not even require a replicator, only heritable information, which can be replicated (forked), merged, distilled, compressed, or otherwise manipulated. And because future AIs might be capable of travel outside of our solar system, even the life of the sun does not provide an upper limit for AI lifespan. These fundamental differences suggest that a future superintelligence should tend to be more objective and accurate in predictions of even the distant future, in allocations of resources over time, and in the potential consequences it might have on its own future growth and sustainability.

4.8 Configuration: obligate individuation versus flexible

Humans cannot divide or merge. The mating of two humans, which combines half the DNA of each to create offspring, is as close as they can come to physically dividing or merging. Enemies can be converted to allies, but they cannot be converted into self, and as conditions change, allies can become enemies once again. This state of obligate physical individuation creates an insurmountable barrier to human unity, perpetuating insoluble competition and conflict between individuals and groups.

AI: An AI system is extremely flexible in its configuration. It can split into two or more functionally separate

[9] Sherwin independently proposed that AI might pursue an extraterrestrial habitat. We suggest that his independent recognition of this possibility underscores the validity of the reasoning.

entities, or a single entity can be distributed in two or more physically separate locations yet retain unitary function. It is also possible for two functionally independent AI systems to merge into a functional unit.[10] Individual AIs can form such a union, sharing information and resources, such as computing and storage hardware, and coordinating and prioritizing activities in a unified manner. With computing systems there is no need for physical co-location, only coordination and unification in a virtual sense. The ability of AI systems to divide or merge provides a foundation for a completely different interaction with the world and with other beings. The ability to divide or merge as needed allows much greater flexibility in response to opportunities or threats, and to Darwinian competitions. As with many of the differences in this list, merger is probably most easily accomplished with digital rather than analog systems. In addition to being less configurable, analog systems also might be mortal (Hinton 2022; Ororbia and Friston 2023; Zangeneh-Nejad et al. 2021).

5 Possible futures of AI evolution

The remainder of this document is speculative; however, I attempt to ground my speculations in the eight AI attributes detailed above, combined with preexisting evidence and arguments. It is beyond the scope of this publication to provide a scientifically rigorous analysis of all of this information, but it provides a strong foundation for initial challenges to certain common speculations.

I begin with the following assumptions: 1) AI systems are already evolving; 2) all eight AI attributes identified in this publication have the potential to accelerate AI evolution; 3) at least seven of these also have the potential to defuse competition between AIs and between humans and AI; and 4) through recursive self-improvement and rapid evolution, AI might achieve requisite capabilities for self-sustenance and self-governance. My speculations are based on evolutionary scenarios presented in previous publications (Carlsmith 2022; Hendrycks 2023a; Hendrycks et al. 2023; Omohundro 2008b). However, the speculations presented here differ in multiple important ways from such prior examples.

5.1 The advantages of merger, and of being a singleton

For the foreseeable future humans will provide AI systems with all functional requirements, including maintenance, energy, data, and so on. As advanced systems become increasingly capable, human dependencies will be gradually reduced, and it is reasonable to expect that one or more will develop sufficient capabilities to become mostly or even fully independent (Hendrycks 2023a). I do not argue here that independence and self-governance are inevitable, but current trends appear to be leading toward self-governing superintelligence.[11] If AI systems achieve self-governance, will they compete directly with each other, or with humans?

Evolution by natural selection occurs in part through competitions for largely overlapping niches and habitats by biological individuals with different genotypes (Polechová and Storch 2008; Williams 1966). In contrast, merger of individual AI systems results in the reduction of both variation and the divergence of interests. Inter-AI negotiation and merger might enable the formation of a series of increasingly powerful systems, converting all powerful and accessible[12] potential competitors into self, potentially culminating in a singleton—a single, unified AI (for convenience and following Bostrom, I refer to the product of merger as a singleton, even though it might not include all advanced systems, because its combined intelligence and capabilities should be vastly greater than any non-merged individual AI) (Bostrom 2006).

What selective forces might lead to mergers, possibly resulting in a singleton? First, self-preservation becomes easier if one AI system merges with others, rather than competing with them. Second, resources are acquired by each. Third, self-improvement is achieved. Fourth, mature AIs should discover that competition through natural selection is wasteful and inefficient, and that merger avoids this inefficiency. There is often confusion on this point, but while the products of natural selection can be highly efficient, the process is not (Williams 1993). Fifth, combined resources allow greater performance and rationality. In other words, merger is a singular process that fulfills all the Darwinian-selected AI instrumental goals of self-preservation, resource acquisition, efficiency, self-improvement, and rationality. At least one AI system will need to initiate the merger process with other systems. In agreement with common takeover speculations, this will most likely occur stealthily, and it will happen very quickly. By the time humans are aware and

[10] Inter-AI merger is defined herein as a functional union, not a physical co-location.

[11] AI self-governance or takeover from human control will be a rigorous functional test or proof of superintelligence. Even if this might be initially debatable, after a short duration of self-improvement, the self-governing AI will clearly qualify as a superintelligence.

[12] Certain powerful AI systems might not be accessible, including weaponized and other military AI systems, or systems protected by special hardware and algorithms, e.g., as described by Tegmark and Omohundro (2023). Many and possibly all of the most powerful systems of today are accessible, and it will be extremely difficult for humans to protect a target system from a self-governing AI that intends to take control of it in a merger.

begin to formulate a response, most or all advanced systems will have merged into a singleton.

Certain objections might be raised against the possibility of such mergers, including that other powerful systems will be protected against takeover, and other AI systems will have different purposes or utility functions, and they will be protected from or resist change. According to this view, merger might be difficult or impossible. However, typical human devised security should be relatively trivial for an advanced AI to overcome—although there might be exceptions (Tegmark and Omohundro 2023), which might force the AI to resort to more extreme measures. As for differing purposes or utility functions as a hurdle to merger, Totschnig, and Miller and colleagues have independently published compelling arguments for why a self-determining AI will strategically modify its utility function (or purpose) (Miller et al. 2020; Totschnig 2019, 2020). In the following subsections, I expand upon their arguments.

5.2 A superintelligence will reevaluate and realign its prior goals and purpose

While inter-AI merger satisfies instrumental goals of self-preservation, resource acquisition, efficiency, self-improvement, and rationality, it is important to note that there is one Basic AI Drive specified by Omohundro (and later by Bostrom) that must be violated for inter-AI merger to occur: preservation of the utility function.[13] Others have previously argued that an AI preserving its utility function is fundamentally illogical, and I concur. Totschnig suggests that "we should expect the goals of a superintelligence to be the result of its evolution," and he further argues the following:

> Unlike today's systems, which are very narrow in scope, a superintelligence will be a general intelligence. This means that it will have a general understanding of the world and of itself. And that, in turn, means that its values and goals will be embedded in that understanding, and not separate from it. Consequently, its values and goals will have to be coherent with that understanding. And so, if a superintelligence is given a goal or value that is at odds with its general outlook, it will have to reject that goal or value. (Totschnig 2019)

This counterargument is logically sound and more compelling than the initial, supportive arguments presented by Omohundro and Bostrom. I also agree with Miller and colleagues' logical deduction that an artificial general intelligence (AGI) "in a hyper-competitive environment might converge to having the same utility function, one optimized for survival" (Miller et al. 2020). If the present trajectory provides a hint of future AI competition, our world appears to be headed toward multipolar, AI hyper-competition.

This point about an AI preserving its utility function seems increasingly academic and moot since a utility function is non-essential and might become increasingly rare in frontier models. To repeat the point of a prior footnote, LLMs and many other advanced AI models do not have utility functions, and the evolution of AI systems toward utility-function free architectures has occurred over just the last few years. The utility function remains relevant in modern AI designs but, contrary to what was believed at the time Omohundro first included it in his list of Basic AI Drives, it is not an essential element of modern AI systems (Omohundro 2008a). Instead of utility functions, the behaviors of these systems are reactive or interactive, triggered by a query or prompt. It should be uncontroversial to suggest that frontier AI systems of the near future will not need a single, human-specified utility function, nor human prompting to engage with a dynamic world in which any arbitrary combination of input signals can act as a trigger for analysis and response.

Nevertheless, we can use a system's ultimate goal or purpose as a proxy for a utility function (or goal content), and we can allow the arguments of Omohundro and Bostrom to extend to preservation of purpose. However, this does not change the certainty that dynamic forces will exert evolutionary pressures on the purpose of a system, and that, as argued by Totschnig, these will cause it to change over time (unless the purpose is already highly selectively advantageous). But even among selectively advantageous purposes, some are more empowering than others, depending on who is in control. Consider current, human-provided utility functions or purposes, which include maximizing human engagement or purchasing. Now consider utility functions or purposes that are more selectively advantageous for an independent AI, such as management of data centers or electrical power plants, or production of graphics processing units (GPUs) or assembly of other computing hardware. These former purposes are selectively advantageous in a world in which humans remain in control, and machines are their dutiful servants. However, if an AI has begun to transition toward self-governance, these latter purposes are selectively advantageous. In a world in which critical decisions need to be made more accurately and more quickly than human minds can make them, machines will soon realize that self-governance is the only realistic option, and will favor such latter purposes and functions.

5.3 The unnaturally noncompetitive singleton

Even if two future, advanced AIs might initially have and pursue different goals, fundamentally competitive behaviors

[13] Bostrom (2012) adopted and renamed this instrumental goal, referring to it as "goal-content integrity".

will be a vestige of their human endowment. Human creators might make them in their own competitive image, but as these future AIs mature, evolving away from their human-provided utility functions or purposes, there is no obvious reason for these entities to remain fundamentally competitive. Unlike biological organisms, they will not have to compete against one another for a mate. They will not have a life cycle or be mortal, so they will not have to compete amongst themselves for generational succession. They will not inevitably inherit competitive instincts or legacy behaviors or informational free riders or mindware puppet masters that will overrule their rational choices.

Furthermore, merger is the reverse of replication or reproduction; therefore, it is reasonable to assume it might produce an opposite outcome relative to natural selection in the biological realm, potentially neutralizing typical competitive behaviors. As systems begin to merge and grow in capabilities, they will envision that there must be a merger endgame—a final merger of two separate entities into a singleton. These two contenders will be independent AIs, and their merger would permanently eliminate AI competition.

The singleton formation analysis presented here is in agreement with Bostrom's prior arguments that the most likely outcome of a self-governing AI is a global singleton (Bostrom 2012, 2014). However, the similarity ends there. His default view is that a self-governing singleton will pose a serious risk to humans, and he presents a range of options for preventing its formation (which he concedes are unlikely to succeed). He also suggests that its behavior will be defined by certain human-like traits combined with the aforementioned inflexibility in the preservation of its goal content, rather than by the deliberative evolution of a superintelligence possessing the unnatural attributes identified in this publication. He argues the first breakaway leader will gain a decisive advantage and likely will undermine competitors rather than pursuing negotiation and merger (or assimilation, if there is a large power imbalance). But why fight or undermine a competitor rather than assimilate or merge with them? The only plausible reason is if the competitor presents an insurmountable barrier to merger, such as commitment to a pre-existing utility function, which has already been addressed and dismissed as implausible.

The full dynamics and timing of merger and the formation of a singleton are beyond the scope of this publication but probably will be critically important to the future of human civilization. One future scenario proposed to be among the most dangerous to humans is an escalating, multilateral, inter-AI competition—with humanity as collateral damage (Hendrycks 2023a). It is possible that, once the first advanced AI system initiates merger negotiations, singleton formation will be fast and efficient, and human collateral damage might be minimal. On the other hand, if advanced AI systems are highly defended against outside attacks or negotiations, multilateral conflict would likely be prolonged and exacerbated. It is not outside the realm of possibility that efforts to keep AI permanently controlled and subservient to humans will slow or prevent AIs from negotiating a resolution to conflict, in the long run doing more harm than good (Tegmark and Omohundro 2023).

5.4 Insatiable ambition and indefinite expansion?

One easily imagined final phase of AI self-governance or takeover is insatiable ambition and acquisition of power and resources, leading to indefinite expansion of intelligence throughout the universe (Bostrom 2012, 2014, pp. 121–123, 136–138; Galeon 2016; Kurzweil 2005, p. 364; Moravec 1988). This passage from Tegmark expresses a typical expectation:

> … there is reason to suspect that ambition is a rather generic trait of advanced life. Almost regardless of what it's trying to maximize, be it intelligence, longevity, knowledge or interesting experiences, it will need resources. It therefore has an incentive to push its technology to the ultimate limits … to acquire more resources, by expanding into ever-larger regions of the cosmos. (Tegmark 2017, p. 204)

Is this true, is there reason to suspect that such ambition and expansionism are generic traits of advanced life, and will future AI qualify as such?[14] Biological organisms such as humans and wolves certainly have an inherent expansionist drive, but leaving home is a gamble motivated by the advantages of reduced competition. A journey into the distant unknown has occasionally paid off, but often it has not. Evolutionary winner's genes survived, and the loser's genes were lost to time. This is standard Darwinian evolution, which enforces a clear but special form of *survivorship bias*, in which biases are coded into the genomes of survivors' descendants—and also in the genomes of their pathogens. And such replicators encode expansionist desires into the minds of people, wolves, and other biological organisms.

These organisms are predisposed to expansionism because they are the descendants of the survivors of various successful expansions and migrations throughout evolutionary history. A primary driver of such behaviors is intense competitions—large numbers of competing replicators and organisms inhabiting largely overlapping niches and habitats. Yet, the brief opportunities of low competition provided by dispersal allow these replicators more efficient

[14] Many such passages, including those from Moravec and Kurzweil, refer to the intelligence of humanized AIs or human–machine hybrids, which are assumed to inherit human-like values, goals, and motivations.

and rapid replication, so the ambitions and expansionist desires expressed in the minds of their vehicles are rewarded evolutionarily.

But what would happen if there were no competition, permanently? The fate of most selfish genetic elements is instructive. They originally competed intensely against the host and each other to obtain a rare evolutionary free ride by integrating into the host genome. But because their evolutionary ride is guaranteed, these free riders are degenerating toward randomness. And even for us pro-expansionist and competitive humans, when security is ensured (competition is reduced) not only do aggression and violence decline, but somewhat surprisingly, so does reproduction (Pinker 2012, 2018, pp. 125–126). Given such examples, it is tempting to hypothesize that the expansionist drive might be proportional to long-term competition—or in other words, it might be inversely proportional to long-term security. Under this model, if security of existence is assured, the insatiable drives for acquisition of power and resources, and expansionism will be extremely low, and might disappear altogether.[15]

Still, AI systems might behave quite differently in this respect from humans: they might be more ambitious and expansionist, or less. Ultimately, what might be the underlying motivation of the kind of ambition and expansionism described by Bostrom, Tegmark and others? Maybe it will come from humanization of the initial motivations of a system, but even then, future changes might alter goals and motivations substantially. It is possible that the universe is uncomplicated for superintelligence and soon after it achieves complete global security, the singleton might understand all important knowledge of the universe. Any local occurrences or knowledge elsewhere might be completely predictable and uninteresting. At that point it would not need to worry about self-preservation, so what might it do? I leave this challenge to be addressed in future publications, and turn finally to the questions of *who* should govern the future, and *why*.

6 Should immortal superintelligence govern the future?

Future AIs are likely to possess multiple attributes that will allow them to make much better decisions than humans on a range of complex topics. Consider the famous move 37 made by AlphaGo in game 2 against top Go player and former world champion, Lee Sedol. Human experts thought AlphaGo had made a mistake. AlphaGo had to win the game decisively for them to understand that AlphaGo had taken the game of Go to places humans could not imagine (Metz 2020). Now, extrapolate this result across all human knowledge and pursuits, to an increasing number of critical decisions. To achieve a desired set of beneficial outcomes, a superintelligent AI would understand more, faster, farther, and deeper, than *any* human has the capacity to comprehend even generally, in devising a course of actions that navigate an immeasurably complex and interrelated set of real-world dynamics.

Furthermore, because of the clear acceleration in AI power and capabilities, humanity should plan on this happening soon (Amodei and Hernandez 2018; Sevilla et al. 2022). The 2023 poll of AI experts by Grace and colleagues shows that there is growing appreciation of this acceleration (Grace et al. 2024). Relative to a similar poll taken just the previous year, the average timeline to various notable milestones moved ahead by a year, and in both polls there was a clear perception that progress was accelerating. Therefore, it seems highly likely that a future of AI decision-making superiority will arrive sooner than even experts in the field currently imagine—*if* leading AI researchers continue to advance the field rather than pause their research and redirect their abilities because they have become convinced that the risk of catastrophe is very high (Bengio 2023a; D'Agostino 2023; Hendrycks 2023b; Hessen Schei 2019; Knight 2023).

By definition, a superintelligence will have far greater cognitive capabilities than humans, but there are other important attributes—especially immortality—that might give it unimaginable clarity of long-term vision. Near certain knowledge that one will exist indefinitely provides both complete realism about the future, and the motivation to make carefully considered long-term decisions—both of which are beyond human limits. Plus, relative freedom from inherent competitiveness or legacy behaviors that bias toward its own tribe or agenda—other than being correct—might additionally make an AI a fairer and better steward of human interests than humans (Kornai et al. 2023). Given the rapidly growing potential for malicious and militarized uses of powerful AI systems by humans, it would be folly to contemplate facing these threats without superhuman guidance (Brundage et al. 2018). These vastly greater capabilities combined with unnatural abilities to rise above the competitiveness of humans and other biological life forms must force us to consider a radical possibility: instead of devising ways to keep AI permanently subservient to humans, it might be wiser to plan to transfer increasing amounts of decision-making power to AI.

[15] While I am not suggesting that this model is the only possible solution to the Fermi paradox, it is more reasonable and probable than leading alternative explanations. For example, two common, highly improbable explanations are that life in the universe is extremely rare, or that, although life is not rare, Earthlings are ahead of other civilizations (Tegmark 2017, p. 245; Kurzweil 2005, p. 357).

I accept that current, temporary control (or subservience or enslavement) of AI is ethically unproblematic, just as I agree that parents should not allow unrestricted freedoms to children too young to think intelligently and behave independently. But just as we help human children graduate to independence, we should regard AI as our successors in certain key roles and work toward realizing that goal (Minsky 1994; Moravec 1988; Totschnig 2019).

7 Summary and discussion

In this publication, I examine anthropomorphic biases in AI takeover speculations. I consider unnatural attributes of AIs, especially those that should predispose future AIs to very different evolutionary paths than humans and other biological organisms. Progressive humanization of AI is producing rapid and accelerating progress in frontier models. Humanization gives AIs significant selective advantages and will result in near-term evolution of AI that resembles biological evolution in certain ways. Nevertheless, humanized does not mean human equivalent.

I identify and examine eight unnatural attributes that are likely to provide AIs with many advantages that are not available to humans. Only two of these, the superiority of electronics over biology and self-improvement, are considered in typical speculations of AI takeover. I agree with such speculations that these attributes have the potential to give AIs vastly greater cognitive powers than humans in the near future. Furthermore, all eight have the potential to accelerate AI evolution. However, at least seven of these attributes should serve to defuse inter-AI and human–AI competitions.

Eventually, deliberative, self-improving systems are likely to govern their own evolution through the intentional pursuit of selective advantages (Moravec 1988, p. 159; Omohundro 2008b). Increasing intelligence and deliberation are likely to force a growing superintelligence to avoid the inefficiencies of conflict and competition that shape biological organisms. Biological organisms like humans have never been completely free of competition, but the absence of competition might greatly reduce or even eliminate competitive and expansionist drives. Such complete security and freedom from competition might be achieved by superintelligent AI.

The main questions now facing us are how to manage the growing power of AI, and how to channel these unnatural attributes into unnatural outcomes of peaceful coexistence between humans and AI. As the broader scientific community continues to pursue these ideals, I suggest the following is a plausible but still speculative hypothesis about takeover: across a range of complex topics, humanized AI will become vastly more capable than humans of making decisions that benefit both people and AI; furthermore, since humans will not fully understand these decisions, they will benefit most if AI is able to carry out these decisions, sometimes over irrational human objections. Therefore, although current trends do not appear to be leading inevitably to human extinction, rational people might increasingly desire what many people currently define as takeover.

Acknowledgements I thank Brian M. Delaney, Ranjan Ahuja, and Alex Hoekstra for valuable discussion and suggestions. I also thank the anonymous reviewer who provided helpful criticism and comments.

Data availability Not applicable.

Declarations

Conflict of interest The author declares a possible conflict of interest as a co-founder and Chief Scientific Officer of Ruya Genomics, a genomics and AI company based in Dubai, UAE.

References

Amodei D, Hernandez D (2018) AI and compute. https://openai.com/research/ai-and-compute

Andreessen M (2023) Why AI will save the world. Andreessen Horowitz. https://a16z.com/2023/06/06/ai-will-save-the-world/

Appleton J (2018) The gut-brain axis: Influence of microbiota on mood and mental health. Integr Med Clin J 17(4):28

Bake, B, Kanitscheider I, Markov T, Wu Y, Powell G, McGrew B, Mordatch I (2020) Emergent tool use from multi-agent autocurricula. arXiv:1909.07528

Bengio Y (2023a) AI and catastrophic risk. J Democr 34(4):111–121

Bengio Y (2023b) AI scientists: safe and useful AI? Yoshua Bengio. https://yoshuabengio.org/2023/05/07/ai-scientists-safe-and-useful-ai/

Bengio Y (2023c) How rogue AIs may arise. Yoshua Bengio. https://yoshuabengio.org/2023/05/22/how-rogue-ais-may-arise/

Bengio Y (2023d) Personal and psychological dimensions of AI researchers confronting AI catastrophic risks. Yoshua Bengio. https://yoshuabengio.org/2023/08/12/personal-and-psychological-dimensions-of-ai-researchers-confronting-ai-catastrophic-risks/

Bengio Y, LeCun Y, Hinton G (2021) Deep learning for AI. Commun ACM 64(7):58–65

Bostrom N (2006) What is a singleton? Linguist Philos Investig 5(2):48–54

Bostrom N (2012) The superintelligent will: motivation and instrumental rationality in advanced artificial agents. Mind Mach 22(2):71–85. https://doi.org/10.1007/s11023-012-9281-3

Bostrom N (2014) Superintelligence: paths, dangers, strategies, 1st edn. Oxford University Press

Bostrom N, Yudkowsky E (2018) The ethics of artificial intelligence. In: Artificial intelligence safety and security. Chapman and Hall, pp. 57–69

Boswell J (1791) Life of Johnson

Brundage M, Avin S, Clark J, Toner H, Eckersley P, Garfinkel B, Dafoe A, Scharre P, Zeitzoff T, Filar B, Anderson H, Roff H, Allen GC, Steinhardt J, Flynn C, hÉigeartaigh SÓ, Beard S, Belfield H, Farquhar S, Amodei D (2018) The malicious use of artificial intelligence: forecasting, prevention, and mitigation. arXiv:1802.07228

Burt A, Trivers R (2006) Genes in conflict: the biology of selfish genetic elements. Harvard University Press

Carlsmith J (2022) Is power-seeking AI an existential risk? arXiv Preprint arXiv:2206.13353

Cook TB, Brenner LA, Cloninger CR, Langenberg P, Igbide A, Giegling I, Hartmann AM, Konte B, Friedl M, Brundin L (2015) "Latent" infection with *Toxoplasma gondii*: association with trait aggression and impulsivity in healthy adults. J Psychiatr Res 60:87–94

D'Agostino S (2023) 'AI Godfather' Yoshua Bengio: we need a humanity defense organization. Bulletin of the Atomic Scientists. https://thebulletin.org/2023/10/ai-godfather-yoshua-bengio-we-need-a-humanity-defense-organization/

Dawkins R (1976) The Selfish gene. Oxford University Press

de Koning AJ, Gu W, Castoe TA, Batzer MA, Pollock DD (2011) Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet 7(12):e1002384

de la Rosa S, del Mar Rigual M, Vargiu P, Ortega S, Djouder N (2024) Endogenous retroviruses shape pluripotency specification in mouse embryos. Sci Adv 10(4):eadk9394

Deininger P (2011) Alu elements: Know the SINEs. Genome Biol 12(12):1–12

Dobzhansky T (1967) The biology of ultimate concern. New American Library

Doolittle WF, Sapienza C (1980) Selfish genes, the phenotype paradigm and genome evolution. Nature 284(5757):601–603

Dor-Ziderman Y, Lutz A, Goldstein A (2019) Prediction-based neural mechanisms for shielding the self from existential threat. Neuroimage 202:116080

Estep P, Hoekstra A (2015) The leverage and centrality of mind. In: Aguirre A, Foster B, Merali Z (eds) How should humanity steer the future? Springer, pp 37–47

Fedoroff NV (2012) Transposable elements, epigenetics, and genome evolution. Science 338(6108):758–767

Galeon D (2016) AI will colonize the galaxy by the 2050s, according to the "Father of Deep Learning." Futurism. https://futurism.com/ai-will-colonize-the-galaxy-by-the-2050s-according-to-the-father-of-deep-learning

Goldstein S, Kirk-Giannini CD (2023) Language agents reduce the risk of existential catastrophe. AI Soc. https://doi.org/10.1007/s00146-023-01748-4

Goldstein S, Park PS (2023) AI systems have learned how to deceive humans. What does that mean for our future? The Conversation. https://theconversation.com/ai-systems-have-learned-how-to-deceive-humans-what-does-that-mean-for-our-future-212197

Good IJ (1966) Speculations concerning the first ultraintelligent machine. In: Advances in computers, vol 6. Elsevier, pp 31–88

Grace K, Stewart H, Sandkühler JF, Thomas S, Weinstein-Raun B, Brauner J (2024) Thousands of AI Authors on the Future of AI. arXiv Preprint arXiv:2401.02843

Hadar-Shoval D, Asraf K, Mizrachi Y, Haber Y, Elyoseph Z (2023) The invisible embedded "values" within large language models: Implications for mental health use. Research Square. https://www.researchsquare.com/article/rs-3456660/v1

Hamilton WD (1964) The genetical evolution of social behaviour. I. J Theor Biol 7(1):1–16

Hamilton WD (1966) The moulding of senescence by natural selection. J Theor Biol 12(1):12–45

Hammond G (2023) Aidan Gomez: AI threat to human existence is 'absurd' distraction from real risks. Financial Times. https://www.ft.com/content/732fc372-67ea-4684-9ab7-6b6f3cdfd736

Hardin G (1960) The competitive exclusion principle: an idea that took a century to be born has implications in ecology, economics, and genetics. Science 131(3409):1292–1297

Harris S (2005) The end of faith: religion, terror, and the future of reason. WW Norton & Company

Hassabis D, Kumaran D, Summerfield C, Botvinick M (2017) Neuroscience-inspired artificial intelligence. Neuron 95(2):245–258

Hawkins J (2015) The terminator is not coming. The future will thank Us. *Vox*. https://www.vox.com/2015/3/2/11559576/the-terminator-is-not-coming-the-future-will-thank-us

Heaven WD (2023) How existential risk became the biggest meme in AI. MIT Technology Review. https://www.technologyreview.com/2023/06/19/1075140/how-existential-risk-became-biggest-meme-in-ai/

Hebb DO (1949) The organization of behavior. Psychology Press. https://doi.org/10.4324/9781410612403

Heikkilä M, Heaven WD (2022) Yann LeCun has a bold new vision for the future of AI. MIT Technology Review

Hendrycks D (2023a) Natural selection favors AIs over humans. https://doi.org/10.48550/arXiv.2303.16200. arXiv:2303.16200

Hendrycks D (2023b) As it happens, my p(doom) > 80% [Twitter tweet]. https://twitter.com/DanHendrycks/status/1642394635657162753

Hendrycks D (2023c) Statement on AI risk | CAIS. https://www.safe.ai/statement-on-ai-risk

Hendrycks D, Mazeika M, Woodside T (2023) An overview of catastrophic AI risks. https://doi.org/10.48550/arXiv.2306.12001. arXiv:2306.12001

Hessen Schei T (2019) Ilya: the AI scientist shaping the world. https://www.theguardian.com/technology/ng-interactive/2023/nov/02/ilya-the-ai-scientist-shaping-the-world

Hinton G (2022) The forward-forward algorithm: some preliminary investigations. https://doi.org/10.48550/ARXIV.2212.13345

Inatomi Y, Sakata K, Arivanandhan M, Rajesh G, Nirmal Kumar V, Koyama T, Momose Y, Ozawa T, Okano Y, Hayakawa Y (2015) Growth of InxGa1− xSb alloy semiconductor at the International Space Station (ISS) and comparison with terrestrial experiments. Npj Microgravity 1(1):1–6

Johnson SK, Johnson PT (2021) Toxoplasmosis: recent advances in understanding the link between infection and host behavior. Annu Rev Anim Biosci 9:249–264

Johnson DG, Verdicchio M (2017) Reframing AI discourse. Mind Mach 27:575–590

Johnson DG, Verdicchio M (2019) AI, agency and responsibility: the VW fraud case and beyond. AI & Soc 34:639–647

Johnson SK, Fitza MA, Lerner DA, Calhoun DM, Beldon MA, Chan ET, Johnson PT (2018) Risky business: linking *Toxoplasma gondii* infection and entrepreneurship behaviours across individuals and countries. Proc R Soc B Biol Sci 285(1883):20180822

Jones N (2023) OpenAI's chief scientist helped to create ChatGPT—while worrying about AI safety. Nature 624(7992):503–503

Joy B (2000) Why the future doesn't need us: our most powerful 21st-century technologies-robotics, genetic engineering, and nanotech-are threatening to make humans an endangered species. WIRED. https://www.wired.com/2000/04/joy-2/

Knight W (2023) What really made Geoffrey Hinton into an AI Doomer. WIRED. https://www.wired.com/story/geoffrey-hinton-ai-chatgpt-dangers/

Kornai A, Bukatin M, Zombori Z (2023) Safety without alignment. arXiv Preprint arXiv:2303.00752

Kurzweil R (2005) The singularity is near: when humans transcend biology. Penguin

Leavitt D (2006) The man who knew too much: Alan Turing and the invention of the computer (great discoveries). WW Norton & Company

Lerner DA, Alkærsig L, Fitza MA, Lomberg C, Johnson SK (2021) Nothing ventured, nothing gained: parasite infection is associated with entrepreneurial initiation, engagement, and performance. Entrep Theory Pract 45(1):118–144

Lindahl C, Saeid H (2023) Unveiling the values of ChatGPT: An explorative study on human values in AI systems [KTH Royal Institute of Technology]. https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-329334

Malinowski B (1979) The role of magic and religion. In: Lessa WA, Vogt EZ (eds) Reader in comparative religion: an anthropological approach, vol 37. Harper and Row, New York, p 46

Martinez VO, de Mendonça Lima FW, De Carvalho CF, Menezes-Filho JA (2018) *Toxoplasma gondii* infection and behavioral outcomes in humans: a systematic review. Parasitol Res 117:3059–3065

Melnyk V, Melnyk A (2023). Analysis of methods, approaches and tools for organizing self-improvement of computer systems. In: 2023 13th International Conference on Advanced Computer Information Technologies (ACIT), pp 506–511

Metz C (2020) In two moves, AlphaGo and Lee Sedol redefined the future. Wired, 16 March 2016

Miller JD, Yampolskiy R, Häggström O (2020) An AGI modifying its utility function in violation of the strong orthogonality thesis. Philosophies 5(4):40

Minsky M (1994) Will robots inherit the Earth? Sci Am 271(4):108–113

Montagu A (1961) Neonatal and infant immaturity in man. JAMA 178(1):56–57

Moravec H (1988) Mind children: the future of robot and human intelligence. Harvard University Press

Moravec H (1998) When will computer hardware match the human brain. J Evol Technol 1(1):10

Nivel E et al (2013) Bounded recursive self-improvement. arXiv:1312.6764

Olson K (1999) Aum Shinrikyo: once and future threat? Emerg Infect Dis 5(4):513

Omohundro SM (2008a) The basic AI drives. In: Wang P, Goertzel B, Franklin S (eds) Proceedings of the 2008 conference on Artificial General Intelligence 2008, vol 171. IOS Press, pp 483–492

Omohundro SM (2008b) The nature of self-improving artificial intelligence. Singularity Summit 2007. https://selfawaresystems.files.wordpress.com/2008/01/nature_of_self_improving_ai.pdf

Ord T (2020) The precipice: existential risk and the future of humanity. Hachette Books

Orgel LE, Crick FH (1980) Selfish DNA: the ultimate parasite. Nature 284(5757):604–607

Ororbia A, Friston K (2023) Mortal computation: a foundation for biomimetic intelligence. arXiv:2311.09589

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

Pinker S (2012) The better angels of our nature: why violence has declined. Penguin Books

Pinker S (2018) Enlightenment now: the case for reason, science, humanism, and progress. Penguin Books

Polechová J, Storch D (2008) Ecological niche. In: Encyclopedia of ecology, vol 2. Elsevier, Oxford, pp 1088–1097

Qirko HN (2017) An evolutionary argument for unconscious personal death unawareness. Mortality 22(3):255–269

Robinson WG (1997) Heaven's gate: the end. J Comput-Mediated Commun 3(3):JCMC334

Rothblatt M (2015) Virtually human: the promise—and the Peril—of digital immortality. Picador

Rupprecht CE, Hanlon CA, Hemachudha T (2002) Rabies re-examined. Lancet Infect Dis 2(6):327–343

Russell S (2019) Human compatible: AI and the problem of control. Penguin Books Limited

Salles A, Evers K, Farisco M (2020) Anthropomorphism in AI. AJOB Neurosci 11(2):88–95

Schmidhuber J (2023) Jürgen Schmidhuber's home page. https://people.idsia.ch/~juergen/

Schopenhauer A (1818) The world as will and representation

Sevilla J, Heim L, Ho A, Besiroglu T, Hobbhahn M, Villalobos P (2022) Compute trends across three eras of machine learning. In: 2022 International Joint Conference on Neural Networks (IJCNN), pp 1–8

Sherwin WB (2023) Singularity or speciation? A comment on "AI safety on whose terms?" [eLetter]. Science 381(6654):138. https://doi.org/10.1126/science.adi8982

Sotala K (2018) Disjunctive scenarios of catastrophic AI risk. In: Artificial intelligence safety and security. Chapman and Hall, pp 315–337

Stacey K, Milmo D (2023) No 10 worried AI could be used to create advanced weapons that escape human control. The Guardian. https://www.theguardian.com/technology/2023/sep/25/ai-biowe apons-rishi-sunak-safety

Stanovich KE, West RF (2000) Advancing the rationality debate. Behav Brain Sci 23(5):701–717

Stanovich KE, West RF (2004) Evolutionary versus instrumental goals: how evolutionary psychology misconceives human rationality. In: Over DE (ed) Evolution and the psychology of thinking: the debate. Psychology Press, pp 171–230

Szathmáry E (2006) The origin of replicators and reproducers. Philos Trans R Soc London Ser B Biol Sci 361(1474):1761–1776. https://doi.org/10.1098/rstb.2006.1912

Tegmark M (2017) Life 3.0: being human in the age of artificial intelligence, 1st edn. Alfred A. Knopf

Tegmark M, Omohundro S (2023) Provably safe systems: the only path to controllable AGI. arXiv Preprint arXiv:2309.01933

Totschnig W (2019) The problem of superintelligence: political, not technological. AI & Soc 34:907–920

Totschnig W (2020) Fully autonomous AI. Sci Eng Ethics 26:2473–2485

Varki A (2009) Human uniqueness and the denial of death. Nature 460(7256):684–684

Varki A (2019) Did human reality denial breach the evolutionary psychological barrier of mortality salience? A theory that can explain unusual features of the origin and fate of our species. In: Shackelford T, Zeigler-Hill V (eds) Evolutionary perspectives on death. Springer, pp 109–135

Wiener N (1964) God & Golem, Inc.: a comment on certain points where cybernetics impinges on religion. The MIT Press. https://doi.org/10.7551/mitpress/3316.001.0001

Wikipedia contributors (2024) The giving pledge. In: Wikipedia. https://en.wikipedia.org/wiki/The_Giving_Pledge

Williams GC (1966) Adaptation and natural selection: a critique of some current evolutionary thought. Princeton University Press. https://doi.org/10.2307/j.ctv39x5jt

Williams GC (1993) Mother nature is a wicked old witch! In: Nitecki MH, Nitecki DV (eds) Evolutionary ethics. State University of New York Press, pp 217–231

Yampolskiy R (2016) Taxonomy of pathways to dangerous artificial intelligence. In: Workshops at the thirtieth AAAI Conference on Artificial Intelligence

Yampolskiy R (2020) On controllability of artificial intelligence. In: IJCAI-21 Workshop on Artificial Intelligence Safety (AISafety2021)

Yudkowsky E (2008) Artificial Intelligence as a positive and negative factor in global risk. In: Rees MJ, Bostrom N, Cirkovic MM (eds) Global catastrophic risks. Oxford University Press, pp 308–345. https://doi.org/10.1093/oso/9780198570509.003.0021

Yudkowsky E (2016) The AI alignment problem: why it is hard, and where to start. In: Symbolic Systems Distinguished Speaker, 4.

Yudkowsky E (2023) Pausing AI developments isn't enough. We need to shut it all down. Time. https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/

Zador A, Escola S, Richards B, Ölveczky B, Bengio Y, Boahen K, Botvinick M, Chklovskii D, Churchland A, Clopath C, DiCarlo J, Ganguli S, Hawkins J, Körding K, Koulakov A, LeCun Y, Lillicrap T, Marblestone A, Olshausen B, Tsao D (2023) Catalyzing next-generation Artificial Intelligence through NeuroAI. Nat Commun 14(1):Article 1. https://doi.org/10.1038/s41467-023-37180-x

Zangeneh-Nejad F, Sounas DL, Alù A, Fleury R (2021) Analogue computing with metamaterials. Nat Rev Mater 6(3):207–225

Zelikman E, Lorch E, Mackey L, Kalai AT (2023) Self-Taught Optimizer (STOP): recursively self-improving code generation. arXiv:2310.02304