How Hard is Artificial Intelligence? Evolutionary Arguments and Selection Effects

Carl Shulman
Nick Bostrom

Abstract

Several authors have made the argument that because blind evolutionary processes produced human intelligence on Earth, it should be feasible for clever human engineers to create human-level artificial intelligence in the not-too-distant future. This evolutionary argument, however, has ignored the observation selection effect that guarantees that observers will see intelligent life having arisen on their planet no matter how hard it is for intelligent life to evolve on any given Earth-like planet. We explore how the evolutionary argument might be salvaged from this objection, using a variety of considerations from observation selection theory and analysis of specific timing features and instances of convergent evolution in the terrestrial evolutionary record. We find that, depending on the resolution of disputed questions in observation selection theory, the objection can be either be wholly or moderately defused, although other challenges for the evolutionary argument remain.

1. Evolutionary arguments for easy intelligence

1.1 Introduction

What can human evolution tell us about the prospects for human-level Artificial Intelligence (AI)?[1] A number of philosophers and technologists, including David Chalmers (2010) and Hans Moravec (1976, 1988, 1998, 1999), argue that human evolution shows that such AI is not just possible but feasible within this century. On these accounts, we can estimate the relative capability of evolution and human engineering to produce intelligence, and find that human engineering is already vastly superior to evolution in some areas and is likely to become superior in the remaining areas before too long. The fact that evolution produced intelligence therefore indicates that human engineering will be able to do the same. Thus, Moravec writes:

---

[1] Here, we mean systems which match or exceed the cognitive performance of humans in virtually all domains of interest; uniformly "human-level" performance seems unlikely, except perhaps through close emulation of human brains (Sandberg and Bostrom, 2008), since software is already superhuman in many fields.

The existence of several examples of intelligence designed under these constraints should give us great confidence that we can achieve the same in short order. The situation is analogous to the history of heavier than air flight, where birds, bats and insects clearly demonstrated the possibility before our culture mastered it.[2]

Similarly, Chalmers sketches the evolutionary argument as follows:

1. Evolution produced human intelligence [mechanically and non-miraculously].
2. If evolution can produce human intelligence [mechanically and non-miraculously], then we can probably produce human-level artificial intelligence (before long).
_____
3. We can probably produce human-level artificial intelligence (before long).

These arguments for the feasibility of machine intelligence do not say whether the path to be taken by human engineers to produce AI will resemble the path taken by evolution. The fact that human intelligence evolved implies that running genetic algorithms is *one* way to produce intelligence; it does not imply that it is the only way or the easiest way for human engineers to create machine intelligence. We can therefore consider two versions of the evolutionary argument depending on whether or not the engineering of intelligence is supposed to use methods that recapitulate those used by evolution.

1.2 Argument from problem difficulty

The argument from problem difficulty tries to use evolutionary considerations indirectly to demonstrate that the problem of creating intelligent systems is not too hard (since blind evolution did it) and then use this as a general ground for thinking that human engineers will probably soon crack the problem too. One can think of this argument as making a claim about the space of possible algorithms to the effect that it is not too difficult to search in this space and find an algorithm that produces human-level intelligence when implemented on practically feasible hardware. (The difficulty depends on unknown facts about the space of algorithms for intelligence, such as the extent to which the shape of the fitness landscape favors hill-climbing.) We can formalize this argument as follows:

1'. Evolution produced human intelligence.
2'. If evolution produced human intelligence, then it is "non-hard" for evolutionary processes to produce human intelligence.
3'. If it is "non-hard" for evolutionary processes to produce human evolution, then it is not extremely difficult for engineers to produce human-level machine intelligence.
4'. If it is not extremely difficult for engineers to produce human-level machine intelligence, it will probably be done before too long.
_____
5'. Engineers will (before long) produce human-level machine intelligence.

While (1') is well established, and we may grant (4'), premises (2') and (3') require careful scrutiny.

Let us first consider (3'). Why believe that it would not be extremely difficult for human engineers to figure out how to build human-level machine intelligence, assuming that it was "non-hard" (in a sense

_____

[2] See Moravec (1976).

that will be explained shortly) for evolution to do so? One reason might be optimism about the growth of human problem-solving skills in general or about the ability of AI researchers in particular to come up with clever new ways of solving problems. Such optimism, however, would need some evidential support, support that would have to come from outside the evolutionary argument. Whether such optimism is warranted is a question outside the scope of this paper, but it is important to recognize that this is an essential premise in the present version of the evolutionary argument, a premise that should be explicitly stated. Note also that if one were *sufficiently* optimistic about the ability of AI programmers to find clever new tricks, then the evolutionary argument would be otiose: human engineers could then be expected to produce (before too long) solutions even to problems that were "extremely difficult" (at least extremely difficult to solve by means of blind evolutionary processes).

Despite the need for care in developing premise (3') in order to avoid a *petitio principii*, the argument from problem difficulty is potentially interesting and possesses some intuitive appeal. We will therefore return to this version of the argument in later sections of the paper, in particular focusing our attention on premise (2'). We will then see that the assessment of evolutionary difficulty turns out to involve some deep and intricate issues in the application of observation selection theory to the historical record.

1.3 Argument from evolutionary algorithms

The second version of the evolutionary argument for the feasibility of machine intelligence does not attempt to parlay evolutionary considerations into a general assessment of how hard it would be to create machine intelligence using some unspecified method. Instead of looking at general problem difficulty, the second version focuses on the more specific idea that genetic algorithms run on sufficiently fast computers could achieve results comparable to those of biological evolution. We can formalize this "argument from evolutionary algorithms" as follows:

> 1'. Evolution produced human intelligence.
> 2'. If evolution produced human intelligence, then it is "non-hard" for evolutionary processes to produce human intelligence.
> 3''. We will (before long) be able to run genetic algorithms on computers that are sufficiently fast to recreate on a human timescale the same amount of cumulative optimization power that the relevant processes of natural selection instantiated throughout our evolutionary past (for any evolutionary process that was non-hard).
> _____
> 4''. We will (before long) be able to produce by running genetic algorithms results comparable to some of the results that evolution produced, including systems that have human-level intelligence.

This argument from evolutionary algorithms shares with the argument from problem difficulty its first two premises. Our later investigations of premise (2') will therefore bear on both versions of the evolutionary argument. Let us take a closer look at this premise.

1.4 Evolutionary hardness and observation selection effects

We have various methods available to begin to estimate the power of evolutionary search on Earth: estimating the number of generations and population sizes available to human evolution[3], creating mathematical models of evolutionary "speed limits" under various conditions[4], and using genomics to measure past rates of evolutionary change.[5] However, reliable creation of human-level intelligence through evolution might require trials on many planets in parallel, with Earth being one of the lucky few to succeed. Can the fact of human evolution on Earth let us distinguish between the following scenarios?

> *Non-hard Intelligence*: There is a smooth path of incremental improvement from the simplest primitive nervous systems to brains capable of human-level intelligence, reflecting the existence of many simple, easily-discoverable algorithms for intelligence. On most planets with life, human-level intelligence also develops.

> *Hard Intelligence*: Workable algorithms for intelligence are rare, without smooth paths of incremental improvement to human-level performance. Evolution requires extraordinary luck to hit upon a design for human-level intelligence, so that only 1 in 10^{1000} planets with life does so.

In either scenario every newly evolved civilization will find that evolution managed to produce its ancestors. The observation selection effect is that no matter how hard it is for human-level intelligence to evolve, 100% of evolved civilizations will find themselves originating from planets where it happened anyway.

How confident can we be that Hard Intelligence is false, and that premise (2') in the evolutionary arguments can be supported, in the face of such selection effects? After a brief treatment of premise (3''), we discuss the theoretical approaches in the philosophical literature, particularly the Self-Sampling Assumption (SSA) and the Self-Indication Assumption (SIA)—because, unfortunately, correctly analyzing the evidence on evolution depends on difficult, unsettled questions concerning observer-selection effects.[6] We note that one common set of philosophical assumptions (SIA) supports easy evolution of intelligence, but that it does so on almost *a priori* grounds that some may find objectionable. Common alternatives to SIA, on the other hand, require us to more carefully weigh the evolutionary data. We attempt this assessment, discussing several types of evidence which hold up in the face of observation selection effects. We find that while more research is needed, the thesis that "intelligence is exceedingly hard to evolve" is consistent with the available evolutionary data under these alternative assumptions. However, the data do rule out many particular hypotheses under which intelligence might be exceedingly hard to evolve, and thus the evolutionary argument should still increase our credence in the feasibility of human-level AI.

2. Computational requirements for recapitulating evolution through genetic algorithms

---

[3] Baum (2004) very roughly estimates that between 10^{30} and 10^{40} creatures have existed on Earth, in the course of arguing that evolutionary search could not have relied on brute force to search the space of possible genomes. However, Baum does not consider the implications of an ensemble of planets in his calculation.

[4] For instance, MacKay (2009) computes information-theoretic upper bounds to the power of natural selection with and without sex in a simple additive model of fitness.

[5] See, e.g., Hawks et al. (2007) on recently accelerating adaptive selection in humans, including comparison of adaptive substitution rates in different primate lineages.

[6] See Grace (2010) for a helpful review of these questions and prominent approaches.

Let us assume (1') and (2'), i.e. that it was non-hard in the sense described above for evolution to produce human intelligence. The argument from evolutionary algorithms then needs one additional premise to deliver the conclusion that engineers will soon be able to create machine intelligence, namely that we will soon have computing power sufficient to recapitulate the relevant evolutionary processes that produced human intelligence. Whether this is plausible depends both on what advances one might expect in computing technology over the next decades and on how much computing power would be required to run genetic algorithms with the same optimization power as the evolutionary process of natural selection that lies in our past. One might for example try to estimate how many doublings in computational performance, along the lines of Moore's law, one would need in order to duplicate the relevant evolutionary processes on computers.

Now, to pursue this line of estimation, we need to realize that not every feat that was accomplished by evolution in the course of the development of human intelligence is relevant to a human engineer who is trying to artificially evolve machine intelligence. Only a small portion of evolutionary optimization on Earth has been selection for intelligence. More specifically, the problems that human engineers cannot trivially bypass may have been the target of a very small portion of total evolutionary optimization. For example, since we can run our computers on electrical power, we do not have to reinvent the molecules of the cellular energy economy in order to create intelligent machines—yet molecular evolution might have used up a large part of the total amount of selection power that was available to evolution over the course of Earth's history.

One might argue that the key insights for AI are embodied in the structure of nervous systems, which came into existence less than a billion years ago.[7] If we take that view, then the number of relevant "experiments" available to evolution is drastically curtailed. There are are some $4\text{-}6*10^{30}$ prokaryotes in the world today[8], but only 10^{19} insects[9], and fewer than 10^{10} human (pre-agricultural populations were orders of magnitude smaller). However, evolutionary algorithms require not only variations to select among but a fitness function to evaluate variants, typically the most computationally expensive component. A fitness function for the evolution of artificial intelligence plausibly requires simulation of "brain development," learning, and cognition to evaluate fitness. We might thus do better not to look at the raw number of organisms with complex nervous systems, but instead to attend to the number of neurons in biological organisms that we might simulate to mimic evolution's fitness function. We can make a crude estimate of that latter quantity by considering insects, which dominate terrestrial biomass, with ants alone estimated to contribute some 15-20% of terrestrial animal biomass.[10] Insect brain size varies substantially, with large and social insects enjoying larger brains; e.g., a honeybee brain has just

---

[7] Legg (2008) offers this reason in support of the claim that humans will be able to recapitulate the progress of evolution over much shorter timescales and with reduced computational resources (while noting that evolution's unadjusted computational resources are far out of reach). Baum (2004) argues that some developments relevant to AI occurred earlier, with the organization of the genome itself embodying a valuable representation for evolutionary algorithms.

[8] See Whitman et al. (1998).

[9] See Sabrosky (1952).

[10] See Schultz (2000).

under 10^6 neurons[11], while a fruit fly brain has 10^5 neurons[12], and ants lie in between with 250,000 neurons. The majority of smaller insects may have brains of only a few thousand neurons. Erring on the side of conservatively high, if we assigned all 10^{19} insects fruit-fly numbers of neurons the total would be 10^{24} insect neurons in the world. This could be augmented with an additional order of magnitude, to reflect aquatic copepods, birds, reptiles, mammals, etc., to reach 10^{25}. (By contrast, in pre-agricultural times there were fewer than 10^7 humans, with under 10^{11} neurons each, fewer than 10^{18} total, although humans have a high number of synapses per neuron.)

The computational cost of simulating one neuron depends on the level of detail that one wants to include in the simulation. Extremely simple neuron models use about 1,000 floating-point operations per second (FLOPS) to simulate one neuron (for one second of simulated time); an electrophysiologically realistic Hodgkin-Huxley model uses 1,200,000 FLOPS; a more detailed multicompartmental model would add another 3-4 orders of magnitude, while higher-level models that abstract systems of neurons could subtract 2-3 orders of magnitude from the simple models.[13] If we were to simulate 10^{25} neurons over a billion years of evolution (longer than the existence of nervous systems as we know them) in a year's runtime these figures would give us a range of 10^{31}-10^{44} FLOPS. By contrast, the Japanese K computer, currently the world's most powerful supercomputer, provides only 10^{16} FLOPS. In recent years it has taken approximately 6.7 years for commodity computers to to increase in power by one order of magnitude. Even a century of continued Moore's law would not be enough to close this gap. Running more or specialized hardware, or longer runtimes, could contribute only a few more orders of magnitude.

This figure is conservative in another respect. Evolution achieved human intelligence yet it was not *aiming* at this outcome—put differently: the fitness functions for natural organisms do not select only for intelligence and its precursors.[14] Even environments in which organisms with superior information-processing skills reap various rewards may not select for intelligence, because improvements to intelligence can and often do impose significant costs, such as higher energy consumption or slower maturation times, and those costs that may outweigh whatever benefits are derived from smarter behaviour. Excessively deadly environments reduce the value of intelligence: the shorter one's expected lifespan, the less time there will be for increased learning ability to pay off. Reduced selective pressure for intelligence slows the spread of intelligence-enhancing innovations, and thus the opportunity for selection to favor subsequent innovations that depend on those. Furthermore, evolution may wind up stuck in local optima that humans would notice and bypass by altering trade-offs between exploitation and exploration or by providing a smooth progression of increasingly difficult intelligence tests.[15] And as mentioned above, evolution scatters much of its selection power on traits that are unrelated to intelligence, such as Red Queen's races of co-evolution between immune systems and parasites. Evolution will continue to waste resources producing mutations that have been reliably lethal, and will

---

[11] See Menzel and Giurfa (2001).
[12] See Truman et al. (1993).

[13] See Sandberg and Bostrom (2008).
[14] See Legg (2008) for further discussion of this point, and of the promise of functions or environments that determine fitness based on a smooth landscape of pure intelligence tests.
[15] See Bostrom (2009) for a taxonomy and more detailed discussion of ways in which engineers may outperform historical selection.

fail to make use of statistical similarities in the effects of different mutations. All these represent inefficiencies in natural selection (when viewed as a means of evolving intelligence) that it would be relatively easy for a human engineer to avoid while using evolutionary algorithms to develop intelligent software.

It seems plausible that avoiding inefficiencies like those just described would make it possible to trim many orders of magnitude from the 10^{31}-10^{44} FLOPS range calculated above pertaining to the number of neural computations that have been performed in our evolutionary past. Unfortunately, it is difficult to find a basis on which to estimate *how* many orders of magnitude. It is difficult even to make a rough estimate—for aught we know, the efficiency savings could be 5 or 10 or 25 orders of magnitude.

The above analysis addressed the nervous systems of living creatures, without reference to the cost of simulating bodies or the surrounding virtual environment as part of a fitness function. It is plausible that an adequate fitness function could test the competence of a particular organism in far fewer operations than it would take to simulate all the neuronal computation of that organism's brain throughout its natural lifespan. AI programs today often develop and operate in very abstract environments (theorem-provers in symbolic math worlds, agents in simple game tournament worlds, etc.)

A skeptic might insist that an abstract environment would be inadequate for the evolution of general intelligence, believing instead that the virtual environment would need to closely resemble the actual biological environment in which our ancestors evolved. Creating a physically realistic virtual world would require a far greater investment of computational resources than the simulation of a simple toy world or abstract problem domain (whereas evolution had access to a physically realistic real world "for free"). In the limiting case, if complete microphysical accuracy were insisted upon, the computational requirements would balloon to utterly infeasible proportions.[16] However, such extreme pessimism seems unlikely to be well founded; it seems unlikely that the best environment for evolving intelligence is one that mimics nature as closely as possible. It is, on the contrary, plausible that it would be more efficient to use an artificial selection environment, one quite unlike that of our ancestors, an environment specifically designed to promote adaptations that increase the type of intelligence we are seeking to evolve (say, abstract reasoning and general problem-solving skills as opposed to maximally fast instinctual reactions or a highly optimized visual system).

Where does premise (3'') stand? The computing resources to match historical numbers of neurons in straightforward simulation of biological evolution on Earth are severely out of reach, even if Moore's law continues for a century. The argument from evolutionary algorithms depends crucially on the magnitude of efficiency gains from clever search, with perhaps as many as thirty orders of magnitude required. Precise estimation of those efficiency gains is beyond the scope of this paper.

---

[16] One might seek to circumvent this through the construction of robotic bodies that would let simulated creatures interact directly with the real physical world. But the cost and speed penalties of such an implementation would be prohibitive (not to mention the technical difficulties of creating robots that could survive and reproduce in the wild!) With macroscopic robotic bodies interacting with the physical world in realtime, it might take millions of years to recapitulate important evolutionary developments.

In lieu of an estimate supporting (3''), one has to fall back on the more general argument from problem difficulty, in which (3'') is replaced by (3') and (4'), premises which might be easier to support on intuitive grounds. But the argument from problem difficulty also requires premise (2'), that the evolution of intelligence on Earth was "non-hard." (This premise was also used in the argument from evolutionary algorithms: if (2') were false, so that one would have to simulate evolution on vast numbers of planets to reliably produce intelligence through evolutionary methods, then computational requirements could turn out to be many, many orders of magnitude higher still.) We now turn to discuss (2') more closely, and theoretical approaches to evaluating it.

3. Two theories of observational selection effects

Does the mere fact that we evolved on Earth let us distinguish between the Hard Intelligence and Non-hard Intelligence scenarios? Related questions arise in philosophy[17], decision theory[18], and cosmology[19], and the two leading approaches to answering them give conflicting answers. We can introduce these approaches using the following example:

> *God's Coin Toss*: Suppose that God tosses a fair coin. If it comes up heads, he creates ten people, each in their own room. If tails, he creates one thousand people, each in their own room. The rooms are numbered 1-10 or 1-1000. The people cannot see or communicate with the other rooms. Suppose that you know all this, and you discover that you are in one of the first ten rooms. How should you reason that the coin fell?

The first approach begins with the Self-Sampling Assumption:

> (SSA) Observers should reason as if they were a random sample from the set of all observers in their reference class.[20]

Here the reference class is some set of possible observers, e.g. "intelligent beings" or "humans" or "creatures with my memories and observations." If the reference class can include both people who discover they are in rooms 1-10 and people who discover they are in rooms 11-1000, then applying SSA will lead you to conclude, with probability 100/101, that the coin fell heads.[21] For if the coin came up

---

[17] See, for instance, the philosophical debate over "Sleeping Beauty" cases, beginning with Elga (2000) and Lewis (2001).

[18] See Piccione & Rubinstein (1997) on the Absentminded Driver problem.

[19] If we consider cosmological theories on which the world is infinite (or finite but exceedingly large) with sufficient local variation, then all possible observations will be made somewhere. To make predictions using such theories we must take into account the indexical information that we are making a particular observation, rather than the mere fact that some observer somewhere has made it. To do so principles such as SSA and SIA must be combined with some measure over observers, as discussed in Bostrom (2007) and Grace (2010).

[20] This approach was pioneered by Carter (e.g., 1983), developed by Leslie (e.g., 1993) and Bostrom (2002a), and is used implicitly or explicitly by a number of other authors, e.g. Lewis (2001). Bostrom (2002a) offers an extension to consider "observer-moments," SSSA.

[21] If the reference class includes only observers who have discovered that they are in one of the first ten rooms, then SSA will not alter our credences in this example.

heads then 100% of the reference class would find itself in your situation, but if it came up tails then only 1% of the reference class would find itself in your situation. On the other hand, prior to the discovery of your room number you should consider heads and tails equally likely, since 100% of your reference class would find itself in your situation either way.

The second approach adds an additional principle, the Self-Indication Assumption:

> (SIA) Given the fact that you exist, you should (other things equal) favor hypotheses according to which many observers exist over hypotheses on which few observers exist.[22]

In the SSA+SIA combination, if we take SIA to apply to members of a reference class that includes all observers indistinguishable from ourselves, the specific reference class no longer matters: a more expansive reference class receives a probability boost from having more observers in it, but this is exactly offset by the probability penalty for making our observations a smaller portion of the reference class. The details of the reference class no longer play a significant role. SIA then gives us the following algorithm: first assign probabilities to possible worlds normally, then multiply the probability of each possible world by the number of observers in situations subjectively indistinguishable from one's own, apply a renormalization constant so that probabilities add up to 1, and divide the probability of each world evenly among the indexical hypotheses that you are each particular observer (indistinguishable from yourself) in that world.

In God's Coin Toss, this algorithm means that before discovering your room number you consider a result of tails 100 times more likely than heads, since conditional on tails there will be one hundred times as many observers in your situation as there would be given heads. After you discover that yours is among the first ten rooms, you will consider heads and tails equally likely, as an equal number of observers will find themselves in your evidential situation regardless of the flip's outcome.

Equipped with these summaries, we can see that SSA offers a formalization of the intuition that the mere fact that we evolved is not enough to distinguish Non-hard Intelligence from Hard Intelligence: if we use a reference class like "humans" or "evolved intelligent beings", then, in both scenarios, 100% of the members of the reference class will find themselves in a civilization that managed to develop anyway. SSA also lets us draw inferences about evolutionary developments that are not so clouded by observation selection effects. For instance, suppose that we are evenly divided (on non-indexical considerations) between the hypotheses that echolocation is found on either 1% or 100% of planets with relevantly similar populations of observers. Upon observing that echolocation exists on Earth, we could again make a Bayesian update as in God's Coin Toss and conclude that common echolocation is 100 times as likely as rare echolocation.

On this account, evolutionary innovations required to produce intelligence will be observed regardless of their difficulty, while other innovations will be present only if they are relatively easy given background conditions (including any innovations or other conditions required for intelligence). Observation

---

[22] We will abbreviate the SSA+SIA combination as SIA for brevity. SIA has been developed repeatedly as a response to the Doomsday Argument, as in Olum (2002) and Dieks (2007), and is closely connected with the "thirder" position on Sleeping Beauty cases as in Elga (2000).

selection might conceal the difficulty or rarity in the development of humans, nervous systems, eukaryotes, abiogenesis, even the layout of the Solar System or the laws of physics. We would need to look at *other* features of the evolutionary record, such as the timing of particular developments, innovations not in the line of human ancestry, and more direct biological data. We explore these lines in sections 5 and 6.

However, this approach is not firmly established, and the SIA approach generates very different conclusions, as discussed in the next section. These divergent implications provide a practical reason to work towards an improved picture of observation selection effects. However, in the meantime we have reason to attend to the results of both of the most widely held current theories.

4. The Self-Indication Assumption (SIA) favors the evolutionary argument

Initially, the application of SIA to the question of the difficulty of evolution may seem trivial: SIA strongly favors more observers with our experiences, and if the evolution of intelligence is very difficult, then it will be very rare for intelligence like ours to evolve in the universe. If, prior to applying SIA, we were equally confident in Non-hard Intelligence and Hard Intelligence, then when we apply SIA we will update our credences to consider Non-hard Intelligence 10^{1000} times as likely as Hard Intelligence, since we would expect 10^{1000} times as many planets to evolve observers indistinguishable from us under Non-hard Intelligence. This probability shift could overwhelm even exceedingly strong evidence to the contrary: if the non-indexical evidence indicated that Hard Intelligence was a trillion trillion times as probable as Non-hard Intelligence, an SIA user should still happily bet ten billion dollars against a cent that the evolution of intelligence is not Hard.

However, when we consider hypotheses on which the evolution of intelligence is increasingly easy the frequency of observations indistinguishable from ours may actually decline past a certain point. We observe a planet where the evolution of humanity took 4.5 billion years after the formation of the Earth. If intelligence arose sufficiently quickly and reliably, then we would expect life and intelligence to evolve early in a planet's lifetime: there would be more planets with intelligence, but fewer planets with late-evolved civilizations like ours. This consideration, in combination with SIA, might seem to favor an intermediate level of difficulty, so that the evolution of intelligence typically takes several billion years with the resources of the Earth and occurs fairly reliably but not always on life-bearing planets.

Similarly, we observe no signs of intelligent alien life. If intelligent life were common, it might have colonized the Earth before humans could develop, or made itself visible, in which case no humans on Earth would make our exact observations. Some combination of barriers, the so-called "Great Filter," must have prevented such alien life from developing near us and preempting our observations.[23] However, other things equal, SIA initially appears to favor explanations of the Great Filter which place the barriers after the evolution of intelligence. Here the thought is that if interstellar travel and communication are practically impossible, or if civilizations almost invariably destroy themselves before space colonization, then observers like us can be more frequent; so if we have even a small initial credence in such explanations of the Great Filter then after applying SIA we will greatly prefer them to "intelligence is rare" explanations. Even if one initially had only 0.1% credence that the explanation of

---

[23] See Hanson (1998a) on the Great Filter. Neal (2007) and Grace (2010) explore the interaction with SIA-like principles.

the Great Filter allowed for reliable evolution of observers like us (e.g., space travel is impossible, or advanced civilizations enforce policies against easily detectable activities on newcomers), application of the SIA would boost the probability of such explanations sufficiently to displace hypotheses that imply that advanced life is extremely rare. This would seem to leave the evolutionary argument for AI on sound footing, from the perspective of an SIA proponent.[24]

However, the preceding analysis assumed that our observations of a fairly old but empty galaxy were accurate descriptions of bedrock reality. One noted implication of SIA is that it tends to undermine that assumption. Specifically, the Simulation Argument raises the possibility that given certain plausible assumptions, e.g. that computer simulations of brains could be conscious, then computer simulations with our observations could be many orders of magnitude more numerous than "bedrock reality" beings with our observations.[25] Without SIA, the Simulation Argument need not bring us to the Simulation Hypothesis, i.e. the claim that we are computer simulations being run by some advanced civilization, since the assumptions might turn out to be false.[26] However, if we endorse SIA, then even if our non-indexical evidence is strongly against the Simulation Hypothesis, an initially small credence in the hypothesis can be amplified by SIA (and the potential for very large simulated populations) to extreme confidence.[27] This would favor hypotheses on which intelligence evolved frequently enough that advanced civilizations would be able to claim a large share of the resources suitable for computation (to run simulations), but increased frequency beyond that would not significantly increase the maximum population of observers indistinguishable from us. The combination of the Simulation Hypothesis and SIA would also independently favor the feasibility of AI, since advanced AI technology would increase the feasibility of producing very large simulated populations.

To sum up this section, known applications of the SIA consistently advise us to assign negligible probability to Hard Intelligence, even in the face of very strong contrary evidence, so long as we assign even miniscule prior probability to relatively easy evolution of intelligence. Since the number of planets with intelligence and the number of observers indistinguishable from us can come apart, the SIA allows for the evolution of intelligence to be some orders of magnitude more difficult than once per solar system, but not so difficult that the great majority of potential resources for creating observers go unclaimed. Drawing such strong empirical conclusions from seemingly almost *a priori* grounds may seem objectionable. However, the Self-Indication Assumption has a number of such implications, e.g. that if

---

[24] Grace (2010) argues that AI might be expected to be able to better overcome barriers to interstellar travel and communication the Great Filter in combination with SIA should reduce our credence in AI powerful enough to engage in interstellar travel. The strength of this update would depend on our credence in other explanations of the Great Filter, and is arguably rendered moot by the analysis of the the interaction of SIA with the Simulation Hypothesis in subsequent paragraphs.

[25] The argument is presented in Bostrom (2003), see also Bostrom & Kulczycki (2011).

[26] Note, per Chalmers (2003) and Bostrom (2003, 2005) that the Simulation Hypothesis is not a skeptical hypothesis, but a claim about what follows from our empirical evidence about the feasibility of various technologies. Most of our ordinary beliefs would remain essentially accurate.

[27] Note that SIA also amplifies our credence in hypotheses that simulator resources are large. If we assign even a small probability to future technology enabling arbitrarily vast quantities of computation, this hypothesis can dominate our calculations if we apply SIA.

we non-indexically assign any finite positive prior probability to the world containing infinitely many observers like us then post-SIA we must believe that this is true with probability 1.[28] Defenders of SIA willing to bite such bullets in other contexts may do the same here, and for them the evolutionary argument for AI will seem on firm footing. However, if one thinks that our views on these matters should be more sensitive to the observational evidence, then one must turn from the SIA and look elsewhere for relevant considerations.

We now move on to more detailed descriptions of Earth's evolutionary history, information that can be combined with SSA to assess the evolvability of intelligence, without the direct bias against Hard Intelligence implied by SIA.

5. SSA and evidence from convergent evolution

Recall that within the SSA framework we reason as though we were randomly selected from the set of all observers in our reference class. If the reference class includes only human-level intelligences, then nearly 100% of the members of the reference class will stem from an environment where evolution produced human-level intelligence at least once. By the same token, if there are innovations that are required for the evolution of human-level intelligence, these should be expected to evolve at least once among the ancestors of the human-level intelligences. However, nothing in the observation selection effect requires that observers find that human-level intelligence or any precursor innovations evolved *more than once* or *outside the line of ancestry* leading up to the human-level intelligences. Thus, evidence of convergent evolution—the independent development of an innovation in multiple taxa—can help us to understand the evolvability of human intelligence and its precursors, and to evaluate the evolutionary arguments for AI.

The Last Common Ancestor (LCA) shared between humans and octopuses, estimated to have lived at least 560 million years in the past, was a tiny wormlike creature with an extremely primitive nervous system; it was also an ancestor to nematodes and earthworms.[29] Nonetheless, octopuses went on to evolve extensive central nervous systems, with more nervous system mass (adjusted for body size) than fish or reptiles, and a sophisticated behavioral repertoire including memory, visual communication, and tool use.[30] Impressively intelligent animals with more recent LCAs include, among others, corvids (crows and ravens, LCA about 300 million years ago)[31], elephants (LCA about 100 million years ago).[32] In other words, from the starting point of those wormlike common ancestors in the environment of Earth,

---

[28] See Bostrom and Cirkovic (2003).

[29] See Erwin and Davidson (2002).

[30] See e.g. Mather (1994, 2008), Finn, Tregenza, and Norman (2009) and Hochner, Shomrat, & Fiorito, G. (2006) for a review of octopus intelligence.

[31] For example, a crow named Betty was able to bend a straight wire into a hook in order to retrieve a food bucket from a vertical tube, without prior training; crows in the wild make tools from sticks and leaves to aid their hunting of insects, pass on patterns of tool use, and use social deception to maintain theft-resistant caches of food; see Emery and Clayton (2004). For LCA dating, see Benton and Ayala (2003).

[32] See Archibald (2003) for LCA dating, and Byrne, Bates, and Moss (2009) for a review arguing that elephants' tool use, number sense, empathy, and ability to pass the mirror test suggest that they are comparable to non-human great apes.

the resources of evolution independently produced complex learning, memory, and tool use both within and without the line of human ancestry.

Some proponents of the evolutionary argument for AI, such as Moravec (1976), have placed great weight on such cases of convergent evolution. Before learning about the diversity of animal intelligence, we would assign some probability to scenarios in which the development of these basic behavioral capabilities (given background conditions) was a major barrier to creating human-level intelligence. To the extent convergent evolution lets us rule out particular ways in which the evolution of intelligence could be hard, it should reduce our total credence in the evolution of intelligence being hard (which is just the sum of our credence in all the particular ways it could be hard).

There is, however, an important caveat to such arguments. A species that displays convergent evolution of intelligence behaviorally may have a cognitive architecture that differs in unobserved ways from that of human ancestors. Such differences could mean that the animal brains embody algorithms that are hard to "scale" or build upon to produce human-level intelligence, so that their ease of evolution has little bearing on AI feasibility. By way of analogy, chess-playing programs outperform humans within the limited domain of chess, yet the underlying algorithms cannot be easily adapted to other cognitive tasks, let alone human-level AI. Insofar as we doubt the "scalability" of octopus or corvid intelligence, despite the appearance of substantial generality, we will discount arguments from their convergent evolution accordingly.

Further, even if we condition on the relevant similarity of intelligence in these convergent lineages and those ancestral to humans, observation selection effects could still conceal extraordinary luck in factors shared by both. First, background environmental effects, such as the laws of physics, the layout of the Solar System, and the geology of the Earth could all be unusually favorable to the evolution of intelligence (relative to simulated alternatives for AI), regardless of convergent evolution. Second, the LCA of all these lineages was already equipped with various visible features—such as nervous systems— that evolved only once in Earth's history and which might therefore have been arbitrarily difficult to evolve. While background conditions such as geology and the absence of meteor impacts seem relatively unlikely to correspond to significant problems for AI designers, it is somewhat less implausible to suppose that early neurons conceal extraordinary design difficulty: while computational models of individual neurons have displayed impressive predictive success, they might still harbor subtly relevant imperfections.[33] Finally, some subtle features of the LCA may have greatly enabled later development of intelligence without immediate visible impact. Consider the case of eyes, which have developed in many different animal lineages with widely varying anatomical features and properties (compare the eyes of humans, octopuses, and fruit flies). Eyes in all lineages make use of the proteins known as opsins, and some common regulatory genes such as PAX6, which were present in the LCA of all the creatures with eyes.[34] Likewise, some obscure genetic or physiological mechanism dating back to the octopus-human LCA may both be essential to the later development of octopus-level intelligence in various lineages and have required extraordinary luck.

---

[33] See Sandberg and Bostrom (2008), for a review.
[34] See Schopf (1992) on the convergent evolution of eyes.

In addition to background conditions shared by lineages, convergent evolution also leaves open the possibility of difficult innovations lying between the abilities of elephants or corvids or octopuses and human-level intelligence, since we have no examples of robustly human-level capabilities evolving convergently. If the evolution of human-level intelligence were sufficiently easy, starting from the capabilities of these creatures on Earth, then it might seem that observers should find that it appeared multiple times in evolution on their planets. However, as human technology advanced, we have caused mass extinctions (including all other hominids) and firmly occupied the ecological niche of dominant tool-user. If the evolution of human-level intelligence typically preempts the evolution of further such creatures, then evolved civilizations will mostly find themselves without comparably intelligent neighbours, even if such evolution is relatively easy.[35] Accurate estimation of the rate at which human-level intelligence evolves from a given starting point would involve the same need for Bayesian correction found in analysis of disasters that would have caused human extinction.[36]

In summary, by looking at instances of convergent evolution on Earth, we can refute claims that certain evolutionary innovations—those for which we have examples of convergence—are exceedingly difficult, given certain assumptions about their underlying mechanisms. Although this method leaves open several ways in which the evolution of intelligence could in principle have been exceedingly difficult, it narrows the range of possibilities. In particular, it provides disconfirming evidence against hypotheses of high evolutionary difficulty between the development of primitive nervous systems and those fairly complex brains providing the advanced cognitive capabilities found in corvids, elephants, dolphins, etc. Insofar as we think evolutionary innovations relevant to AI design will disproportionately have occurred after the development of nervous systems, the evidence from convergent evolution remains quite significant.

To reach beyond the period covered by convergent evolution, and to strengthen conclusions about that period, requires other lines of evidence.

6. SSA and clues from evolutionary timing
The Earth is approximately 4.54 billion years old. Current estimates hold that the expansion of the Sun will render the Earth uninhabitable (evaporating the oceans) in somewhat more than a billion years.[37] Assuming that no other mechanism reliably cuts short planetary windows of habitability, human-level intelligence could have evolved on Earth hundreds of millions of years later than it in fact did.[38]

---

[35] Preemption might not occur if, for instance, most evolved intelligences were unable to manipulate the world well enough to produce technology. However, most of the examples of convergent evolution discussed above do have manipulators capable of tool use, with the possible exception of cetaceans (whales and dolphins).

[36] See Cirkovic, Sandberg, and Bostrom (2010) for an explanation of this correction.

[37] See Dalrymple (2001) and Adams and Laughlin (1998).

[38] The one-billion-year figure is best seen as an upper bound on the remaining habitability-window for the Earth. It is quite conceivable (had humans not evolved) that some natural process or event would have slammed this window shut much sooner than one billion years from now, especially for large mammalian-like life forms. However, there are grounds for believing that whatever the fate would have been for our own planet, there are many Earth-like planets in the universe whose habitability window exceeds 5 billion years or more. The lifetime and size of the habitable zone depends on the mass of the

Combined with a principle such as SSA, this evidence can be brought to bear on the question of the evolvability of intelligence.

6.1 Uninformative priors plus late evolution suggest intelligence is rare

Brandon Carter (1983, 1989) has argued that the near-coincidence between the time it took intelligence to evolve on Earth, and Earth's total habitable period, suggests that the chances of intelligent life evolving on any particular Earth-like planet are in fact far below 1.[39]

Following Carter, let us define three time intervals: $\bar{t}$, "the expected average time ... which would be intrinsically most likely for the evolution of a system of 'intelligent observers', in the form of a scientific civilization such as our own" (Carter 1983, p. 353); t_e, the time taken by biological evolution on this planet $\approx 4 \times 10^9$ years; and t_0, the period during which Earth can support life $\approx 5.5 \times 10^9$ years using the above estimate.

Carter's argument then runs roughly as follows: at our present stage of understanding of evolutionary biology, we have no real way of directly estimating $\bar{t}$. Also, there is no a priori reason to expect $\bar{t}$ to be on the same timescale as t_0. Thus, we should use a very broad starting probability distribution over $\bar{t}$ — a distribution in which only a small portion of the probability mass is between 10^9 years and 10^{10} years, leaving a large majority of the probability mass in scenarios where either: (a) $\bar{t} \ll t_0$, or (b) $\bar{t} \gg t_0$.

Carter suggests that we can next rule out scenarios in which $\bar{t} \ll t_0$ with high probability, since if technological civilizations typically take far less than 4×10^9 years to evolve, our observations of finding ourselves as the first technological civilization on Earth, recently evolved at this late date, would be highly uncommon. This leaves only scenarios in which either $t \approx t_0$ (a small region), or $\bar{t} \gg t_0$ (a large region). Due to observer-selection effects, intelligent observers under either of these scenarios would observe that intelligent life evolved within their own world's habitable period (even if, as in Hard Intelligence, $\bar{t}$ is many orders of magnitude larger than t_0). Thus, at least until updating on other information, we should deem it likely that the chance of intelligent life evolving on our planet within the sun's lifetime was very small.

---

star. Stellar lifetimes scale as $M^{-2.5}$ and their luminosity as $M^{3.5}$, where M is their mass in solar masses; see Hansen et al. (1994). Thus, a star 90% of the sun's mass would last 30% longer and have a luminosity of 70%, allowing an Earth analogue to orbit 0.83 AU from the star with the same energy input as Earth. The interaction between stellar mass and the habitabile zone is more complex, requiring assumptions about climate, but models typically find that the timespan a terrestrial planet can remain habitable is greater for less heavy stars, with increases of several billion years for stars only marginally less massive than the sun; see Kasting et al. (1993) and Lammer et al. (2009). Lighter stars are considerably more common than heavier stars. Sunlike G-class stars of 0.8-1.04 solar masses make up only 7.6% of main-sequence stars, while the lighter 0.45-0.8 solar mass K-class stars make up 12.1%, and the even lighter M-class red dwarfs 76.45%; see LeDrew (2001). A randomly selected planet will therefore be more likely to orbit a lighter star than the sun, assuming the number of planets formed per system is not vastly different G-class and K-class stars.

[39] This section draws on a discussion in Bostrom (2002a).

6.2 Detailed timing suggests that there are fewer than eight "hard steps"

However, knowledge of the Earth's habitable lifetime can also be used to attempt to place probabilistic upper bounds on the the number of improbable "critical" steps in the evolution of humans. Hanson (1998b) puts it well:

> Imagine that someone had to pick five locks by trial and error (i.e., without memory), locks with 1, 2, 3, 4, and 5 dials of ten numbers each, so that the expected time to pick each lock was .01, .1, 1, 10, and 100 hours respectively. If you had just a single (sorted) sample set of actual times taken to pick the locks, say .00881, .0823, 1.096, 15.93, and 200.4 hours, you could probably make reasonable guesses about which lock corresponded to which pick-time. And even if you didn't know the actual difficulties (expected pick times) of the various locks, you could make reasonable guesses about them from the sample pick-times.
>
> Now imagine that each person who tries has only an hour to pick all five locks, and that you will only hear about successes. Then if you heard that the actual (sorted) pick-times for some success were .00491, .0865, .249, .281, and .321 hours, you would have a harder time guessing which lock corresponds to which pick-time. You could guess that the first two times probably correspond to the two easiest locks, but you couldn't really distinguish between the other three locks since their times are about the same. And if you didn't know the set of lock difficulties, these durations would tell you very little about the hard lock difficulties.
>
> It turns out that a difficulty of distinguishing among hard steps is a general consequence of conditioning on early success. … For easy steps, the conditional expected times reflect step difficulty, and are near the unconditional time for the easiest steps. The conditional expected times for the hard steps, on the other hand, are all pretty much the same.

For example, even if the expected pick-time of one of the locks had been a million years, you would still find that its average pick-time in successful runs is closer to .2 or .3 than to 1 hour, and you wouldn't be able to tell it apart from the 1, 10, and 100 hours locks. Perhaps most usefully, Carter and Hanson argue that the expected time between the picking of the last lock and the end of the hour has approximately the same time distribution as the expected time between one "hard step" and another.[40] Therefore, if we knew the "leftover" time L at the end of the final lock-picking, but did not know how many locks there had been, we would be able to use that knowledge to rule out scenarios in which the number of "hard steps" was much larger than $n = 1$ hour $/ L$.

Thus, to start with the simplest model, if we assume that the evolution of intelligent life requires a number of steps to occur sequentially (so that, e.g., nervous systems have no chance of evolving until multicellularity has evolved), that *only* sequential steps are needed, that some of these steps are "hard steps" in the sense that their expected average time exceeds Earth's total habitable period (assuming the steps' prerequisites are already in place, and in the absence of observer-selection effects), and that these

---

[40] Carter (1983) proves this analytically for the special case in which all hardsteps are of the same difficulty; Hanson (1998b) verifies, via Monte Carlo simulations, that it approximately holds with hard steps of varied difficulties. Aldous (2010) makes several additional generalizations of the result.

"hard steps" have a constant chance of occurring per time interval,—then we can use the gap t_e and t_0 to obtain an upper bound on the number of hard steps. Carter estimates this bound to be 3, given his assumption that t_0 is about 10^{10} years (based on earlier longer estimates of the habitable period). Hanson (1998b), using a model with only hard sequential steps (with constant unconditional chance of occurrence after any predecessor steps), calculates that with 1.1 billion years of remaining habitability (in accordance with more recent estimates) and $n = 7$, only 21% of planets like Earth with evolved intelligence would have developed as early.

The same bounds hold (more sharply, in fact) if some steps have a built-in time lag before the next step can start (e.g., the evolution of oxygen-breathing life requiring an atmosphere with a certain amount of oxygen, produced by anaerobic organisms over hundreds of millions of years). These bounds are also sharpened if some of the steps are allowed to occur in any order (Carter, 1983). Thus, the "hard steps" model rules out a number of possible "hard intelligence" scenarios: evolution typically may take prohibitively long to get through certain "hard steps", but, between those steps, the ordinary process of evolution suffices, even without observation selection effects, to create something like the progression we see on Earth. If Earth's remaining habitable period is close to that given by estimates of the sun's expansion, observation selection effects could not have given us hundreds or thousands of steps of acceleration, and so could not, for example, have uniformly accelerated the evolution of human intelligence across the last few billion years in the model.[41] [42]

---

[41] As with the evidence from convergent evolution, there are caveats about the types of scenario this evidence can disconfirm. While the Hanson and Carter models have been extended to cover many branching possible routes to intelligence, the extended models still do not allow us to detect a certain kind of rapid "dead-end" that preempts subsequent progress. For example, suppose that some early nervous system designs are more favorable to the eventual evolution of human-level intelligence, but that whichever nervous system arises first will predominate, occupying the ecological niches that might otherwise have allowed a new type of nervous system to emerge. If the chance of developing any nervous system is small, making it a hard step, then the dead-end possibility does not affect the conclusions about planets with intelligent life. However, if the development of nervous systems occurs quickly with high probability, but producing nervous systems with the right scalability is improbable, then this will reduce the proportion of planets that develop human-level intelligence without affecting the typical timelines of evolutionary development on such planets. Conceivably, similar dead-ends could afflict human engineers—although humans are better able to adopt new approaches to escape local optima.

[42] One objection to the use of the Carter model is that it assumes hard steps are permanent. But in fact, *contra* the model, the organisms carrying some hard step innovation could become extinct, e.g. in an asteroid bombardment. If such events were to frequently "reset" certain hard steps, then scenarios with long delays between the first resettable step and the evolution of intelligent life would be less likely. Success would require both that hard steps be achieved and that disasters not disrupt hard steps in the interim (either via lack of disasters, or through relevant organisms surviving). With a near-constant probability ($1-p$) of relevant catastrophe per period, the probability of avoiding catastrophe for a duration of time t would be p^t. Thus, all else equal, allowing for the possibility that hard steps are not permanent reduces the expected time to complete the resettable steps, conditioning on successful evolution of intelligent life. Carter's model would then underestimate the number of hard steps. However, this

Moreover, in addition to providing information about the total number of hard steps, the model can also give probabilistic bounds on how many hard steps could have occurred in any given time interval. For example, it would allow us to infer with high confidence that at most one hard step has occurred in the 6 million years since the human/chimp common ancestors. This may help narrow the bounds on where AI engineering difficulty can be found.

7. Conclusions

Proponents of the evolutionary argument for AI have pointed to the evolution of intelligent life on Earth as a ground for anticipating the creation of artificial intelligence this century. We explicated this argument in terms of claims about the difficulty of search for suitable intelligent cognitive architectures (the argument from problem difficulty) and, alternatively, in terms of claims about the availability of evolutionary algorithms capable of producing human-level general intelligence when implemented on realistically achievable hardware (the argument from evolutionary algorithms).

The argument from evolutionary algorithms requires an estimate of how much computing power it would take to match the amount of optimization power provided by natural selection over geological timescales. We explored one way of placing an upper bound on the relevant computational demands and found it to correspond to more than a century's worth of continuing progress along Moore's Law—an impractically vast amount of computing power. Large efficiency gains are almost certainly possible, but they are difficult to quantify in advance. It is doubtful that the upper bound down calculated in our paper could be reduced sufficiently to enable the argument from evolutionary algorithms to succeed.

The argument from problem difficulty avoids making specific assumptions about amounts of computing power required or the specific way that human-level machine intelligence would be achieved. This version of the argument replaces the quantitative premise about computational resource requirements and evolutionary algorithms with a more intuitive appeal to the way that evolutionary considerations might tell us something about the problem difficulty of designing generally intelligent systems. But in either of its two versions, the evolutionary argument relies on another assumption: that the evolution of human intelligence was not exceedingly hard (2').

We pointed out that this assumption (2') cannot be directly inferred from the fact that human intelligence evolved on Earth. This is because an observation selection effect guarantees that all evolved human-level intelligences will find that evolution managed to produce them—independently of how difficult or improbable it was for evolution to produce this result on any given planet.

---

underestimation is less severe when we account for the fact that longer time periods allow more chances for the hard steps to occur. The chance of completing n hard steps in time t is proportional to t^n. So when we examine planets where intelligence evolved we would tend to find intelligence evolving after a longer-than-usual gap between disasters, or a series of disasters which spared some lineages embodying past hard steps. Over long time scales, the exponential increase of the catastrophe effect will dominate the polynomial increase from more opportunities for hard steps. However, with plausible rates of catastrophes, the effect of increased opportunity greatly blunts the objection.

We showed how to evaluate the possible empirical support for (2') from the alternative standpoints of two leading approaches to anthropic reasoning: the Self-Sampling Assumption and the Self-Indication Assumption. The Self-Indication Assumption strongly supports the evolutionary argument and the feasibility of AI (among other strong and often counter-intuitive empirical implications). The implications of the Self-Sampling Assumption depend more sensitively on the details of the empirical evidence. By considering additional information about the details of evolutionary history—notably convergent evolution and the timing of key innovations—the Self-Sampling Assumption can be used to make probabilistic inferences about the evolvability of intelligence and hence about the soundness of the evolutionary argument.

This further SSA analysis disconfirms many particular ways in which the evolution of intelligence might be hard, especially scenarios of extreme hardness (with many difficult steps), thus supporting premise (2') (although Carter's model also has implications counting against very easy evolution of intelligence). Of particular interest, two lines of evidence count against extreme evolutionary hardness in developing human-level intelligence given the development of nervous systems as we know them: fairly sophisticated cognitive skills convergently evolved multiple times from the starting point of the earliest nervous systems; and "hard step" models predict few sequential hard steps in our very recent evolutionary history. Combined with the view that evolutionary innovations in brain design especially are diagnostic of AI design difficulty, these observations can avert some of the force of the objection from observation selection effects.

Thus, with one major approach to anthropic reasoning (SIA) providing strong support for (2'), and the other (SSA) offering a mixed picture and perhaps moderate support, observation selection effects do not cripple the evolutionary argument via its premise (in either of its versions) of non-hard evolution of intelligence.

Extensive empirical and conceptual uncertainty remains. Further progress could result from several fields. Computer scientists can explore optimal environments for the evolution of intelligence, and their computational demands—or how easily non-evolutionary programming techniques can replicate the functionality of evolved designs in various domains. Evolutionary biologists and neuroscientists can untangle questions about evolutionary convergence and timing. And physicists, philosophers, and mathematicians can work to resolve the numerous open questions in observational selection theory. Considering how recent many of the relevant ideas and methodologies are, and the character of results thus far obtained, it seems likely that further epistemic truffles are to be found in these grounds.[43]

References

Adams, F.C. & Laughlin, G. (1998) The Future of the Universe, *Sky and Telescope*, **96** (2), p. 32.

Aldous, D.J. (2010) The Great Filter, Branching Histories and Unlikely Events. [http://www.stat.berkeley.edu/~aldous/Papers/GF.pdf]

Archibald, J.D. (2003), Timing and biogeography of the eutherian radiation: fossils and molecules compared, *Molecular Phylogenetics and Evolution*, **28 (**2), pp. 350-359.

Baum, E. (2004) *What is Thought?*, MIT Press.

Benton, M.J. & Ayala, F.J. (2003) Dating the Tree of Life, *Science*, **300** (5626), pp. 1698-1700.

Bostrom, N. (1996) Investigations into the Doomsday Argument. [http://anthropic-principle.com/preprints/inv/investigations.html]

Bostrom, N. (2001) The Doomsday Argument, Adam & Eve, UN++, and Quantum Joe, *Synthese*, **127** (3), pp. 359-387.

Bostrom, N. (2002a) *Anthropic Bias: Observation Selection Effects in Science and Philosophy*. New York: Routledge.

Bostrom, N. (2002b) Self-Locating Belief in Big Worlds: Cosmology's Missing Link to Observation, *Journal of Philosophy*, **99** (12), pp. 607-623.

Bostrom, N. & Cirkovic, M. (2003) The Doomsday argument and the Self-Indication Assumption, *Philosophical Quarterly*, **53** (210), pp. 83-91.

Bostrom, N. (2003) Are You Living In a Computer Simulation?, *Philosophical Quarterly*, **53** (211), pp. 243-255.

Bostrom, N. (2005) The Simulation Argument: Reply to Weatherson, *Philosophical Quarterly*, **55** (218), pp. 90-97.

Bostrom, N. (2007) Observation selection effects, measures, and infinite spacetimes, in Carr, B. (ed.) *Universe or Multiverse?*, Cambridge: Cambridge University Press.

Bostrom, N. & Sandberg, A. (2009) The Wisdom of Nature: An Evolutionary Heuristic for Human Enhancement, in Savulescu, J. and Bostrom, N. (eds.) *Human Enhancement*, Oxford: Oxford University Press.

Bostrom, N. & Kulczycki, M. (2011) A Patch for the Simulation Argument, *Analysis*, **71** (1), pp. 54-61.

Byrne, R.W., Bates, L.A. & Moss, C.J. (2009) Elephant cognition in primate perspective, *Comparative Cognition & Behavior Reviews*, **4**, pp. 65–79.

Carter, B. (1983) The anthropic principle and its implications for biological evolution, *Phil. Trans. R. Soc. Lond. A*, **310** (1512), pp. 347-363.

Carter, B. (1989) The anthropic selection principle and the ultra-Darwinian synthesis, in Bertola, F. and Curi, U. (eds) *The Anthropic Principle*, Cambridge: Cambridge University Press.

Chalmers, D.J. (2005) The Matrix as metaphysics, in Grau, C. (ed.) *Philosophers Explore the Matrix*, Oxford: Oxford University Press.

Chalmers, D. (2010) The Singularity: A Philosophical Analysis, *Journal of Consciousness Studies*, **17** (9-10), pp. 7-65.

Çirkoviç, M.M., Sandberg, A. & Bostrom, N., (2010) Anthropic Shadow: Observation Selection Effects and Human Extinction Risks, *Risk Analysis*, **30** (10), pp. 1495–1506.

Dalrymple, G.B. (2001) The age of the Earth in the twentieth century: a problem (mostly) solved, *Special Publications, Geological Society of London*, **190** (1), pp. 205–221.

Dieks, D. (2007) Reasoning About the Future: Doom and Beauty, *Synthese*, **156** (3), pp. 427-439.

Elga, A. (2000) Self-locating Belief and the Sleeping Beauty problem, *Analysis*, **60** (2), pp. 143-147.

Emery, N.J., & Clayton, N.S. (2004). The Mentality of crows: convergent evolution of intelligence in corvids and apes. Science, 306(5703), 1903 - 1907.

Erwin, D.H. & Davidson, E.H. (2002) The last common bilaterian ancestor, *Development*, **129** (13), pp. 3021–3032.

Finn, J.K., Tregenza, T., & Norman, M.D. (2009) Defensive tool use in a coconut-carrying octopus, *Current Biology*, **19** (23), pp. R1069-R1070.

Goodman, M. et al. (2009) Phylogenomic analyses reveal convergent patterns of adaptive evolution in elephant and human ancestries, *Proceedings of the National Academy of Sciences*, **106** (49), pp. 20824-20829.

Grace, C. (2010) *Anthropic Reasoning in the Great Filter*. BSc(Hons). Australian National University.

Groombridge, B. and Jenkins, M.D. (2000) *Global biodiversity: Earth's living resources in the 21st century*, Cambridge: World Conservation Press.

Hansen, C.J. & Kawaler, S.D. (1994) *Stellar Interiors: Physical Principles, Structure, and Evolution*, Birkhäuser.

Hanson, R. (1998a) The great filter—are we almost past it?.
[http://hanson.gmu.edu/greatfilter.html]

Hanson, R. (1998b) Must early life be easy? The rhythm of major evolutionary transitions.
 [http://hanson.gmu.edu/hardstep.pdf]

Hasegawa, M., Kishino, H. & Yano, T. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA, *Journal of Molecular Evolution*, **22**, pp. 160-174.

Hawks, J. et al (2007) Recent acceleration of human adaptive evolution, *Proc Natl Acad Sci*, **104** (52), pp. 20753-8.

Hochner, B., Shomrat, T. & Fiorito, G. (2006) The Octopus: A Model for a Comparative Analysis of the Evolution of Learning and Memory Mechanisms, *Biol. Bull.*, **210**, pp. 308-317.

Jacobsen, S.B. (2003) How Old Is Planet Earth? *Science,* **300,** pp. 1513-1514.

Kasting, J.F., Whitmire, D.P. & Reynolds, R.T. (1993) Habitable Zones around Main Sequence Stars, *Icarus* **101**, pp. 108-128.

Lammer, J.H. et al. (2009) What makes a planet habitable? *Astron Astrophys Rev* **17**, pp. 181–249.

LeDrew, G. (2001) The Real Starry Sky, *Journal of the Royal Astronomical Society of Canada*, **95**, No. 1, pp. 32–33.

Legg, S. (2008) *Machine Super Intelligence* (PhD thesis, Department of Informatics, University of Lugano).

Leslie, J. (1993) Doom and Probabilities, *Mind*, **102** (407), pp. 489-91. [http://www.jstor.org/pss/2253981]

Lewis, D.K. (2001) Sleeping Beauty: reply to Elga, *Analysis*, **61** (271), pp. 171-176.

Li, Y., Liu, Z., Shi, P. & Zhang, J. (2010) The hearing gene *Prestin* unites echolocating bats and whales. *Current Biology*, **0** (2), pp. R55-R56.

MacKay, D.J.C. (2009) *Information theory, inference, and learning algorithms*, Cambridge : Cambridge Univ. Press.

Mather, J.A. (1994) 'Home' choice and modification by juvenile Octopus vulgaris (Mollusca: Cephalopoda): specialized intelligence and tool use? *Journal of Zoology*, **233**, pp. 359-368.

Mather, J.A. (2008) Cephalopod consciousness: Behavioural evidence, *Consciousness and Cognition*, **17** (1), pp. 37-48.

Menzel, R. and Giurfa, M. (2001) Cognitive architecture of a mini-brain: the honeybee, *Trends Cog. Sci.* **5** (2), p. 62.

Moravec, H. (1976) The role of raw power in intelligence. [http://www.frc.ri.cmu.edu/users/hpm/project.archive/general.articles/1975/Raw.Power.html]

Moravec, H. (1998) When will computer hardware match the human brain?, *Journal of Transhumanism*, **1**.

Moravec, H. (1999) *Robots: Mere Machine to Transcendent Mind*, Oxford: Oxford University Press.

Neal, R.M. (2007) Puzzles of Anthropic Reasoning Resolved Using Full Non-indexical Conditioning, *Technical Report No. 0607,* Department of Statistics, University of Toronto. [http://www.cs.toronto.edu/~radford/ftp/anth2.pdf].

Olum, K.D. (2002) The doomsday argument and the number of possible observers, *Philosophical Quarterly*, **52** (207), pp. 164-184.

Orgel, L.E. (1998) The origin of life—a review of facts and speculations, *Trends in Biochemical Sciences*, **23** (12, 1), pp. 491-495.

Piccione, M. & Rubinstein, A. (1997) The absent-minded driver's paradox: Synthesis and responses, *Games and Economic Behavior*, **20** (1), pp. 121-130.

Sabrosky, C.W. (1952) How many insects are there?, in U.S. Dept. of Agr. (eds) *Insects: The Yearbook of Agriculture*, Washington, D.C.: U.S.G.P.O.

Sandberg, A. & Bostrom, N. (2008) Whole brain emulation: A roadmap. *Technical report 2008–3*, Future for Humanity Institute, Oxford University. [http://www.fhi.ox.ac.uk/Reports/2008-3.pdf]

Schopf, W.J. (ed.) (1992) *Major Events in the History of Life*, Boston: Jones and Barlett.

Schröder, K.P. & Connon Smith, R. (2008) Distant future of the Sun and Earth revisited, *Monthly Notices of the Royal Astronomical Society,* **386** (1), pp.155–163.

Schultz, T.R. (2000) In search of ant ancestors, *Proceedings of the National Academy of Sciences*, **97** (26), pp. 14028–14029.

Truman, J.W., Taylor, B.J., and Award, T.A. (1993) Formation of the adult nervous system, in Bate, M., and Arias, A.M. (eds.) *The Development of Drosophila Melanogaster*, Cold Spring Harbor: Cold Spring Harbor Laboratory Press.

Whitman, W.B., Coleman, D.C., and Wiebe, W.J. (1998) Prokaryotes: the unseen majority, *Proceedings of the National Academy of Sciences*, **95** (12), pp. 6578–6583.