Revisiting the Evolution Anchor in the Biological Anchors Report

A review of criticisms and an alternative estimate based on the thermodynamic approach

This document is the work of Janvi Ahuja and Victoria Schmidt as part of the Epoch FRI Mentorship Programme 2023. We worked on for ~10 hours a week for two months.Tegan McCaslin mentored the project, and Rose Hadshar and Angelina Li provided significant feedback and advice as our peer reviewers.

The Forecasting TAI timelines with biological anchors report produces an estimate of the compute needed to develop a transformative model using 2020 architectures and algorithms. It uses six different biological frameworks to estimate the compute needed to develop a transformative model, one of which is an evolution based framework. The evolution anchor estimates the amount of computation done by all animals throughout evolution, from the earliest animals with neurons to modern day humans. In this report, we look into criticisms of the evolution anchor and summarise their effect sizes. We then expand on one such criticism and discuss some of our own criticisms with the biological anchors framework as a whole. The executive summary is a two page summary of the whole report.

This report may be useful to you if you:

- Are interested in biological anchors and defer to the report to determine your AI timelines (and put some weight on the evolutionary anchor). In this case I would recommend reading the executive summary and reading further on areas of interest. Note, Cotra has posted an update to her original draft and has been interviewed more recently on her timelines
- Are interested in the upper bound estimate of the biological anchors report and want to investigate the most conservative anchor, or otherwise particularly interested in evolutionary anchor. In this case, I might recommend reading the whole report

Executive summary

Motivation statement

- The biological anchors report has influenced many views on when TAI will be developed. One survey by Clarke and McCaffary found it was the second most cited source that is deferred to on TAI timelines (where the first is “inside view”).
- At the outset of the fellowship, our goal was to expand on Nuño Sempere’s criticism pertaining to the cost of simulating the environment, but we found this difficult to make traction on (see here)
- Instead, we decided to collate all the criticisms on the evolutionary anchor and their effect sizes.
- We also decided to expand on one of the alternative approaches to the evolutionary anchor as we found a way to.

What we did

- Summarised critiques of the evolution anchor
- Proposed a best-guess for an upper bound based on the thermodynamic approach
- Proposed broader criticisms

Summary of criticisms and their effect sizes

Critique and approach to incorporate this into the evolution anchor	Expected effect size	Updated evolution anchor	Reference
Original estimate	*N/A*	*1E41 FLOP*	*Ajeya Cotra*
Environment simulation: Add costs of simulating an environment and coupling architectures with that environment	Upwards: not quantified	N/A	Jennifer Lin
Environment simulation: Add environmental simulation cost to the original estimate	+5E27 - ≥4E29 FLOP	1E41 FLOP	Nuño Sempere
Environment simulation: Simulate whole Earth *molecular simulation*	1E60 FLOP	1E60 FLOP	meanderingmoose
Environment simulation: Simulate whole Earth *thermodynamic approach*	1E45 FLOP	1E45 FLOP	Ege Erdil
Anthropic critiques	+up to 6 OOM	1E41 - 1E47 FLOP	Ege Erdil
Paradigm shift	possibly +>>30 years	N/A	Jennifer Lin
	drastically shortened timelines; not quantified here	discard bio anchors completely	Elizier Yudkowsky
Missing architecture search space	Upwards; not quantified	N/A	Jennifer Lin
Evolutionary algorithms are inefficient	Downwards; not quantified	decreasing the weight of the evolution anchor to 3%	Marius Hobbhahn

Proposing a best-guess for an upper bound to the thermodynamic approach

- The thermodynamic approach estimates the total amount of energy received by the earth from the Sun and converts this into FLOP. Erdil used the Landauer principle for this conversion, which is the theoretical lower limit of energy consumption of computation. As we expect that most energy was not converted as efficiently as the theoretical lower limit we propose two alternatives:
    - Using the conversion rate it takes for the brain to convert joules into FLOP
    - Using the conversion rate it takes for the human body to convert joules into FLOP
- As we can expect the average energy-to-information processor to be less efficient than the human brain or body, we expect this is still a conservative upper limit. Our model in the form of a google sheet is available here.

- These approaches result in upper bound estimates which are 4-6 OOMs smaller than the Landauer's principle approach. In addition, it's worth noting that the upper bounds for both of these estimates are lower than (though not far from) Cotra's estimate (1E41), at 2E40 for the caloric approach and 6E40 for the brain energy consumption approach.
- We developed more informative estimate for the evolution anchor, narrowing down the range to 2E40 FLOP as an upper bound (original anchor was 1E41 FLOP.
- We expand on fundamental issues we have with the evolution anchor and how it is derived in Ajeya Cotra's report. These include:
    - Weighing FLOP estimates against developments in compute capacity
    - Noting that bounding parameter estimates is difficult and sometimes arbitrary
    - Noting that some of the parameter ranges vary by many orders of magnitude
    - Noting that FLOP conversion to intelligence is abstract and weird
- Finally, we provide an overview of what our work could mean for TAI timelines estimates. This is shown below:
    - If you:
        - Believe that Cotra's model and framework for calculating TAI is reasonable
        - Believe that the thermodynamic estimate for the evolution anchor is better than the brain computation method Cotra uses, and
        - Believe that our best guess proposal for an improvement upon the thermodynamic estimate is better than the original Landauer approach
            - You might change your best guess for a FLOP estimate for an upper bound to be 1.79E40 FLOP instead of 1e41 FLOP.
    - If you:
        - Believe one of the other criticisms/adjustments noted in section two is legitimate, you might
            - Use the naive estimated effect sizes to update your estimate of FLOP needed to develop TAI
            - Choose investigate it further and update accordingly
            - Consider Nuño Sempere's propagation of beliefs given additional uncertainty surrounding the evolution anchor
    - If you:
        - Believe any of the major reasons to be sceptical listed above you might
            - Downweight on the evolution anchor and the entire biological anchors report
            - Consider other ways to assess AI timelines
            - Consider Nuño Sempere's propagation of beliefs given additional uncertainty surrounding the evolution anchor

1. Introduction

Background

- The evolution anchor is one of six biological anchors used in the TAI[1] timelines report by Ajeya Cotra, which estimates when the computation required to train a TAI model will become affordable. It then uses these estimates to arrive at a probability of training a transformative model in each year beyond 2020.
- Cotra takes the human brain as proof of concept for training a general intelligence (GI) and comes up with 6 different biological anchors to estimate the computation required to train a human brain. These anchors are: Neural network anchors with three different horizon lengths (short, medium, long), the Genome anchor, the Lifetime anchor, and the Evolution anchor.
- Figure 1 provides an overview how the estimates made by Cotra sum up to her final model. It is modified from Anson Ho's summary of the report.

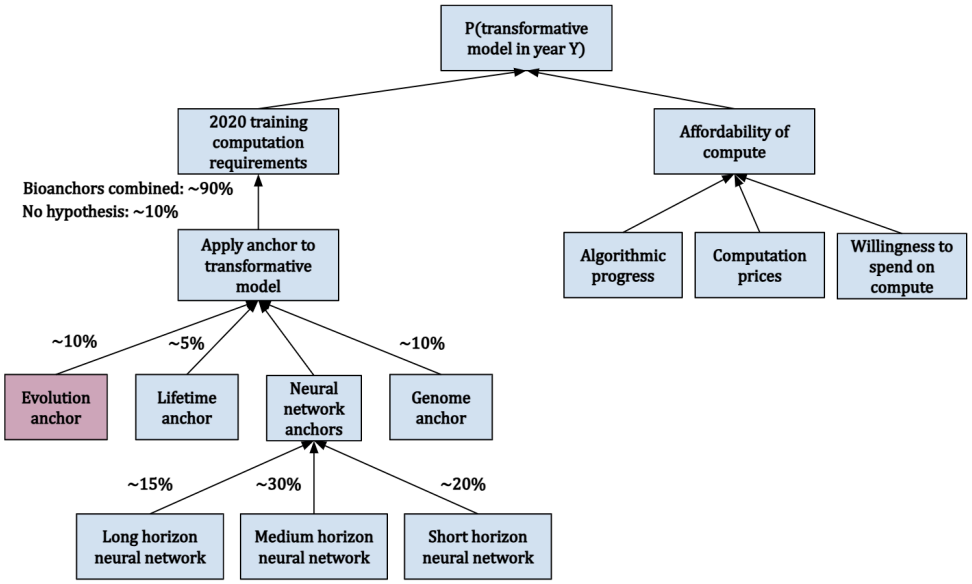

Figure 1: Visualising inputs of the P(transformative model in year Y) estimate made in Cotra's report. Taken and modified from Ho's report. Focus of this report is highlighted in pink.

- In this report, we expand on the Evolution anchor, which aims to calculate all computation[2] performed by neurons (measured in FLOP) in the history of evolution. Evolution produced human intelligence, the only development of GI that we are currently aware of. The argument here is that architectures and optimization algorithms developed by humans are at least as efficient as natural selection. So, the amount of compute that evolution required in order to develop human intelligence serves as a soft upper bound to what humans should need in order to develop a TAI.

[1] TAI: transformative artificial intelligence
[2] Using an estimate for the computational capacity of the human brain from "How Much Computational Power Does It Take to Match the Human Brain?" by Joe Carlsmith.

- Cotra uses the following parameters for her calculation:
    - The first neurons developed about 1E9 years ago, or **1E16** seconds.
    - The "average ancestor" performed about as many FLOP/s as a nematode, **~1E4** FLOP/s.
    - The "average population size" was about **1E21** individuals at any given time.
- This results in the following estimate for FLOP used in evolution
    - (1E16 seconds) * (1E4 FLOP/s*1E21) = **1E41 FLOP**
- Note that are several assumptions the evolutionary anchor approach makes:
    - Human-level intelligence ≈ transformative intelligence[3]
    - Transformative artificial intelligence would take the same amount of computation as human-level intelligence to develop, or less
    - "What it took" to develop human intelligence can be measured by all the brain activity carried out by all neuron-having organisms that have lived so far
    - Brain activity can be accurately modelled as simulated neuron FLOP and is done reasonably
- Cotra gives each of her six anchors a weighting to form her distribution of the training requirements for a transformative model. The evolution anchor is weighted at 10%, and estimates ~1e41 FLOP are needed to train a transformative model. As the evolution anchor serves as an upper bound for computational requirements, attributing it a higher weight results in higher probabilities for longer timelines, while downweighting it shortens expected timelines. Figure 2 visualises the weighting of each anchor and the FLOP each one stipulates.

[3] Cotra's definition of a TAI is a "*"software" (i.e. a computer program or collection of computer programs) that has at least as profound an impact on the world's trajectory as the Industrial Revolution did*." For more, see "How do we define transformative AI" and "transformative model"?, part 1 of Cotra's report.

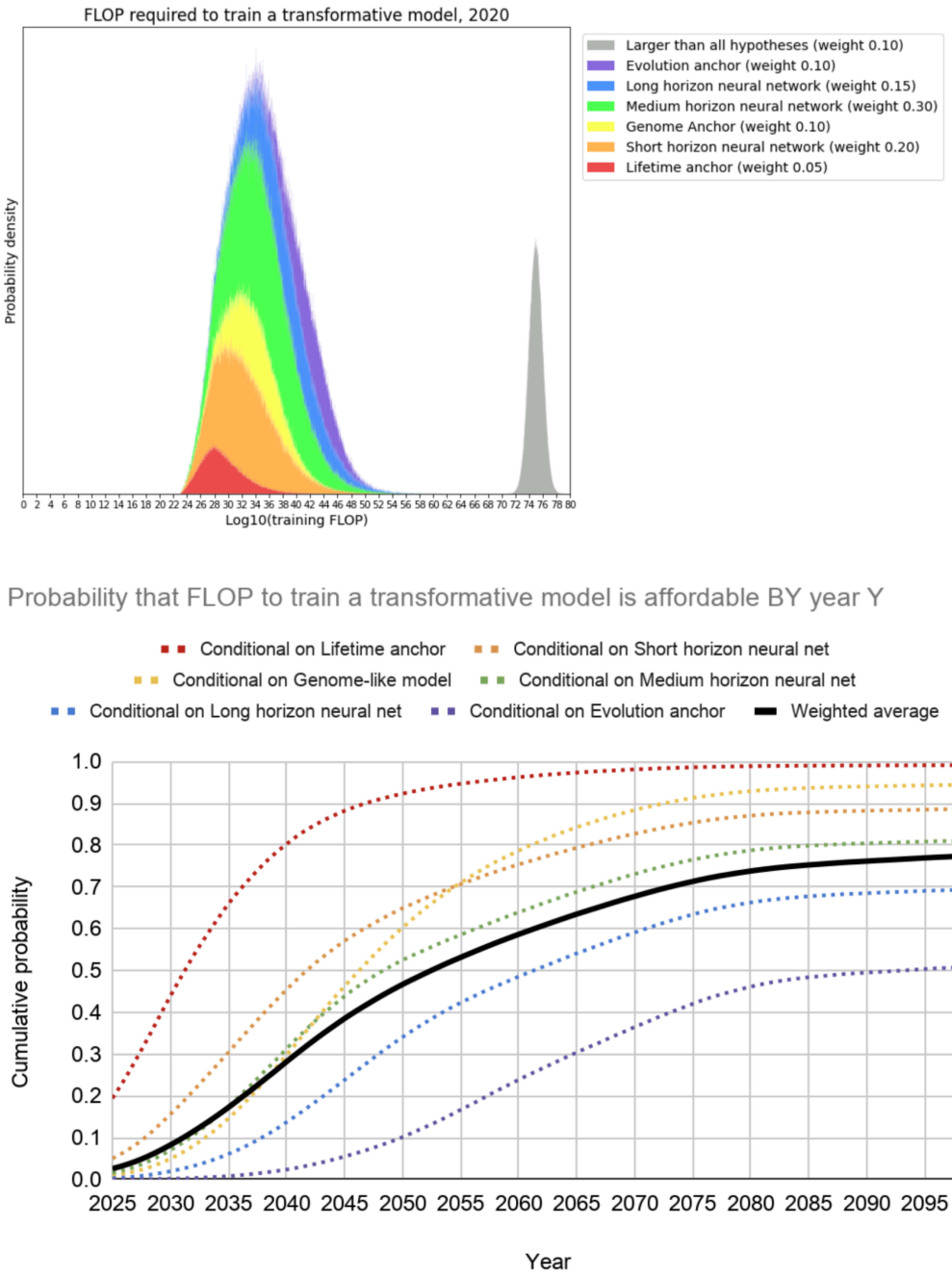

Figure 2: Distribution of the weights assigned to each anchor (represented by probability density) alongside the FLOP needed for training a transformative model (from Cotra's report).

- For further detail on each of the anchors and on Cotra's report as a whole, please refer to the report itself (part one linked here) as well as these excellent summaries: Grokking "Forecasting TAI with biological anchors" by Anson Ho and Biological Anchors: A Trick That Might Or Might Not Work by Scott Alexander.

Motivation

- **Why is the biological anchors report important?**
  - The biological anchors report by Ajeya Cotra was written to address a key uncertainty in planning for mitigating risks from advanced artificial intelligence.

Specifically, it was written to understand when TAI will be developed, by understanding when the compute necessary for a transformative model will become affordable.
    - The report has had a substantial effect within the field of AI risk research, influencing many views on when TAI will be developed. One survey by Clarke and McCaffary found it was the second most cited source that is deferred to on TAI timelines (where the first is "inside view"). It should noted that it's likely that most people don't defer to rigid outputs of the model but rough estimates.
    - In addition, it has influenced other models such as this piece by Davidson.
    - It should be noted that I think the biological anchors report is now less important and informative than it was 2 years ago. This is due to the fact we have more evidence of TAI capabilities growing, and don't need a naive prior to the same degree, and can instead use current trends to extrapolate further

- **Why are we focusing on the evolution anchor in particular?**
    - As an upper bound for FLOP required to train a TAI, the evolution anchor informs the conservative timelines estimate for when TAI will be developed. Those who are more sceptical of TAI development may trust values skewed closer to this bound.
    - We noted many criticisms surrounding the evolution anchor in particular, but no summary of these.
    - In addition, we wanted to scrutinise one anchor of the biological anchors report to understand how sensitive it is to subjective considerations.
    - It is worth noting that our original project was to investigate a specific criticism made by Nuño Sempere on the evolution anchor. In particular, that the evolution anchor does not account for the cost of simulating a complex environment, which is part of what it took to develop intelligence on earth. However, we found that this criticism was hard to make meaningful progress beyond Sempere's original estimates[4].

[4] This is because of several reasons: (1) Converting the cost of simulating an environment into FLOP is difficult, as the environment was not generated using operations/exchanges in information. (2) Thus, the path forward instead might have looked like understanding the cost of simulating environments using current day simulation systems in FLOP, and using this to estimate the cost of simulating the Earth's history. We spent ~5 hours in total investigating this, and found it difficult to do. In particular, our uncertainty surrounded scaling of costs— it would involve many, many more dynamic interactions than are currently simulated by systems today, and it wasn't clear to us what the scaling cost of simulation looks like in terms of FLOP. (3) Sempere suggests a method of *"(i) Coming up with estimates of what the least fine-grained world that we would expect might be able to produce intelligent life if we simulated natural selection in it (ii) Calculating how much compute it would take to in fact simulate it".* We spent ~15 hours investigating (i), and found this difficult to make traction on. A complete version of this might involve simulating a reduced size environment, as well as a reduced population (given the reduced environment) and calculating the FLOP it would take to produce this. This seemingly beats the point of the evolution anchor, as we are no longer anchored on *what it in fact took* evolution to develop intelligence, but instead on what we speculate it could possibly take. Given this, we no longer thought this was a valuable path to pursue for the scope of this project. Specifically, we did not think we could make marginal progress on Sempere's initial estimates/bounds in a reliable manner.

2. Summary of criticisms

- We have collated a non-exhaustive list of the most important existing criticisms of the evolution anchor from our perspectives (we spent ~ three hours cumulatively looking for these).
- There are varying suggestions on how to update in response to the arguments made against the evolution anchor or how it is deployed. These include additions to the compute estimates, entirely new ways to derive an evolutionary anchor, and downweighting this particular anchor in the final timelines estimate. Some critiques are confined to qualitative arguments and do not provide quantitative suggestions with regards to the evolution anchor.
- When we reference posts, they may have sections other than the section we discuss in this criticism. Also note, there are other criticisms of the biological anchors report as a whole, which also apply to parameters and mechanisms used by and using the evolutionary anchor. We don't address these here.
- There is an important difference in estimating "what it in fact took evolution to develop frontier human-level intelligence" and "what it would take human beings to simulate human-level intelligence in an evolution-flavoured way using modern AI and simulation technology." This distinction is not explicitly made in the criticisms, but would change how to approach modelling the amount of compute needed to develop human-level intelligence.
- The table below provides an overview of the critiques and their expected effects on the evolution anchor. For more detail on how the numbers were derived, see the corresponding chapter above.

Critique and approach to incorporate this into the evolution anchor	**Expected effect size**	**Updated evolution anchor**	**Reference**
Original estimate		*1E41 FLOP*	*Ajeya Cotra*
Environment simulation: Add costs of simulating an environment and coupling architectures with that environment	Upwards: not quantified	N/A	Jennifer Lin
Environment simulation: Add environmental simulation cost to the original estimate	+5E27 - ≥4E29 FLOP	1E41 FLOP	Nuño Sempere
Environment simulation: Simulate whole Earth *molecular simulation*	1E60 FLOP	1E60 FLOP	meanderingmoose
Environment simulation: Simulate whole Earth *thermodynamic approach*	1E45 FLOP	1E45 FLOP	Ege Erdil
Anthropic critiques	+up to 6 OOM	1E41 - 1E47 FLOP	Ege Erdil

Paradigm shift	possibly >30 years	N/A	Jennifer Lin
	Drastically shortened timelines; not quantified here	Discard bio anchors completely	Elizier Yudkowsky
Missing architecture search space	Upwards; not quantified	N/A	Jennifer Lin
Evolutionary algorithms are inefficient	Downwards; not quantified	Decreasing the weight of the evolution anchor to 3%	Marius Hobbhahn

2.1 Missing environment simulation

- The evolution anchor currently calculates the computational power required to create human intelligence by adding up the activity of all neurons over the course of evolutionary history. However, this model overlooks the role of interactions with the environment and other agents and feedback loops in the development of intelligence.
- Critiques such as the ones made by Jennifer Lin und Nuño Sempere argue that, by not accounting for these environmental factors, the evolution anchor may underestimate the actual computational work that went into shaping human intelligence**.**

2. 1. 1. Adding the cost of simulating the environment to Cotra's brain computation estimate

A Bio anchors Review by Jennifer Lin

- Summary
  - Jennifer Lin's post suggests that in order to rerun evolution, at least three additional things are needed:
    - (1) A complex environment, which mimics the Earth over the course of evolution (ostensibly over time, seasons, biomes).
    - (2) A search space of architectures that simulates the search space of architectures explored by evolution to derive animal brains (see more details on this in Chapter 2.2).
    - (3) A way to couple such architectures to the environment, to mimic environmental selection pressures which led to human-level intelligence (by the simulation of relevant parts of bodies, such as sensors and effectors).
  - Lin notes that she does not believe this will be a technological bottleneck to progress towards developing TAI, due to the state of virtual reality simulation today.
  - However, Lin does note that the compute required to run such a simulation could take up a significant fraction of the total cost of rerunning evolution, but does not suggest how much.
- Thoughts

    - The overall argument seems sound, but lacks any further specifications on what Lin thinks is an appropriate environment (granularity, size, time), how to actually couple architectures/brains and what that would mean in terms of compute power involved.
- Estimated effect size
    - Compute requirements go up for environment simulation, the characterisation of an architecture search space, and for coupling architectures with the environment (e.g. by simulating bodies).
    - Adjust the evolution anchor upward, amount not specified.

A concern about the "evolutionary anchor" of Ajeya Cotra's report by Nuño Sempere

- Summary
    - Nuño Sempere's blog post suggests a similar critique to Lin. He notes some simulated environments which might pose as useful anchors, including Minecraft (~5E11 FLOP/s[5]) and a simulation of Earth with a precision comparable to current atmospheric simulations (~4.1E13 FLOP/s[6]).
- Thoughts
    - It is not clear which level of granularity is appropriate to rerun the evolution of humans.
        - One idea we had was to anchor on how much detail human senses can perceive (or the senses of the human ancestor with the highest sensual capacity).
            - This approach does not account for other species that might not develop into GI themselves, but significantly influence the development of human-level intelligence (e.g. microorganisms).
- Estimated effect size
    - In order to compare the two examples, we chose to compare the peak performances of the hardware used to run Minecraft and atmospheric simulations over the course of 1E9 years (in simulation time).
        - Actual FLOP requirements will be lower, as utilization rate[7] rates do not reach 100% and simulations do not need to be run for 1E9 years in order to simulate 1E9 years.
    - Additional FLOP requirement to run a simulation for 1E9 years is between 5E27 FLOP (run the hardware required for Minecraft for 1E9 years) and 4E29 FLOP (run the hardware used for current atmospheric simulation for 1E9 years; to be expanded to the whole of Earth).

[5] This page claims that recommended requirements are an Intel Core i5 for the CPU, and a Radeon 7 200 for the GPU. The Intel CPU has a processing speed of 37.73 GFLOPs and the Radeon of 467.2 GFLOPs (copied from NuñoSemperes blog post). Note that this is maximum hardware capacity required, not actual performance utilized.

[6] On the AfES (AGCM for Earth Simulator), it takes ~4E16 FLOP to simulate Earth's atmosphere for a day with a run time of 1518 s. The theoretical peak performance of the system is ~4.1E13 FLOP/s with a utilization rate of 58.4%.

[7] In reality, algorithms only use a certain percentage of the maximum hardware capacity. Utilization rates are dependent on algorithms, hardware systems and the compatibility between those.

2. 1. 2. Simulate the Earth from scratch

- Another approach is to estimate the compute needed to simulate the actual quality of the environment for Earth, including the habitat as well as any relevant evolutionary processes, and then make some appropriate downwards adjustments.

[Examining Evolution as an Upper Bound for AGI Timelines] by Meandering moose

- Summary
    - Meandering moose remarks that considering only neural computation implies that this is the only relevant computation done by evolution. Instead, the amount of computation done by evolution includes the entirety of all organisms and their environment. As the DNA is the actually changing unit, they suggest molecular-level simulation in order to get a true upper bound for the computation involved in evolutionary history.
        - They assume, for the purpose of their estimate, that the cost of simulating an atom for one second is equivalent to 1 FLOP (which is probably significantly too low). They consider the cost of this for 1E10 people with 1E28 atoms each, multiplied by 1E6 to account for other lifeforms and the environment. They arrive at a staggering 1E60 FLOP[8] required in order to simulate the last billion years of evolution's history.
        - Meandering moose also notes that the simulation might require more or less granularity, for example at a cellular or quantum level respectively.
- Thoughts
    - Simulation cost of 1 FLOP/s per atom is with high probability significantly underestimating the computational cost of molecular simulation.[9]
    - The factor 1E6 used to capture simulation of the environment and other non-human lifeforms seems arbitrary and is not further explained.
    - Arguably, not all atoms in all humans have been relevant for the development of human-level intelligence.
- Estimated effect size
    - Another approach to estimate an evolution anchor based on molecular-level simulation.
    - 1E10 people * 1E28 atoms each * 1E6 to account for environment/other lifeforms * 1E16 seconds = 1E60 FLOP

Evolution anchor thoughts by Ege Erdil

- Summary

[8] The original report states 1E70 FLOP, but this is probably a typo.
[9] Even if the simulation of one atom for one second took 1 FLOP, simulation costs increase non-linearly with the number of atoms involved. If simulating 1 atom for 1 second on x processors took 1 second, simulating 5 atoms for 1 second would take significantly longer than 5 seconds. Actual compute costs depend on the quantum-mechanical computing method used, but all involve correlation terms accounting for atom-atom-interactions between all atoms. A system that is used to run molecular simulations for scientific research is Cologne University's CHEOPS, a high perfomance cluster with a theoretical peak performance of 1E14 FLOP/s.

- Ege Erdil proposes a “thermodynamic approach”. In this approach, a FLOP estimate is derived by treating the Earth as a computing unit with an energy budget (4E24 joules/year).
    - This energy budget is the balance between the energy Earth receives from the sun and the energy it radiates back to space and is converted into FLOP using Landauer’s principle.
    - Landauer’s principle, or the Landauer limit, is the “theoretical lower limit of energy consumption of computation”.
    - Computation requires energy, and the Landauer limit allows to calculate the maximum amount of computation theoretically achievable with a given amount of energy.
        - Computing processes involve information erasure[10], and information erasure increases entropy. This increase in entropy can be expressed as a rise in temperature and therefore allows the conversion of energy into the maximum amount of compute theoretically doable with that energy.
        - According to the Landauer limit, the lower bound to the energy cost to erase one bit of information is ~3E-23 J (at T = 3 K). With Earth’s energy budget, 1E47 bit erasures can theoretically be performed every year.
    - Erdil then does rough downward adjustments to the actual FLOP available to evolution:
        - Each FLOP likely took evolution 100-1000 bit erasures to perform: this shifts the actual FLOP output downward by 2-3 OOM[11]
        - actual physical systems do not operate at the Landauer limit: 4-5 OOM
        - most of Earth’s energy budget is not used up for evolution: 2-3 OOM
        - Earth’s actual surface area has a temperature of about 300 rather than 3 Kelvin: 2 OOM[12]
        - This adds up to 1E36 FLOP/year, resulting in 1E45 FLOP performed by evolutionary processes over the course of the last billion (or 1E9) years.
- Thoughts
    - This approach seems well bound as it relies on Earth’s energy budget, a number we’re sure of within orders of magnitude.
    - However, it relies on an unusual conversion. In particular, it is unclear what exactly the Landauer limit calculates in regards to evolution. My current best guess is that the Landauer limit conservatively guesses how many information processing bits could have happened on Earth, which likely refers to all neuron brain processing and maybe some other information related tasks.

[10] Reversible computing is an exception.
[11] OOM: orders of magnitude
[12] The Landauer limit is calculated by this formula: E=k*T*ln(2). E is the energy per bit of information, T the temperature an k the Boltzman constant. As T feeds linearly into the equation, an increase of T by 2 OOM (from 3 K to 300 K), results in energy going up by 2 OOM.

    - What's not clear is what the line is between information processing and not-information processing tasks.
        - Naively, it seems that all neuronal activity are information processing tasks. However, it is unclear whether other activities might also be considered as information processing such as biomatter decay, extinction, or the loss of genes in a population.
    - Our current view is that it captures all neuronal activity (as Cotra's model does), plus some significant surplus. This makes it a more reliable upper bound for the same calculation Cotra is making, but note it likely does not address other costs, such as the cost of simulating the environment.
        - The issue raised here is that it might take more energy to simulate a real-world process than it takes for that process to actually occur. This means that in order to simulate Earth, we might need more energy than Earth's energy budget can provide, and can therefore not assume that the computation performable with Earth's energy budget is sufficient to simulate human evolution one-on-one.
    - Given this, we think it is at least an improvement on the Cotra model. We expand on this approach in detail on section 3.
- Estimated effect size
    - Set the evolution anchor at 1E45 FLOP.

2.2 Missing architecture search space

A Bio anchors Review by Jennifer Lin

- Summary
    - Jennifer Lin proposes an alternative way to consider the computational cost of developing intelligence which involves considering the size of the architecture search space.
    - An architecture search space refers to the set of all possible configurations or architectures that can be considered for a particular computational model, such as a neural network. In the case of evolution, it refers to the set of all possible brain configurations that were considered, including all those that were produced, before the human brain was produced. See footnote for more detail on architecture search spaces[13].
    - Lin suggests ascertaining the size of an architecture search space encompasses brain architectures developed by evolution, and then calculating how expensive it would be for a range of optimisation algorithms to search this space to find the human brain or some form of general intelligence.

[13] An architecture search space refers to the set of all possible configurations or architectures that can be considered for a particular computational model. This space outlines the range of variations allowed in the model's design and its architectural elements. The goal of searching through this architecture space is to find the optimal configuration that performs best for a given task on predefined performance metrics.
The architecture search space defines the boundaries and constraints within which the search for an optimal model occurs. It can range from varying the number of neurons in a single layer to allowing for entirely different types of network topologies. Complexity and size of the search space significantly impact the computational resources required to find an optimal model.

    - She suggests that one way to set an upper bound might be by iterating through the space with a brute grid search algorithm.
    - Lin notes many features[14] which might make the search space to find the human brain much larger, and difficult to specify. She notes that the brain's nontrivial topology is a feature that has not been thoroughly explored in neural architecture search (NAS) research yet, and might be a conceptual barrier to determining a search space.
- Thoughts
    - We note that this criticism did not explicitly propose an estimate in how this might change the evolution anchor. We did not have the expertise to quantify this or explore it further.
- Estimated effect size
    - Not quantified.

2.3 Anthropic critiques

- The anthropic argument elaborates on the fact that we suffer an observation selection effect when regarding evolution as a process that eventually turned out to produce GI. In other words: evolution only seems likely to produce GI in worlds where evolution actually did produce GI.
- It might well be that the evolution of humans contained one or several "hard steps", which are very unlikely to occur within the time in which Earth is habitable. Maybe we just got really lucky and don't know about it. For a more elaborate summary of the key insights, check out Mark Xu's summary on Shulman and Bostrom's paper "How Hard is Artificial Intelligence?".
    - In order to quantify how this effects the compute necessary for a rerun of human evolution, Ege Erdil combined Hanson's grabby aliens model[15] and a consideration from John Schulman[16], concluding that an update accounting for anthropic considerations is likely to be within 6 OOM.
    - Erdil also makes the argument that all capabilities relevant for GI have developed recently (within the last 10 million years[17]). With the assumption from Hanson's model that hard steps are on average spaced equally apart throughout evolutionary history (meaning that the last one would have been around 300 million years ago), no anthropic update would be needed at all.
- Thoughts
    - The numbers given here highly depend on Hanson's model.

---

[14] From Lin's post: *"The space of architectures searched over by evolution seems much larger to me than the space of 2020 algorithms, in a way that we don't even know how to specify. For example, the brain seems to have a much more complicated network topology than 2020 algorithms -- with an interplay of feedforward and feedback processes playing a key role in some theories of how intelligence might work! Other differences include that the dynamical update rule for how the state of a system changes from one moment to the next seems more complicated in brains than in neural networks, the architecture of the brain can change over the lifetime of an organism instead of being fixed at initialization, and so on. (See Marblestone, Wayne, Kording (2016) for a review.)"*

[15] Hanson's median scenario proposes a grabby civilization origin every 1E20 planets. Hanson estimates about 6 "hard steps", with a plausible range of 3-12.

[16] "[...] doing N independent parallel computation[s] and selecting one of them is way less useful than doing an N times longer serial computation."

[17] This is about the time when the clade represented by humans, chimpanzees and bonobos split from the ancestors of gorillas.

    - If one buys Hanson's model and believes that relevant GI capacities developed within the last 10 million years, anthropic considerations can be discarded altogether.
- Estimated effect size
    - Depending on when capabilities relevant for GI have developed: no adjustment, or an increase of up to 6 OOM to the evolution anchor.

2.5 Missing paradigm shifts

- Cotra's report relies heavily on the assumptions that 2020 algorithms are capable of scaling to TAI given more compute, and that algorithmic progress is gradual.[18]
- In Biology-Inspired AGI Timelines: The Trick That Never Works, Eliezer Yudkowsky argues against this approach and declares that progress in algorithms is not gradual, but highly depends on paradigm shifts such as the invention of transformers or the development of Long Short-Term Memory (LSTM).
    - On this view, anchoring on SOTA algorithms doesn't make much sense, as TAI algorithms will look nothing like anything we are able to imagine at this point. Taking current SOTA algorithms and extrapolating from there is equivalent to stating that algorithms such as alpha-beta and Eliza are capable of achieving similar results as AlphaGo and GPT-3, if you throw vast amounts of compute at them, which is obviously not true.
    - Yudkowsky claims that paradigm shifts are expected to drastically shorten TAI timelines, and suggests discarding biological anchors completely.
    - In Biological Anchors: A Trick That Might Or Might Not Work, a review of Yudkowsky's AI safety dialogues, Scott Alexander compares the scaling of current algorithms with compute power to assess the development of TAI to the attempt of a Victorian scientist to predict the invention of spaceships by extrapolating trends in ship size, anchoring on the size of the moon. Focusing on some measurable, but arbitrary property (compute power/ship size) that seems related at first glance can lead to the neglect of some fundamental step (paradigm shift/invention of rockets).
- Counters to the argument made by Yudkowsky include comments made by Carl Shulman and Vanessa, describing how software progress in AI has been mostly dependent on available compute:
    - Hardware drives more improvement than software. Hardware improvements allow for the increase in available compute that drives software improvements. In other words: algorithmic innovation is not a bottleneck, as the best algorithms for a given level of compute are found relatively quickly.
    - This implies that while paradigm shifts occur, they dot not bottleneck AI progress, as they are induced by and happen shortly after (relevant) increases in available compute.

---

[18] "*I am committed to a weaker claim, which is that it is likely that researchers could figure out how to combine 2020 architectures and algorithms with an amount of computation between human lifetime computation and evolution computation (as well as arbitrary amounts of training data) to train a transformative model within a few years of trying. My model wouldn't be a useful tool for thinking about TAI timelines if you assigned a small probability to this [...].*"See pages 5-7 and 18 of part 4 of Cotra's report, respectively, and the chapter "Seeing continued progress and no major counterexamples to DL scaling well" in her two-year update.

- Jennifer Lin considers TAI to be scaled from a “paradigm that’s more efficient than deep learning” and can imagine that relevant breakthroughs might take >>30 years to occur.
- Thoughts
    - Intuitively, it makes a lot of sense to assume that paradigm shifts should influence TAI timelines and they are not accounted for in Cotra’s estimates.
    - It seems essential exploring wether paradigm shifts have been transformative for AI progress independently or only so in combination with significant hardware progress. If the latter is true, and if paradigm shifts occur quickly after increases in available compute power, their influence on AI timelines should be neglectible compared to hardware progress.
    - What this argument means for someone’s timelines depends on wether they buy the initial assumption that 2020 algorithms are capable of scaling to TAI with sufficient compute.
        - If yes: paradigm shifts are expected to shorten TAI timelines significantly.
        - If no: in this case, it might not make much sense to rely on Cotra’s timeline estimates in the first place. If paradigm shifts are prerequisite for TAI development, they might constitute an additional bottleneck and lengthen timelines, or accelerate TAI development by introducing unprecedented, drastic changes.
- Estimated effect size
    - *Intuitions differ strongly; from possibly shifting timelines up to >>30 years ito the future (Lin) to discarding bioanchors completely (Yudkowsky)*

3. Proposing a best-guess for an upper bound to the thermodynamic approach

- We decided to explore the thermodynamic approach further as we found it more convincing than the original evolutionary anchor estimate as an upper bound, whilst still seeming more methodologically tractable within the scope of our project timelines, given our technical backgrounds (unlike, for example, the architecture search space approach).
- In this section we aim to explain the thermodynamic approach in more detail, and explain in a google sheets model.
- In addition, we were able to expand on the approach and lower the estimated upper bound, thus narrowing the range.

The thermodynamic approach in more detail

- The thermodynamic approach, introduced by Ege Erdil, aims to estimate the total amount of energy received by the earth (the Earth’s “energy budget”) and convert this into FLOP. Erdil used the Landauer principle for this conversion. This principle is described below:
- *"The Landauer's principle is a physical principle pertaining to the lower theoretical limit of energy consumption of computation. It holds that an irreversible change in*

information stored in a computer, such as merging two computational paths, dissipates a minimum amount of heat to its surroundings. In intuitive terms, this is the energy cost of information." (source)

- The Landauer principle outlines the minimum amount of energy needed to run an operation without decreasing the overall entropy of a system (ie. without breaking the second law of thermodynamics).
- Thus, given some set of energy, if you assume that all of that energy was put into running operations or information processing, you can come up with a number of operations that could have been possibly run.
- Operations can also be seen as FLOP (the operations a computer undertakes) to calculate the maximum amount of FLOP that could have theoretically happened.
- But, using all of the energy budget of earth here is a huge overestimate, as not all of energy was put into running operations/information processing (much was wasted, much was converted into things like light, thermal, or kinetic energy and not into information).
- In addition, even energy that was put into running operations was not likely to be converted at the lower theoretical limit the Landauer limit suggests, and was likely much less efficient.

Determining a more informative range

- One plausible way to find a more informative range was using human energy to FLOP conversion instead of the Landauer approach
- This is based on the assumption that humans are at least more efficient at converting energy into FLOP than the average joule of energy deposited on Earth over the last billion years. This is likely as most energy is not used to process information at all (around 23% of solar energy is reflected back into space, and 29% is absorbed in the atmosphere by water vapour, dust, and ozone).
    - The two approaches we consider are:
        - The caloric approach: determines the energy required to produce one FLOP by taking into account two factors: the average daily caloric intake of a human and the average FLOP/s performed by the brain. It divides the total daily calories by the number of seconds in a day to find out how much energy the human body uses every second. Using this information, it then calculates the energy cost of a single FLOP.
        - The brain energy approach determines the energy required to produce one FLOP by taking into account two factors: the average amount of energy used by a human brain and the average FLOP/s performed by the brain. Using this information, it then calculates the energy cost of a single FLOP.
- In order to compare real-world processes with what human-build computers can do, we convert energy into FLOP. So we treat the energy absorbed by the Earth in the last billion years as if it was converted into FLOP at the level of efficiency of the human body or a human brain, to arrive at a measure we can express in terms of computational capacity. This approach is not intuitive and might be fundamentally flawed. See more under “FLOP conversion in weird” in part 1 of the discussion section.

- If we assume the average joule used by humans yields more FLOP than the average joule used by Earth overall, it follows that using human energy to FLOP conversion to calculate the number of FLOP used by evolution serves as an upper bound. This means this conversion can be used as an upper bound instead of the Landauer conversion, providing a smaller, more informative range.

Results

The model, along with the calculations is available here. The table below shows our key results.

Approach	Upper bound estimate	Lower bound estimate	Janvi and Victoria's best guess estimate
Landauer's principle	6.19E46	6.19E35	6.19E41
Caloric approach	1.67E40	1.79E35	1.79E37
Brain Energy approach	6.07E40	1.73E36	8.67E37

These approaches result in upper bound estimates which are 4-6 OOMs smaller than the Landauer's principle approach. In addition, it's worth noting that the upper bounds for both of these estimates are lower than (though not far from) Cotra's estimate (1E41), at 2E40 for the caloric approach and 6E40 for the brain energy consumption.

4. Discussion

Part 1: Major reasons to be sceptical of the evolution anchor framework and our results

There are some considerations which make us suspicious about the evolution anchor and its ability to act as an upper bound or input into the calculation of TAI timelines.

Weighing FLOP estimates against developments in compute capacity

It is difficult to intuitively buy some of the estimates that Cotra's approach and our own best-guesses come up with. In particular, it's hard to take ridiculously large numbers seriously when we've seen so much development with relatively fewer FLOP.

Epoch estimates it took 2.1e+25 FLOP (in total) to train GPT-4 (though this hasn't been reported by OpenAI. Another example with reported numbers is Megatron LM which took 3.9e+24). The evolution anchor assumes that this level of computation is carried out in one second by our average ancestral population. If I were to blindly use this model, I would expect that our current day systems should only be able to do about as much as 1 second of the whole nematode population right now, or through the output framing, what the frontier nematode can do. Daniel Kokotajlo has written some more on how powerful this much compute might be.

Bounding parameter estimates is difficult and sometimes arbitrary

One of the reasons some parameters vary significantly is because bounding estimates is difficult, and somewhat arbitrary.

For a minimum viable environment, we considered different approaches to determine which environmental properties have been most definitely significant for the development of human-level intelligence, while ensuring that our estimate remained conservative and pertained to the “upper-bound quality” that is the main function of the evolution anchor.

Difficult-to-answer questions that came up while considering reasonable cut-offs included:

- Granularity cut-off
    - Rerunning human evolution probably does not require the simulation of an environment on a molecular quantum level.
    - What granularity of detail is necessary?
        - Reducing granularity decreases complexity.
        - Which details are redundant (regarding the environment and other agents/organisms)?
        - How does reducing the granularity affect dynamics such as weather, climate, atmospheric composition and others? How to compensate for that?
    - We can anchor on what human senses can perceive, but that neglects the development of other lifeforms that have been highly influential for human evolution. How to draw a line between relevant and non-relevant organisms?
- Geographic cut-off
    - Rerunning human evolution probably did not take all the space Earth provides.
    - We can look at an isolated dome of a defined space, but conditions in that dome would still depend on the environment outside that dome. Alternatively, we could downsize the radius of our simulated Earth.
    - Which size is appropriately conservative?
        - Reduced space reduces sizes and numbers of biomes.
        - Reducing space reduces total number of organisms and thereby the frequency of mutations.
    - We could compensate for lesser biomes by changing them frequently. How often should these changes occur and how stark should they be?
- Temporal cut-off
    - Not all of evolutionary history was relevant for the development of human-level intelligence.
    - Throughout Earth’s evolutionary history, there have been phases of relative stagnation where nothing changed much. In a rerun of human evolution, these periods could be cut out.
        - When exclusively looking at the evolution of intelligence, there should be more stagnant phases where much changed, but nothing relevant for the development of human-level intelligence happened.
    - When did the most significant steps for GI development occur? How could we avoid cutting those off?

These considerations are not centred around the actual evolutionary history of humans, but rather around what we guess could have been the most significant parts of it - with low confidence on all parameters.

Some of the parameters can vary by many orders of magnitude

As seen in the “parameter variability score” in the model, many of the estimates can vary by orders of magnitude.

This is just in the calculation of the evolution anchor alone. The bio anchors report and model use many parameters which vary by orders of magnitude. I would be interested in seeing a report which considers the effect of changing these parameter guesses on timelines, because I expect this might lead to some really weird/non-intuitive results.

One of the most striking examples is the calculation of how many FLOP a brain processes, which is taken from Carlsmith’s report. This value affects all of the anchors. Carlsmith notes in his report that estimates for this value vary from 1e12 to 7e21, more than 9 OOMs.

FLOP conversion is weird

The conversion we’re doing is very weird. FLOP is not a default unit of measurement with which to calculate the development of intelligence as it happens within evolutionary processes. It is an open question whether it is suitable to summarise evolutionary processes or brain functions into operations and FLOP.

Even if we knew the true FLOP it took to develop intelligence, it is possible that this number does not translate well to how much FLOP is needed by a transformative model in the future. The number might be an underestimate because other factors which did not have to do with information processing/computation contributed significantly to evolution of intelligence, and might have thus been completely missed by the estimate. Equally, it might be an overestimate because of the costs we’ve already incurred.

The analogy is made by AI Impacts on discontinuous progress, and used by SSC here displays this well. It is summarised as “The [mistake] is thinking that spaceship progress depends on some easily-measured quantity (size) instead of on fundamental advances (eg. figuring out how rockets work).” I am not even trying to make the full version of this claim. TAI progress may depend on some easily-measured quantity, and might not depend on fundamental advances. However, FLOP does not seem like it be a suitable “easily-measured quantity” to describe what it took evolution to develop intelligence.

In practice, it is very likely that we will derive intelligence/make progress in a way that is completely in a way that is incomparable to how evolution first derived human-level intelligence. One reason to think this is that, unlike evolution, we have the specific goal to create intelligent systems and are optimising to do so. In addition to this, our optimisation algorithms are different, and seem to perform better than evolutionary algorithms (see Marius Hobbhahn’s critique).

Part 2: What your take-aways might look like

- If you:
    - Believe that Cotra's model and framework for calculating TAI is reasonable
    - Believe that the thermodynamic estimate for the evolution anchor is better than the brain computation method Cotra uses, and
    - Believe that our best guess proposal for an improvement upon the thermodynamic estimate is better than the original Landauer approach
        - You might change your best guess for a FLOP estimate for an upper bound to be 1.79E40 FLOP instead of 1e41 FLOP.
- If you:
    - Believe one of the other criticisms/adjustments noted in section two is legitimate, you might
        - Use the naive estimated effect sizes to update your estimate of FLOP needed to develop TAI
        - Choose investigate it further and update accordingly
        - Consider Nuño Sempere's propagation of beliefs given additional uncertainty surrounding the evolution anchor
- If you:
    - Believe any of the major reasons to be sceptical listed above you might
        - Downweight on the evolution anchor and the entire biological anchors report
        - Consider other ways to assess AI timelines
        - Consider Nuño Sempere's propagation of beliefs given additional uncertainty surrounding the evolution anchor