Evolution provides no evidence for the sharp left turn

149

Ω 44

by **Quintin Pope** 11th Apr 2023

Sharp Left Turn | AI Takeoff | Object-Level AI Risk Skepticism | Evolution | AI | Frontpage

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Does human evolution imply a sharp left turn from AIs?

Arguments for the sharp left turn° in AI capabilities often appeal to an "*evolution -> human capabilities*" analogy and say that evolution's outer optimization process built a much faster human inner optimization process whose capability gains vastly outstripped those which evolution built into humans. Such arguments claim we will see a similar transition while training AIs, with SGD creating some 'inner thing' which is not SGD and which gains capabilities much faster than SGD can insert them into the AI. Then, just like human civilization exploded in capabilities over a tiny evolutionary time frame, so too will AIs explode in capabilities over a tiny "SGD time frame".

Evolution's sharp left turn happened for evolution-specific reasons

I think that "*evolution -> human capabilities*" is a bad analogy for "*AI training -> AI capabilities*". Let's compare evolution to within lifetime learning for a single generation of an animal species:

- A generation is born.
- The animals of the generation learn throughout their lifetimes, collectively performing many billions of steps of learning.
- The generation dies, and all of the accumulated products of within lifetime learning are lost.
- Differential reproductive success slightly changes the balance of traits across the

species.

The only way to transmit information from one generation to the next is through evolution changing genomic traits, because death wipes out the within lifetime learning of each generation.

Now let's look at the same comparison for humans:

- A generation is born.
- The humans of the generation learn throughout their lifetimes, collectively performing many billions of steps of learning.
- **The current generation transmits some fraction of their learned knowledge to the next generation through culture.**
- The generation dies, **but only some** of the accumulated products of within lifetime learning are lost.
- Differential reproductive success slightly changes the balance of genomic traits across humanity.

Human culture allows some fraction of the current generation's within lifetime learning to transmit directly to the next generation. In the language of machine learning, the next generation benefits from a kind of knowledge distillation, thanks to the prior generation providing higher quality 'training data' for the next generation's within-lifetime learning.

This is extremely important because within-lifetime learning happens much, *much* faster than evolution. Even if we conservatively say that brains do two updates per second, and that a generation is just 20 years long, that means a single person's brain will perform ~1.2 billion updates per generation. Additionally, the human brain probably uses a stronger base optimizer than evolution, so each within-lifetime brain update is also probably better at accumulating information than a single cross-generational evolutionary update. Even if we assume that only 1 / 10,000th of the information learned by each generation makes its way into humanity's cross-generational, persistent endowment of cultural information, that still means culture advances ~100,000 times faster than biological evolution.

I think that "*evolution -> human capabilities*" is a very bad reference class to make predictions about "*AI training -> AI capabilities*". We don't train AIs via an outer optimizer over possible inner learning processes, where each inner learning process is initialized from scratch, then takes billions of inner learning steps before the outer

optimization process takes one step, *and then is deleted after the outer optimizer's single step*. Such a bi-level training process would **necessarily** experience a sharp left turn once each inner learner became capable of building off the progress made by the previous inner learner (which happened in humans via culture / technological progress from one generation to another).

However, this sharp left turn does *not* occur because the inner learning processes suddenly become much better / more foomy / more general in a handful of outer optimization steps. It happens because you devoted billions of times more optimization power to the inner learning processes, *but then deleted each inner learner shortly thereafter*. Once the inner learning processes become able to pass non-trivial amounts of knowledge along to their successors, you get what looks like a sharp left turn. But that sharp left turn only happens because the inner learners have found a kludgy workaround past the crippling flaw where they all get deleted shortly after initialization.

In my frame, we've already figured out and applied the sharp left turn to our AI systems, in that we don't waste our compute on massive amounts of incredibly inefficient neural architecture search, hyperparameter tuning, or meta optimization. For a given compute budget, the best (known) way to buy capabilities is to train a single big model in accordance with empirical scaling laws such as those discovered in the Chinchilla paper, not to split the compute budget across millions of different training runs for vastly tinier models with slightly different architectures and training processes. In fact, we can be even more clever and use small models to tune the training process, before scaling up to a single large run, as OpenAI did with GPT-4.

(See also: Gwern on the blessings of scale.)

It's true that we train each new AI from scratch, rather than reusing any of the compute that went into previous models. However, the situation is very different from human evolution because each new state of the art model uses geometrically more compute than the prior state of the art model. Even if we could perfectly reuse the compute from previous models, it wouldn't be nearly so sharp an improvement to the rate of progress as occurred in the transition from biological evolution to human cultural accumulation. I don't think it's plausible for AI capabilities research to have the same sort of hidden factor of ~billion resource overhang that can be suddenly unleashed in a short-to-humans timescale.

The capabilities of ancestral humans increased smoothly as their brains increased in scale

and/or algorithmic efficiency. Until culture allowed for the brain's within-lifetime learning to accumulate information across generations, this steady improvement in brain capabilities didn't matter much. Once culture allowed such accumulation, the brain's vastly superior within-lifetime learning capacity allowed cultural accumulation of information to vastly exceed the rate at which evolution had been accumulating information. This caused the human sharp left turn.

However, the impact of scaling or algorithmic improvements on the capabilities of individual brains is still continuous. which is what matters for predicting how suddenly AI capabilities will increase as a result of scaling or algorithmic improvements. Humans just had this one particular bottleneck in cross-generational accumulation of capabilities-related information over time, leading to vastly faster progress once culture bypassed this bottleneck.

Don't misgeneralize from evolution to AI

Evolution's sharp left turn happened because evolution spent compute in a shockingly inefficient manner for increasing capabilities, leaving vast amounts of free energy on the table for any self-improving process that could work around the evolutionary bottleneck. Once you condition on this specific failure mode of evolution, you can easily predict that humans would undergo a sharp left turn at the point where we could pass significant knowledge across generations. I don't think there's anything else to explain here, and no reason to suppose some general tendency towards extreme sharpness in inner capability gains.

History need not repeat itself. Human evolution is not an allegory or a warning. It was a series of events that happened for specific, mechanistic reasons. If those mechanistic reasons do not extend to AI research, then we ought not (mis)apply the lessons from evolution to our predictions for AI.

This last paragraph makes an extremely important claim that I want to ensure I convey fully:

- IF we understand the mechanism behind humanity's sharp left turn with respect to evolution

- AND that mechanism is inapplicable to AI development

- THEN, there's no reason to reference evolution *at all* when forecasting AI development rates, not as evidence for a sharp left turn, not as an "illustrative example" of some mechanism / intuition which might supposedly lead to a sharp left turn in AI development, not for *anything*.

Here's an analogy to further illustrate the point:

> Imagine that we were trying to figure out how to build very reliable cars. We've so far built a number of car prototypes, but none have reached the full load-bearing capacity of even a single strong human, never mind the vastly superhuman transport capacity that the laws of physics seem to permit.
>
> Someone raises the concern that, once we try to scale current prototypes to the superhuman limit, they'll tend to spontaneously combust, despite the fact that none of the prototypes have ever done so. As evidence for such an event, the person points to the fact that a previous car building effort, led by EVO-Inc., actually had built cars that did sometimes explode randomly.
>
> Concerned, we investigate EVO-Inc.'s car building effort, hoping to avoid whatever failure plagues their cars. Only, upon investigating EVO-Inc., it turns out that they're actually run by insane space clowns, and the reason their cars occasionally explode is because they used armed landmines in place of hubcaps.

My point is that other car builders can learn ~zero lessons from EVO-Inc.[1] The mechanism behind their cars' spontaneous detonation is easily avoided by not using landmines as hubcaps. The organizational-level failures that led to this design choice on EVO-Inc.'s part are also easily avoided by not being insane space clowns. We should not act like there might be some general factor of "explodeyness" which will infect other car building efforts, simply by virtue of those efforts tackling a similar problem to the one EVO-Inc. failed at.

EVO-Inc's failures arose from mechanisms which do not apply to human organizations tackling similar problems. EVO-Inc. didn't use landmines as hubcaps because they were run by greedy, myopic executives who cut corners on safety to increase profits. They didn't do so because they were naive optimists who failed to understand why building non-exploding cars is hard like computer security or rocket science, and who failed to apply proper security mindset to their endeavors. EVO-Inc used landmines as hubcaps

because they were run by insane space clowns who did insane space clown things.

Human car builders may have to tackle problems superficially similar to the spontaneous combustion of the EVO-Inc. cars. E.g., they may have to design the fuel tanks of their cars to avoid combustion during a crash. However, those efforts *still* should not take lessons from EVO-Inc. E.g., if other car builders were to look at crash data from EVO-Inc.'s cars, and naively generalize from the surface-level outcomes of an EVO-Inc. car crash to their own mechanistically different circumstances, they might assume that supersonic fragments posed a significant risk during a crash, and then add ballistic armor between the driver and the wheels, despite this doing nothing to prevent a car's fuel tank from igniting during a crash.

I think our epistemic relationship with evolution's example should be about the same as the human car builders' epistemic relationship with EVO-Inc. Evolution's combined sharp left turn and alignment failures happened because evolution is a *very* different process compared to human-led AI development, leading to evolution-specific mechanisms, which no sane AI developer would replicate.

In order to experience a sharp left turn that arose due to the same mechanistic reasons as the sharp left turn of human evolution, an AI developer would have to:

1. Deliberately create a (very obvious[2]) inner optimizer, whose inner loss function includes no mention of human values / objectives.[3]
2. Grant that inner optimizer ~billions of times greater optimization power than the outer optimizer.[4]
3. Let the inner optimizer run freely without any supervision, limits or interventions from the outer optimizer.[5]

This is the AI development equivalent of using landmines as hubcaps. It's not *just* that this is an insane idea from an alignment perspective. It's also an insane idea from just about any other perspective. Even if you're only trying to maximize AI capabilities, it's a terrible idea to have such an extreme disparity in resources between the inner and outer loops.

AI researchers have actually experimented with bi-level optimization processes such as neural architecture search and second-order meta learning. Based on current results, I don't think anything approaching multiple orders of magnitude difference in resource use between the inner and outer optimizers is plausible. It's just not efficient, and we have better approaches. From the GPT-4 paper:

> A large focus of the GPT-4 project was building a deep learning stack that scales predictably. The primary reason is that for very large training runs like GPT-4, it is not feasible to do extensive model-specific tuning. To address this, we developed infrastructure and optimization methods that have very predictable behavior across multiple scales. These improvements allowed us to reliably predict some aspects of the performance of GPT-4 from smaller models trained using 1, 000× – 10, 000× less compute.

Even if we could magically repurpose all of the compute used throughout OpenAI's tuning of the GPT-4 architecture / training process, I doubt it would even amount to as much compute as they used in the final GPT-4 training run, much less exceed that quantity by orders of magnitude. Modern training practices simply lack that sort of free energy.

See also: Model Agnostic Meta Learning proposed a bi-level optimization process that used between 10 and 40 times more compute in the inner loop, only for Rapid Learning or Feature Reuse? to show they could get about the same performance while removing almost all the compute from the inner loop, or even by getting rid of the inner loop entirely.

Fast takeoff is still possible

The prior sections argue that we should not use an evolutionary analogy as evidence that an inner learner will sufficiently outperform the outer optimizer that constructed it so as to cause a massive spike in capabilities as a result of the same mechanisms that drove the sharp left turn in human evolution.

However, introducing new types of positive feedback loops across multiple training runs may lead to fast takeoff, but it would be a mechanistically different process than the evolutionary sharp left turn, meaning there's no reason to assume takeoff dynamics mirroring those of human evolution. There are two specific mechanisms that I think could produce a fast takeoff:

- AIs contributing to AI capabilities research, producing a positive feedback loop with a sharp upwards kink around the time that AI contributions exceed human contributions.

- AIs deliberately seeking out new training data that grant them useful capabilities. E.g., an AI trying to improve its bioengineering capabilities may set up a very fast cycle of gathering and analyzing new biological data, which significantly outpaces the rate of human scientific innovation.

If fast takeoff is still plausible, why does the specific type of positive feedback loop matter? What changes, as a result of considering various AI-specific fast takeoff mechanisms, as opposed to the general expectation of sudden transitions, as implied by the evolution analogy? Here are four alignment-relevant implications:

1. **Takeoff is less abrupt.** Both of the above mechanisms are vaguely similar to how human cultural development allowed us to jump forwards in capabilities by feeding the outputs of one generation into the "training data" of the next generation. However, I expect that neither mechanism will produce as much of a relative jump in AI capabilities, as cultural development produced in humans. Neither mechanism would suddenly unleash an optimizer *multiple* orders of magnitude faster than anything that came before, as was the case when humans transitioned from biological evolution to cultural development.
2. **Takeoff becomes easier to navigate.** These specific mechanisms of capabilities advance probably both allow for iteration and experimentation. We currently have examples of both AI capabilities advances and of online learning / exploration processes. We can run experiments on current systems to assess the alignment risks posed by both these sources of capabilities improvement.
3. **Capabilities gains are less general.** "capabilities generalize further than alignment" is a common refrain in discussions about the sharp left turn. Usually, this claim is justified by making an analogy to how human capabilities started to quickly generalize across many domains simultaneously.

   However, the process responsible for human breadth of generality was not some small architectural modification evolution made to the human brain. It was humanity's cross-generational process of expanding and improving our available "training data" to cover a broader and broader range of capabilities across many domains (a process we sometimes call "science"). The evolutionary analogy thus offers no reason to expect sudden jumps in generality without corresponding extensions of the training data.

Without this evolutionary analogy, why should we even elevate the very specific claim that '*AIs will experience a sudden burst of generality* ***at the same time*** *as all our alignment techniques fail.'* to consideration at all, much less put significant weight on it?

4. **Alignment probably generalizes pretty well.** Speaking of alignment techniques failing, I expect alignment techniques to mostly generalize across capabilities jumps caused by either of the above mechanisms for sudden capabilities gain.

Will alignment generalize across sudden capabilities jumps?

The previous section argued that the mechanisms driving the sharp left turn in human evolution are not present in AI development, and so we shouldn't generalize from the results of human evolution to those of AI development, even when considering positive feedback loops whose surface-level features are reminiscent of the sharp left turn in human evolution.

This section will first reference and briefly summarize some past writing of mine arguing that our "misalignment" with inclusive genetic fitness isn't evidence for AI misalignment with our values. Then, I'll examine both mechanisms for a possible fast takeoff that I described above from an "inside view" machine learning perspective, rather than assuming outcomes mirroring those of human evolutionary history.

Human "misalignment" with inclusive genetic fitness provides no evidence for AI misalignment

I previously wrote a post, Evolution is a bad analogy for AGI: inner alignment°, arguing that evolutionary analogies between human values and inclusive genetic fitness have little to tell us about the degree of values misgeneralization we should expect from AI training runs, and that analogies to human within-lifetime learning are actually much more informative[6].

I also wrote this subsection° in a much longer post°, which explains why I think evolution is mechanistically very different from AI training, such that we cannot easily infer lessons about AI misgeneralization by looking at how human behaviors differ between the modern and ancestral environments.

Very briefly: "human behavior in the ancestral environment" versus "human behavior in

the modern environment" isn't a valid example of behavioral differences between training and deployment environments. Humans weren't "trained" in the ancestral environment, then "deployed" in the modern environment. Instead, humans are continuously "trained" throughout our lifetimes (via reward signals and sensory predictive error signals). Humans in the ancestral and modern environments are different "training runs".

As a result, human evolution is not an example of:

> We trained the system in environment A. Then, the trained system processed a different distribution of inputs from environment B, and now the system behaves differently.

It's an example of:

> We trained a system in environment A. Then, we trained a *fresh version* of the same system on a different distribution of inputs from environment B, and now the *two different systems* behave differently.

The near-total misalignment between inclusive genetic fitness and human values is an easily predicted consequence of this (evolution-specific) bi-level optimization paradigm, just like the human sharp left turn is an easily predicted consequence of the (evolution-specific) extreme resource disparity between the two optimization levels. And just like evolution provides no reason to assume our own AI development efforts will experience a sharp left turn, so to does evolution not provide any reason to assume our AI development efforts will show extreme misgeneralization between training and deployment.

Capabilities jumps due to AI driving AI capabilities research

For the first mechanism of AIs contributing to AI capability research, I first note that this is an entirely different sort of process than the one responsible for the human sharp left turn. Evolution made very few modifications to the human brain's architecture during the timeframe in which our cultural advancement catapulted us far beyond the limits of our ancestral capabilities. Additionally, humans have so far been completely incapable of changing our own architectures, so there was never a positive feedback loop of the sort that we might see with AIs researching AI capabilities.

Because of this large difference in underlying process between this possible fast takeoff mechanism and the evolutionary sharp left turn, I think we should mostly rely on the current evidence available from AI development for our predictions of future AI development, rather than analogies to our evolutionary history. Additionally, I claim that alignment techniques already generalize across human contributions to AI capability research. Let's consider eight specific alignment techniques:

- Reinforcement learning from human feedback
- Constitutional AI
- Instruction prompt tuning
- Discovering Language Model Behaviors with Model-Written Evaluations
- Pretraining Language Models with Human Preferences
- Discovering Latent Knowledge in Language Models Without Supervision
- More scalable methods of process based supervision
- Using language models to write their own instruction finetuning data

and eleven recent capabilities advances:

- Optimally training language models using the Chinchilla scaling laws
- Transcending Scaling Laws with 0.1% Extra Compute
- Better tuning of training and architectural hyperparameters (example)
- Retrieval mechanisms for language models, such as RETRO
- 1 bit Adam for efficiently sharing gradient info across GPUs
- Doing more than one epoch on high quality text
- (Possibly) an improvement on the Adam optimizer
- Distributed training across many low-memory GPUs
- Stable, 8-bit transformer implementations
- Applying layer norms to query and key outputs of attention layers to stabilize training.
- The Hyena operator as a replacement for attention, to (maybe?) scalable sub-quadratic sequence processing architectures

I don't expect catastrophic interference between any pair of these alignment techniques and capabilities advances. E.g., if you first develop your RLHF techniques for models trained using the original OpenAI scaling laws, I expect those techniques to transfer pretty well to models trained with the Chinchilla scaling laws.

I expect there is *some* interference. I expect that switching your architecture from a vanilla transformer to a RETRO architecture will cause issues like throwing off whatever RLHF hyperparameters you'd found worked best for the vanilla architecture, or complicate analysis of the system because there's now an additional moving part (the retrieval mechanism), which you also need to track in your analysis.

However, I expect we can overcome such issues with "ordinary" engineering efforts, rather than, say, RLHF techniques as a whole becoming entirely useless for the new architecture. Similarly, whatever behavioral analysis pipeline you'd developed to track models based on the vanilla architecture can probably be retrofitted for models based on the RETRO architecture without having to start from scratch.

Importantly, the researchers behind the capabilities advances were *not* explicitly optimizing to maintain backward compatibility with prior alignment approaches. I expect that we can decrease interference further by just, like, *bothering to even try and avoid it.*

I'd like to note that, despite my optimistic predictions above, I do think we should carefully measure the degree of interference between capabilities and alignment techniques. In fact, doing so seems very *very* important. And we can even start right now! We have multiple techniques for both alignment and capabilities. You can just choose a random alignment technique from the alignment list, a random capabilities technique from the capabilities list, then see if applying the capabilities technique makes the alignment technique less effective.

The major exception to my non-interference claim is for alignment techniques that rely on details of trained models' internal structures, such as mechanistic interpretability. CNNs and transformers require different sorts of interpretability techniques, and likely have different flavors of internal circuitry. This is one reason why I'm more skeptical of mechanistic interpretability as an alignment approach[7].

Capabilities jumps due to AI iteratively refining its training data

I think the second potential fast takeoff mechanism, of AIs continuously refining their

training data, is riskier, since it allows strange feedback loops that could take an AI away from human-compatible values. Additionally, most current models derive values and goal-orientated behaviors much more from their training data, as opposed to their architecture, hyperparameters, and the like.

E.g., I expect that choosing to use the LION optimizer in place of the Adam optimizer would have very little impact on, say, the niceness of a language model you were training, except insofar as your choice of optimizer influences the convergence of the training process. Architecture choices seem 'values neutral' in a way that data choices are not.

I still think the risks are manageable, since the first-order effect of training a model to perform an action X in circumstance Y is to make the model more likely to perform actions similar to X in circumstances similar to Y. Additionally, current practice is to train language models on an enormous variety of content from the internet. The odds of any given subset of model data catastrophically interfering with our current alignment techniques cannot be that high, otherwise our current alignment techniques wouldn't work on our current models.

However, second order effects may be less predictable, especially longer term second-order effects of, e.g., training future models on the outputs of current models. Such iterative approaches appear to be gaining popularity, now that current LMs are good enough to do basic data curation tasks. In fact, one of the linked alignment approaches, ConstitutionalAI, is based on using LMs to rewrite texts that they themselves will then train on. Similar recent approaches include:

- Large Language Models Can Self-Improve
- Language Models Can Teach Themselves to Program Better
- The Wisdom of Hindsight Makes Language Models Better Instruction Followers

Although this potential fast takeoff mechanism more closely resembles the mechanisms of cultural development responsible for the human sharp left turn, I think there are still important differences that make a direct extrapolation form human evolutionary history inappropriate. Most prominently, a data refinement fast takeoff wouldn't coincide with exploiting the same sort of massive resource overhang that came into play during the human sharp left turn.

Additionally, I expect there are limits to how far AIs can improve their training data

without having to run novel experiments and gather data different from their initial training data. I expect it will be difficult to extend their competency to a new domain without actually gathering new data from that domain, similar to how human scientific theory only progresses so far in the absence of experimental data from a new domain.

Conclusion

I think that evolution is a bad analogy for AI development. I previously argued° as much in the context of inner alignment concerns, and I've also argued° that evolution is actually very mechanistically different from the process of training an AI.

Our evolutionary history has all sorts of difficult-to-track details that *really* change how we should derive lessons from that history. In this post, the detail in question was the enormous disparity between the optimization strength of biological evolution versus brain-based within lifetime learning, leading to a giant leap in humanity's rate of progress, once within lifetime learning could compound over time via cultural transmission.

I've started to notice a common pattern in evolutionary analogies, where they initially suggest concerning alignment implications, which then seem to dissolve once I track the mechanistic details of what actually happened in the evolutionary context, and how that would apply to AI development. At this point, my default reaction to any evolutionary analogy about AI alignment is skepticism.

---

1. ^ Other than "don't take automotive advice from insane space clowns", of course.
2. ^ If you suspect that you've maybe *accidentally* developed an evolution-style inner optimizer, look for a part of your system that's updating its parameters ~a billion times more frequently than your explicit outer optimizer.
3. ^ - "inner optimizer" = the brain.
   - "inner loss function" = the combination of predictive processing and reward circuitry that collectively make up the brain's actual training objective.
   - "inner loss function includes no mention human values / objectives" because the brain's training objective includes no mention of inclusive genetic fitness.

4. ^ Reflects the enormous disparity in optimization strength between biological evolution and human within-lifetime learning, which I've been harping on about this whole post.
5. ^ Evolution doesn't intervene in our within-lifetime learning processes if it looks like we're not learning the appropriate fitness-promoting behavior.
6. ^ It's not even that I think human within-lifetime learning is *that* informative. It's just that I think "being more informative than evolution" is such a stupidly low bar that human within-lifetime learning clears it by a mile.
7. ^ I do think there's a lot of value in mechanistic interpretability as a source of evidence about the mechanics and inductive biases of SGD. For example, this paper discovered "name mover heads", attention heads that copy a speaker's name to the current token in specific contexts, and also discovered "backup name mover heads", which are attention heads that don't normally appear to act as name mover heads, but when researchers ablated the primary name mover heads, the backup name mover heads changed their behavior to act as name mover heads.