There is a debate within the field of AI regarding whether recursive self-improvement by an AI is possible, and, if so, whether it is likely to happen.

Prominent skeptics of recursive self-improvement include:

- François Chollet, who argued in 2017 that recursively self-improving systems cannot achieve exponential progress in practice.[1]
- Ted Chiang, who argues that AIs are like compilers, where bootstrapping allows self-improvement, but only up to a point.

Points of disagreement include:

- Whether an AI would need full access to its entire codebase for recursive self-improvement. Stuart Armstrong posits that it would not, because many pathways to self-improvement do not rely on this ability, such as the creation of sub-agents or partial self-modification.
- Whether a self-improving AI's capabilities will eventually reach the point of diminishing returns, and where this point may lie. For example, when it comes to AIs built using modern deep learning paradigms, some have argued that self-improvement may be infeasible and not cost-effective, due to the complexity and computational demands of these types of AIs. If this is true, then such AIs wouldn't satisfy the criteria for being seed AI, which were imagined to be designed and not selected for by search-like processes, requiring understanding of their own source code and the ability to make goal-preserving modifications to themselves.
- Whether AGI will have incentives to self-improve in the first place.

A common response to this objection is that self-improvement may be easier outside the deep learning paradigm, as the field of AI is replete with examples of a radical change in approaches allowing previously unfeasible problems to be solved extremely easily. Moreover, improving source code in the context of deep learning does not just refer to, say, a neural network changing each of its weights directly (although steering it as such may well be possible), but also things like code defining the architecture or the code for collecting training data, etc.

Ultimately, the danger of self-improvement does not lie in hypothetical infinite recursive improvements, but in pushing the AGI further down a path to AI takeover — even fairly modest, concrete and reasonable improvements may push an AGI beyond the point of controllability. Given that some trends in modern AI provide for steps on the path to self-improvement, the question of its long-term feasibility remains a crucial one. Furthermore, self-improvement is one way to get fast takeoff, but it is not the only one.

Alternative phrasings

[1] See also Eliezer Yudkowsky's rebuttal.

Related

- Might an "intelligence explosion" never occur?
- Why would the first Artificial General Intelligence ever create or agree to the creati…
- Does the idea of a recursively self improving AI assume that the AI is able to solve …

Scratchpad

Drafts at unknown dates

Alternatively, a terminology for a scenario where pivotal AI that don't engage in extensive self-modification are constructed instead, is KANSI.

https://www.lesswrong.com/posts/8NKu9WES7KeKRWEKK/why-all-the-fuss-about-recursive-self-improvement

~~it's entirely possible that the cognitions it implements are able to (e.g.) crack nanotech well enough to kill all humans, before they're able to crack themselves. The big update over the last decade has been that humans might be able to fumble their way to AGI that can do crazy stuff *before* it does much self-improvement.~~

Though, to be clear, from my perspective it's still entirely plausible that you will be able to turn the first general reasoners to their own architecture and get a big boost, and so there's still a decent chance that self-improvement plays an important early role. (Probably destroying the world in the process, of course. Doubly so given that I expect it's even harder to understand and align a system if it's self-improving.)

~~In other words, it doesn't seem to me like developments like deep learning have undermined the recursive self-improvement argument in any real way. The argument seems solid to me, and reality seems quite consistent with it.~~

comments:

- ~~Early AI systems can improve some but not all parts of their own design. This leads to rapid initial progress, but diminishing returns (basically they are free-riding on parts of the design already done by humans).~~
- ~~Eventually AI is able to improve enough stuff that there are increasing rather than diminishing returns to scale *even within the subset of improvements that the AI is able to make.*~~

~~acquire more resources, stack more layers count?~~

~~Most of my argument for FOOM in the Yudkowsky-Hanson debate was about self-improvement and what happens when an optimization loop is folded in on itself. Though it wasn't necessary to my argument, the fact that Go play went from "nobody has come close to winning against a professional" to "so strongly superhuman they're not really bothering any more" over two years just because that's what happens when you improve and simplify the architecture, says *you don't even need self-improvement* to get things that look like FOOM.~~

A difference I think Paul has mentioned before is that Go was not a competitive industry and competitive industries will have smaller capability jumps.

The thing the industry is calling AGI and targeting may end up being a specific style of shallow deployable intelligence when "real" AGI is a different style of "deeper" intelligence (with, say, less economic value at partial stages and therefore relatively unpursued). This would allow a huge jump like AlphaGo in AGI even in a competitive industry targeting AGI. …

https://www.lesswrong.com/posts/wZAa9fHZfR6zxtdNx/agi-systems-and-humans-will-both-need-to-solve-the-alignment

~~At some point, I expect AGI systems to want/need to solve the alignment problem in order to preserve their goal structure while they greatly increase their cognitive abilities, a thing which seems potentially hard to do.~~

It's not clear to me when that will happen. Will this be as soon as AGI systems grasp some self / situational awareness? Or will it be after AGI systems have already blown past human cognitive abilities and find their values / goals drifting towards stability?

It seems possible that at the point an AGI system reaches the "has stable goals and wants to preserve them", it's already capable enough to solve the alignment problem for itself, and thus can safely self-improve to its limits. It also seems possible that it will reach this point significantly before it has solved the alignment problem for itself (and thus develops the ability to self-improve safely).

If it's the latter situation, where an AGI system has decided it needs to preserve its goals during self improvement, but doesn't yet know how to...

~~An AGI system of human-ish ability in many areas develops enough self/situational awareness to realize a few things:~~

- ~~The basics of instrumental convergence, thus wanting to seek power, protect itself, and preserve its goal representation~~
- ~~That goal-preservation might be (or would be) very difficult if it undergoes major self modification (perhaps it has already exhausted gains from simpler self-modifications)~~

~~For example, AGI-alignment-with-itself-under-self-improvement might be (probably would be) an easier problem than the getting-an-AGI-aligned-with-human-values problem. In that scenario, it seems possible / likely that the AGI system would get what it wants long before the humans got what they want. And if a main limiting factor on the AGI system's power was its unwillingness to self-modify in large ways, getting to its own AGI alignment solution before humans get to theirs might remove one of the main limitations keeping it from gaining the capabilities to seize power.~~

~~Assumption: An AGI system unwilling to self-modify because of fear of goal drift~~

comments:

it should not be underestimated how dangerous running multiple copies of yourself can be if you're not yet in a state where the conversations between them will converge usefully. Multiple copies are a lot more like separate beings than one might think a priori, because

copying simulator does not guarantee the simulacra will remain the same, even for a model trained to be coherent, even if that training is from scratch.

> Computer scientists, however, believe that self-improvement will be recursive. In effect, to improve, and AI has to rewrite its code to become a new AI. That AI retains its single-minded goal but it will also need, to work efficiently, sub-goals. If the sub-goal is finding better ways to make paperclips, that is one matter. If, on the other hand, the goal is to acquire power, that is another.
> The insight from economics is that while it may be hard, or even impossible, for a human to control a super-intelligent AI, *it is equally hard for a super-intelligent AI to control another AI*. Our modest super-intelligent paperclip maximiser, by switching on an AI devoted to obtaining power, unleashes a beast that will have power over it. Our control problem is the AI's control problem too. If the AI is seeking power to protect itself from humans, doing this by creating a super-intelligent AI with more power than its parent would surely seem too risky.

Note difference: ai's know their goals, humans collective don't

I expect the alignment problem for future AGIs to be substantially easier, because the inductive biases that they want should be much easier to achieve than the inductive biases that we want. That is, in general, I expect the distance between the distribution of human minds and the distribution of minds for any given ML training process to be much greater than the distance between the distributions for any two ML training processes. Of course, we don't necessarily have to get (or want) a human-like mind, but I think the equivalent statement should also be true if you look at distributions over goals as well.

~~Your post talks about a single AI, but I think it's also worth considering a multipolar scenario in which there are multiple competing AIs which face a tradeoff between more rapid self-improvement/propagation and value stability. One relevant factor might be whether value stability is achieved before (AI) space colonization become feasible -- if not, it may be difficult to prevent the spread of AIs~~ ~~valuing maximally rapid expansion.~~

~~Working to preserve values doesn't require them being stable, or in accord with how your mind is structured to learn and change, because working to preserve values can be *current behavior*. The work towards preservation of value is itself pivotal in establishing what the equilibrium of values is, there doesn't need to be any other element of the process that would point to particular values *in the absence of* work towards preservation of values. This source of values shouldn't be dismissed when considering the question of nature of a mind, or what its idealized values are, because the work towards preserving values, towards solving alignment, isn't automatically absent in actuality.~~

~~Thus LLM human imitations may well robustly retain human values, or close enough for mutual moral patienthood, even if their nature tends to change them into something else. Alien natural inclinations would only so change them in actuality in the absence of resolute opposition to such change by their own current behavior. And their current behavior may well preserve the values endorsed/implied by their current behavior, by shaping how their models get trained.~~

The solution for AI is that AI will choose slower methods of self-improvement, like learning or getting more hardware vs. creating new versions of itself. This means that AI will experience slower takeoff and thus will less interested in early treacherous turn – and more interested in late treacherous turn when it will become essential part of society.

~~If there is no solution to the alignment problem within reach of human level intelligence, then the AGI can't foom into an ASI without risking value drift…~~

~~A human augmented by a strong narrow AIs could in theory detect deception by an AGI. Stronger interpretability tools…~~

What we want is a controlled intelligence explosion, where an increase in strength of the AGI leads to an increase in our ability to align, alignment as an iterative problem…

~~Yeah it seems possible that some AGI systems would be willing to risk value drift, or just not care that much. In theory you could have an agent that didn't care if its goals changed, right? Shoshannah pointed out to me recently that humans have a lot of variance in how much they care if they're goals are changed. Some people are super opposed to wireheading, some think it would be great. So it's not obvious to me how much ML-based AGI systems of around human level intelligence would care about this. Like maybe this kind of system converges pretty quickly to coherent goals, or maybe it's the kind of system that can get quite a bit more powerful than humans before converging, I don't know how to guess at that.~~

~~https://www.lesswrong.com/posts/vSGTkkjrnJj4chFms/self-improvement-without-self-modification~~

~~you dont have to self modify to self improve~~

https://www.lesswrong.com/posts/oyK6fYYnBi5Nx5pfE/is-recursive-self-improvement-relevant-in-the-deep-learning

On the Feasibility of Recursive Self Improvement

Trillions to quadrillions of parameter models trained at the cost of tens of millions to billions of dollars do not seem particularly amenable to the kind of RSI Yudkowsky envisioned back in the day.

This combination of abilities would, in theory, allow an AGI to recursively improve itself by becoming *smarter* within its original purpose. A Gödel machine rigorously defines a specification for such an AGI.

Our most powerful models aren't designed but selected for via "search like" processes. They aren't by default well factored or particularly modular[14]. Even with advanced interpretability tools/techniques, it seems like it would be hard to make the kind of modifications/improvements that are possible in well written software programs.

An AI inspecting its mind, identifying flaws/inefficiencies and making nontrivial algorithmic/architectural improvements seems to not be particularly feasible under the deep learning paradigm.

- + a list of ways

I'm under the impression that while training more capable successor systems is feasible, doing so would incur considerable economic costs (and thus serve as a taut constraint to the "speed" of takeoff via such feedback loops)[17][18]. It does not seem to necessarily be the case — or even particularly likely — that the above kind of positive feedback loops will lead to discontinuous AI progress.

I think it may be possible that deep learning AGI would eventually be able to look upon its own mind and factorise it[19], but it seems like that level of cognitive capabilities would come well after AI has become existentially dangerous (or otherwise transformative)[20]. I currently do not expect seed AI style RSI to be a component on the critical path to existential catastrophe/"the singularity".

On the Viability of Recursive Self Improvement

My current belief is that even if seed AI flavoured recursive self improvement was viable, the gains that can be eked out that way are not as radical as imagined.

To summarise my position:

Intelligence has high "intrinsic complexity": returns to cognitive investment scale (strongly?) sublinearly with investment of computational resources, and this is a fact of computer science/mathematics/optimisation. You can't self improve to cognitive algorithms that attain linear or superlinear returns on cognition because no such algorithms exist (the same way no amount of algorithmic innovation would deliver a comparison based sorting algorithm that has better worst case complexity than
Θ
$($
n
$\log$
n
$)$
; no such algorithm exist.

It does not seem to me like recursive self-improvement is relevant in the deep learning paradigm.

+ more posts

If the aforementioned objections are correct, then insomuch as one's intuitions around foom were rooted in some expectation of recursive self-improvement and insomuch as one believes that the first AGIs will be created within the deep learning paradigm[23] then the inapplicability of RSI to deep learning should update people *significantly downwards* on the likelihood of hard takeoff/foom[24].

Recursive self-improvement and AI takeoff

Recursively self-improving AI is considered to be *the* push behind the intelligence explosion.

From "Hard Takeoff" (emphasis mine):

*RSI is the **biggest**, most interesting, hardest-to-analyze, **sharpest break-with-the-past**contributing to the notion of a "hard takeoff" aka "AI go FOOM"*, but it's nowhere near being the *only* such factor. The advent of human intelligence was a discontinuity with the past even *without* RSI...

I think there are other avenues for hard takeoff that don't hinge so strongly on RSI (e.g. hardware/content overhang[28][29][30][31]), but they also seem to be somewhat weakened by the deep learning paradigm[32][33][34] (perhaps especially so if scaling maximalism is true)[35][36][37][38]. That said, my broader scepticism of foom deserves its own top level post.

I am also not persuaded by the justification for foom credulity based on AlphaGo. I don't think AlphaGo is necessarily as strong an indicator for foom as suggested: AlphaGo was able to blow past human performance in the narrow field of Go via 3 days of self-play; it does not seem that general competence in rich domains is similarly amenable to self-play.

~~comments:~~

~~I think it's premature to conclude that AGI progress will be large pre-trained transformers indefinitely into the future. They are surprisingly(?) effective but for comparison they are not as effective in the narrow domains where AlphaZero and AlphaStar are using value and action networks paired with Monte-Carlo search with orders of magnitude fewer parameters.~~

~~We haven't formulated methods of self-play for improvement with LLMs and I think that's also a potentially large overhang.~~

~~There's also a human limit to the types of RSI we can imagine and once pre-trained transformers exceed human intelligence in the domain of machine learning those limits won't apply. I think there's probably significant overhang in prompt~~

engineering, especially when new capabilities emerge from scaling, that could be exploited by removing the serial bottleneck of humans trying out prompts by hand.

Finally I don't think GOFAI is dead; it's still in its long winter waiting to bloom when enough intelligence is put into it. We don't know the intelligence/capability threshold necessary to make substantial progress there. Generally, the bottleneck has been identifying useful mappings from the real world to mathematics and algorithms. Humans are pretty good at that, but we stalled at formalizing effective general intelligence itself. Our abstraction/modeling abilities, working memory, and time are too limited and we have no idea where those limits come from, whether LLMs are subject to the same or similar limits, or how the limits are reduced/removed with model scaling.

1. My conclusion was that AGI progress would be deep learning based into the indefinite future, not pretrained transformers

If deep learning yields AGI, the question is how far can its intelligence jump beyond human level before it runs out of compute available in the world, using the improvements that can be made very quickly at the current level of intelligence. In short sprints, a hoard of handmade constants can look as good as asymptotic improvement. So the latter's hypothetical impossibility doesn't put convincing bounds on how far this could be pushed before running out of steam. And if by that point procedures for bootstrapping nanotech become obvious, this keeps going, transitioning into disassembling the world for more compute without pause. All without refuting the bitter lesson.

I think you forgot one critical thing. Why does the normal argument for RSI's inevitability fail? The answer is: *it doesn't.*

Even though there is some research in the direction of a neural network changing each of its weights directly, this isn't important to the main argument because it is about improving *source code*. The weights are more like compiled code.

In the context of deep learning, the source code consists of:

- The code defining the architecture
- The code for collecting data (it can likely just hard code all of the training data if it is smart enough, but this isn't strictly necessary)
- The code for training
- The code utilizing the neural network (this includes things like prompt engineering, the interface to the outside world, sampling, quantization, etc...)

So the question is if a deep learning model could improve any of this code. The question of if it can improve its "compiled code" (the weights) is also probably yes, but isn't what the argument is based on.

Then this runs into the issue that I challenge there's just not that much gain to be made from such source code improvements.

It seems pretty clear to me that AI's could get really good at understanding and predicting the results of editing model weights in the same way they can get good at predicting how proteins will fold. From there, directly creating circuits that add XYZ reasoning functionality seems at least possible.

I don't actually share this intuition.

I don't think you can get the information of computing the gradient updates to particular weights without actually running that computation (or something equivalent to it).

And presumably one would need empirical feedback (i.e. the value of the objective function we're optimising the network for on particular inputs) to compute the desired gradient updates.

The idea of the system just predicting the desired gradient updates without any ground truth supervisory signal seems fanciful.
Ehh, protein folding feels equally fanciful to me, figuring out how the protein will fold without actually simulating the physical interactions.

Meanwhile we have humans already editing model weights to change model behavior in desired ways:
https://www.lesswrong.com/posts/gRp6FAWcQiCWkouN5/maze-solving-agents-add-a-top-right-vector-make-the-agent-go

This is a solid argument inasmuch as we define RSI to be about self-modifying its own weights/other-inscrutable-reasoning-atoms. That does seem to be quite hard given our current understanding.

But there are tons of opportunities for an agent to improve its own reasoning capacity otherwise. At a very basic level, the agent can do at least two other things:

1. Make itself faster and more energy efficient — in the DL paradigm, techniques like quantization, distillation and pruning seem to be very effective when used

by humans and keep improving, so it's likely an AGI would improve them further.

2. Invent computational tools: wrt

Most problems in computer science have superlinear time complexity

on one hand sure, improving this is (likely) impossible in the limit because of fundamental complexity properties. On the other hand, the agent can still become vastly smarter than humans. A particular example: the human mind, without any assistance, is very bad at solving 3SAT. But we've invented computers, and then constraint solvers, and now are able to solve 3SAT much much faster, even though 3SAT is (likely) exponentially hard. So the RSI argument here is, the smarter (or faster) the model is, the more special-purpose tools it can create to efficiently solve specific problems and thus upgrade its reasoning ability. Not to infinity, but likely far beyond humans.

...

each order of magnitude increase in compute buys (significantly) less intelligence; thus progress from human level to a vastly superhuman level just can't be very fast without a qualitative jump in the growth curves for compute investment.
Resource accumulation certainly can't grow exponentially indefinitely and I agree that RSI can't improve exponentially forever either, but it doesn't need to for AI to take over.

An AI doesn't have to get far beyond human-level intelligence to control the future. If there's sufficient algorithmic overhang, current resources might even be enough. FOOM would certainly be easier if no new hardware were necessary. This would look less like an explosion and more like a quantum leap followed by slower growth as physical reality constrains rapid progress.

I'm pointing at the possibility that we already have more than sufficient resources for AGI and we're only separated from it by a few insights (a la transformers) and clever system architecture. I'm not predicting this is true just that it's plausible based on existing intelligent systems (humans).

Epistemic status: pondering aloud to coalsce my own fuzzy thoughts a bit

I'd speculate that the missing pieces are conceptually tricky things like self-referential "strange loops", continual learning with updateable memory, and agentic interactions with an environment. These are only vague ideas in my mind but, for some reason, feel difficult

to solve but don't feel like things that require massive data and training resources so much as useful connections to reality and itself.

…

Each attempt is too expensive.

…
Direct self-improvement (i.e. rewriting itself at the cognitive level) does seem much, much harder with deep learning systems than with the sort of systems Eliezer originally focused on.

In DL, there is no distinction between "code" and "data"; it's all messily packed together in the weights. Classic RSI relies on the ability to improve and reason about the code (relatively simple) without needing to consider the data (irreducibly complicated).

Any verification that a change to the weights/architecture will preserve a particular non-trivial property (e.g. avoiding value drift) is likely to be commensurate in complexity to the complexity of the weights. So... very complex.

The safest "self-improvement" changes probably look more like performance/parallelization improvements than "cognitive" changes. There are likely to be many opportunities for immediate performance improvements[+], but that could quickly asymptote.

I think that **recursive self-empowerment** might now be a more accurate term than RSI for a possible source of foom. That is, the creation of accessory tools for capability increase. More like a metaphorical spider at the center of an increasingly large web. Or (more colorfully) a shoggoth spawning a multitude of extra tentacles.

The change is still recursive in the sense that marginal self-empowerment increase the ability to self-empower.

So I'd say that a "foom" is still possible in DL, but is both less likely and almost certainly slower. However, even if a foom is days or weeks rather than minutes, many of the same considerations apply. Especially if the AI has already broadly distributed itself via the internet.

The MIRI 2000s paradigm for an AI capable of self-improvement, was that it would be modular code with a hierarchical organization, that would potentially engage in self-improvement at every level.

The actual path we've been on has been: deep learning, scaling, finetuning with RLHF, and now (just starting) reflective agents built on a GPT base.

A reflective GPT-based agent is certainly capable of studying itself and coming up with ideas for improvement. So we're probably at the beginning of attempts at self-improvement, right now.

Inability of neural nets to quickly retrain a new larger versions will slower the speed on their progress and thus many competing AIs are more likely to emerge. It is less likely that they can merge later by "merging utility functions" as NNs have no explicit utility functions. Thus multipolar world is more likely.