Unhobbling Is All You Need? On Aschenbrenner's *Situational Awareness*

Ronan McGovern, Founder, Trelis LTD
ronan@trelis.com

Until recently, Leopold Aschenbrenner worked on OpenAI's safety team. He recently left and has published a ~165 page pdf on the near term future of AI.[1]

Situational Awareness – The Decade Ahead

That is the title of the paper, and it is wide-ranging. It goes all of the way from AI training techniques, through the computation power and pricing of the latest GPUs, all the way to the history of the Manhattan project and the Chinese Communist Party.

Superintelligence is coming this decade, according to Aschenbrenner. The US government should take control of all leading AI companies, so this once in a lifetime technology does not fall any further into the hands of our adversaries.

Aschenbrenner is not a "doomer" – i.e. someone who thinks a takeover by AI is imminent and will end humanity. Neither is he an "accelerationist" – at least not of the "e/acc" variety that wishes to power ahead with AI progress in all manners and in all locations. Rather, he believes that:

AI will – by the end of the decade – take over from humans (although – if managed correctly – not in a way that threatens humanity).

AI will be the greatest military weapon of all time, and history will be determined by the AI powers of the West versus the AI powers of China.

As a consequence, he believes:

AI activities should be taken under government control and made secret. Rather than allowing knowledge to freely be obtained by adversaries – as he believes it currently is – this knowledge and development should be locked down, just as in the Manhattan project.

Only with this approach, can the West stay ahead militarily and protect the free world.

I commend and recommend the piece.

1. Can we Extrapolate AI Progress?

Leopold feels that achieving super-intelligence is primarily a matter of quantitative improvements (namely, bigger faster computers). And, where qualitative discoveries are required for super

[1] Aschenbrenner, L. (2024). Situational Awareness: The Decade Ahead. Retrieved from https://situational-awareness.ai/wp-content/uploads/2024/06/situationalawareness.pdf

intelligence, he believes the incentives are there for those to happen as a matter of course. For him, it is merely a case of extrapolating the past into the future.

For me, there remains some thing (or perhaps many things) that are qualitatively different about human intelligence and what we have with AI today. This idea of building larger and faster computers has been impactful – much more so than was expected. But, I don't see how recent progress can be extrapolated to achieve the type of planning, hypothesising, criticising and decision making that humans can do today. I believe we can achieve such capabilities, but their discovery is neither inevitable nor predictable, nor will such discoveries be derived from what we have recently learned.

1.1 Aschenbrenner's pathway to super intelligence.

The core argument goes:

The size of AI models – as measured by the number of computing operations (e.g. multiplications) done to train them – has grown by over 1000X over the last few years.

The ability of these models has grown from passing pre-school tests to advanced high school tests.

If these patterns are to continue, AI models will – by 2027 – be trained with a further 1000+ times more computing operations (aka having "more compute") AND will be passing advanced postgraduate/expert tests.

Aschenbrenner further estimates that a) the AI having more compute, b) the AI being able to access more tools (e.g. use a computer) and c) incorporating various feedback and critical thinking loops, will allow these AIs to become drop-in remote workers.

1.2 Can intelligence be viewed in terms of computing power?

A core piece of Aschenbrenner's argument is that – to believe in Artificial General Intelligence (i.e. superhuman intelligence across a wide range of fields) – one only needs to trust the graphs of increasing computing power versus time. By this he means, to see AGI in our future, one only need extrapolate the progress on a) the size of computers, b) our algorithmic efficiency in using those computers and c) ancillary techniques like having AI use tools and closing the feedback loop to achieve error correction.

For Aschenbrenner, everything can be measured and projected in terms of "effective compute" or "effective size/speed" of the computers being used to train/run the AI model.

"our uncertainty over what it takes to get AGI should be over OOMs (of effective compute), rather than over years"

But, can capabilities be measured solely in terms of compute equivalent?

The human brain is still many orders of magnitude more efficient in terms of power consumption than computing clusters. By this logic, the "effective compute" of a human brain is still far beyond that of large language models. This suggests that human intelligence is not so much quantitatively

different, but qualitatively different. This suggests one cannot measure that qualitative difference with a single "compute equivalent" scalar.

AI models are saturating benchmarks, i.e. AI models are reaching near perfect scores on most of the tests that we give them (high school, university etc.). Yet, with those benchmarks saturated, we still know that there are large gaps in what a human and a language model can do (e.g. planning and decision-making in complex environments). Said differently, there are large shortcomings in how we define and measure performance. Increasing effective computing power does not address this core shortcoming that we cannot measure human intelligence or properly define what it is.

Aschenbrenner argues that improvements in effective computing power will come – not just from faster computers – but improvements in algorithms AND enabling AI to use computers and have error correction mechanisms. Whatever one might think of extrapolating raw computing power, it is even less clear how to justify the extrapolation – from the past – of as-of-yet unknown improvements and mechanisms

1.3 Is Aschenbrenner extrapolating or measuring compute growth?

Aschenbrenner argues one need only trust trend lines to predict future computing capabilities from the past. Yet, his paper appears to be doing something else. "Predictions" for 2026-2027 are not just a projection of past growth onto the future. Rather, these "predictions" incorporate the GPU and datacenter plans of Nvidia, OpenAI and Microsoft. This makes such statements on 2027 less of an extrapolation and more a recognition of current production plans. (And credit is due for highlighting and clearly presenting the scale of plans around datacenter buildouts that are forthcoming).

Moreover, beyond this decade, Aschenbrenner defaults to asymptoting progress not accelerating progress:

"In essence, we're in the middle of a huge scaleup reaping one-time gains this decade, and progress through the [orders of magnitude] will be multiples slower thereafter."

That surprised me! What is the core rational behind acceleration and then deceleration? What is his underlying theory here?

1.4 Unhobbling is all you need?

The paper is reductive because:

> A) It conflates computing power with intelligence

and

> B) where computing power falls short as an explanation for progress, these shortcomings are overcome with what Aschenbrenner calls "unhobbling" – unknown-as-of-yet techniques such as allowing AIs access to computers and providing them with a system 2 feedback loop for error correction.

Per Aschenbrenner:

"there's likely some long tail of capabilities required for automating AI research—the last 10% of the job of an AI researcher might be particularly hard to automate. This could soften takeoff some, though my best guess is that this only delays things by a couple years. Perhaps 2026/27-models speed are the proto-automated-researcher, it takes another year or two for some final unhobbling, a somewhat better model, inference speedups, and working out kinks to get to full automation, and finally by 2028 we get the 10x acceleration (and super-intelligence by the end of the decade)."

Quite simply, we cannot explain what human intelligence is, or how to replicate it. Yes, increasing computing speed and model size is leading to meaningful performance improvements. But, to say that we can achieve capabilities that we can't describe (human intelligence) via methods that we have yet to discover (unhobbling) is not a meaningful prediction.

Further Remarks

What is a drop-in remote worker?

To a degree, we already have those. Servers (that don't necessarily use large language models) do plenty of automated, agentic and remote work already.

AIs doing AI research

Aschenbrenner suggests, on the matter of AI being able to do AI research:

"And the job of an AI researcher is fairly straightforward, in the grand scheme of things: read ML literature and come up with new questions or ideas, implement experiments to test those ideas, interpret the results, and repeat."

I agree that AI (and deterministic software) can be used to further accelerate research. As Aschenbrenner separately points out, AI researchers have the exact knowledge set needed to know how to automate their own AI research.

But, Aschenbrenner acknowledges the following:

> *"Still, in practice, I do expect somewhat of a long tail to get to truly 100% automation even for the job of an AI re- searcher/engineer; for example, we might first get systems that function almost as an engineer replacement, but still need some amount of human supervision."*

If that's the case, surely – at every step of AI improvement – AI will also allow the human researcher to do much more? Why does there have to be a limit to how useful the tool can be for the human researcher? Why does it have to asymptote? If it does asymptote, what is it that unlocks the automation of that very last percentage? What is that "unlock"? If it is something not yet known, then how can its discovery be predicted or extrapolated?

Moreover, recursive planning and decision making is unstable with next token prediction – absent an error correction mechanism. What is the underlying error correction mechanism that we will discover? Can it also be simply extrapolated from the recent past? Why is that likely? Why might such a discovery not equally well come from elsewhere…? For example, from a person or group not thinking about scaling compute?

What it is that would take the AI model from co-pilot to pilot. What is it? If we don't know what it is, how do we know how and when we'll get there? I don't see that qualitative explanation when I look a linear extrapolation graph.

Lack of Divergence in Performance of Models by OpenAI, Google, Mistral and Anthropic

It's interesting how the performance of AI models by OpenAI, Google, Mistral and Anthropic are largely the same. The lack of divergence in capabilities suggests that they are all on to the same thing – using more powerful computers to run neural nets has provided stronger model capabilities. This lack of differentiation supports (although certainly doesn't prove) the hypothesis that there has primarily been just one discovery here.

On this point, Aschenbrenner remarks:

"I'd expect labs' approaches to diverge much more, and some to make faster progress than others—even a lab that seems on the frontier now could get stuck on the data wall while others make a breakthrough that lets them race ahead."

But why hasn't there been divergence so far? Aschenbrenner claims that the top companies have only stopped publishing their work recently, so we are only about to see divergence.

I am skeptical of this. The core transformer paper is from 2017, and we are seeing companies release models at a clip of one or two per year. To the degree the large companies have stopped or slowed publishing, this has been a trend over the last few years, not just the last six months. And, over even the last six months, there are smaller companies (Mistral, Databricks, Snowflake) releasing models that are close to the leading edge. There is also a lot of movement in staff between large companies, as there are no non-competes in California. My guess is that knowledge is flowing between companies and also into the public domain.

When an industry is experiencing progress by growing (primarily) along one dimension, namely, the dimension of increasing computing power, there is a strong incentive to focus on that scaling so as not be left behind. Diversifying to other approaches may not be game-theory optimal because discovery is a fat-tailed type of game requiring a large sample size (i.e. a large number of independent research efforts/experiments).

That said, recent progress in AI is fast and it is obvious. This means AI is attracting talent and there is there is both an urge and unwritten rule to follow the AI equivalent of Moore's law. And, when humans focus – as they did in the Manhattan project – it can be easy to underestimate what can be achieved in terms of progress.

So, on the one hand I expect Google, OpenAI and Microsoft to mirror one another in capabilities. On the other hand, it's hard to rule out some divergence – on the basis that Google and Microsoft are a bit like Bell Labs, i.e. they have a meaningful enough number of independent research projects to play the fat-tailed, large-sample-size, kind of game. Still, these organizations have a strong commercial bias that makes it hard to truly maximise the number of independent initiatives.

AIs with Internal Thoughts

It's an interesting research idea to write out some conversations including internal monologue and see if training on those helps the model to perform better. Someone must have tried this. Although, I have previously written about humans have the advantage of their inner thoughts largely being free. Any "inner thoughts" by LLMs are likely to be heavily censored.

Aschenbrenner brings up this idea of developing an AI model to have inner thoughts:

"For example, I think it's pretty plausible that we'll bootstrap our way to AGI via AI systems that "think out loud" via chain-of-thought."

I like the hypothesis, and we should try it! Does this work? I don't know. Could there be many other ideas to try? Definitely. Are those ideas extrapolated/derived from the idea of using bigger/faster computers to train/run AI models? – I don't think so.

Does Nvidia Price Continue to Skyrocket?

Aschenbrenner's conclusion is that Nvidia is still undervalued. Perhaps he is right. Compounding growth is easily underestimated.

On the other hand, perhaps we underestimate the force of competition coming in from ARM and other chip companies. Perhaps we also underestimate how the efficiency of inference and building models will improve (the Phi-3 3.5 billion parameter model is not far off being as good as Llama 3 70B – so there is room for massive model compression, which reduces the GPUs needed).

My best guess is therefore to opt for a straddle! Either Nvidia will do a further 5X in the next three years OR it will fall to half of its current value or less. That's a guess, I haven't made any such trade (yet).

Natural Gas to Power Clusters

Aschenbrenner argues that natural gas is the most practical way of building computing clusters at scale to build out AI. He would like to have carbon free energy, but advancing and controlling superintelligence is more important.

I wonder if he has written a piece on carbon dioxide and its impact on climate change? I would be interested to read more.

I think super intelligence will require qualitatively different knowledge than just scaling compute, but I agree with him that advancing knowledge, intelligence (and GDP – I would add) is underrated relative to concerns over natural gas.

In Conclusion: Breadth and Depth

Aschenbrenner's depth and breadth of topics is impressive, much broader than my own and I've been on the planet for longer. Further, I have no doubt – given he worked at OpenAI – that he has

much greater specific knowledge of OpenAI's capabilities today. Meanwhile, I only have what is public.

It has been over a year now (yes, the proverbial self-acclaimed overnight expert) where I've been working with AI and I've heard – "Well, maybe there are capabilities in OpenAI that you just don't know about". While I can't rule that out, the landscape is now so competitive that this has become less likely (because companies are under commercial pressure to release their best models… the price matching and price cutting behaviour is indicative of that).

If I'm wrong, it's perhaps more likely because:

a) We just don't understand intelligence or human intelligence. And, when faced with what has been significant progress over the past years, it's hard to say anything but that artificial human intelligence is coming soon and inevitable.

b) human brains are purely predictive. Hypothesising and criticising are figments of my imagination that – in reality – are simply a form of pattern matching. The data defining those patterns need only be captured, and the models need only be large enough to smoothly fit those patterns, and the compute large enough to achieve that smooth fit through optimization.

c) maybe the idea that intelligence is just pattern matching is the only big idea. Once we find that idea (and apply lots of compute), there's nothing else for us to independently discover because everything else can be trivially derived/discovered. To date, I haven't seen anything to suggest there is a theory of everything. Even the algorithmic improvements and error correction loops that Aschenbrenner suggests seem independent (and not extrapolatable) from the idea of using more compute. Our history so far is one where ideas and discoveries bear a strong degree of unpredictability. You do well with discoveries not by concentrating bets but by diversifying (more specifically, lowering the costs of experimenting). But, maybe this is wrong, maybe artificial intelligence is not many ideas, but rather Artificial Intelligence (AGI) is just one. So we should concentrate humanity's bets and lock up innovation in one single government Project if we believe it is "The One".

d) even if – to date – computers and AI have created new levels of abstraction and new jobs, we somehow reach a point at which this is no longer the case. There is no further abstracted management task because we have "automated 100%". I don't see a way to explain what this turning point might be or how it would work. But I can't rule it out.

e) I may be right that future progress cannot be extrapolated from the past, but, if all efforts are wrapped into one Western AI project, just having that level of focus – that level of Manhattan intensity – might be enough to radically advance artificial intelligence. One should not underestimate the progress (and damage?) possible in coordinated mission-driven efforts. And that would be the classic question of whether one should centralise or diversify.