Situational Blindness: A policy highway to AGI-induced catastrophe

Beginnings

I went to sleep after I read the full “situational awareness” blog post (https://situational-awareness.ai/). I dreamed of my grandmother. She video called me and my dad, and of course I got shivers down my spine, because my grandmother shouldn’t be calling anyone; she passed away last year. In the call, just as in life, she was warm and articulate, and immediately jumped into topical conversation in her pleasant and familiar way, while she bustled around her house cooking. She looked at me kindly through her horn rimmed glasses, her extensive smile lines crinkling as she spoke. And she slowly faded from view as the call went on, until the video was of an empty kitchen and the call hung up.

I woke up with tears streaming down my face. Because of course I love my grandmother a great deal, and I miss her. The dream also left me with a profound sense of eeriness, dislocation, and something else, a nagging feeling that it was connected to what I had been reading, that I should pay attention to the symbolism underneath the surface of my dream.

I did so, and this essay is the result.

Part 1: Something Familiar (Introduction, and commentary on Leopold Aschenbrenner’s Introduction section)

Leopold Aschenbrenner is a young man. He writes with the barely contained breathless enthusiasm of the true believer who is stretching out his hands to a crowd of onlookers, ready to pull them into giddy flights of intellect that he has trailed in the morning sky. He lets you know, right there, at the beginning, that you are soon to be an initiate to secrets only the elect few

have reckoned with. As well put together as his multi-chapter writing is, its most interesting aspect is the insight it seems to lend into his psychology and that of his fellow aspirants. To reframe a line from the essay: if these are the attitudes of the people in charge of developing the world's most advanced technology, we're in for a wild ride.

I say this not to be flippant but because I share a certain kinship with Leopold. When I was younger, I possessed an absolute certainty in my own cleverness and ultimately, correctness. Being correct was more important to me than almost anything else. My own certainty was infectious; it earned me positions of authority that more circumspect people missed out on, and often my results justified these promotions.

Yet as I've cast a critical eye on some of my own 'correct' conclusions over the years, I can see the seams: here a logical leap, there an obvious patch of emerging complexity I disregarded because I assumed that when the time came I would simply 'figure it out.' I got away with my assuredness because my thoughts *were* often reasonable and correct, but the domain I worked in also lent itself to rigorous analysis and had a lot of history and precedent and connections to academic literature and commercial implementations that made it tractable for a clever kid to contribute to.

Leopold's confidence is familiar to me but seems less merited. The domain he's working in is in its infancy. His own experience is limited. He has trendlines but no context or real precedents; the precedent he chooses is flawed. He ignores wide swathes of crucial social, economic and political theory. His geopolitical sections are jingoistic caricatures that nonetheless read as self-assured as his technical sections. Despite writing chapters of text, he rushes to his conclusions.

To get to the truth then, we need to slow down, to work through and elaborate the chain of reasoning as Leopold's posited AGI might do, in a way that he himself does not. Here's a roadmap for that:

- In Part 2, we'll examine how Leopold's projections fail to consider any of the obvious social implications of the the timelines he proposes, which will confound the projections themselves
- In Part 3, we'll look at how he sets up potential obstacles to his proposed timeline as straw men that he can blow over with mere intuition and builds a scary historical analogy based on a misapprehension of the way knowledge diffuses in his own field
- In Part 4, we'll review how his proposal for the government to subsidize the infrastructure of the US' biggest and most profitable tech companies in the name of democracy would actually lead to a democratic collapse at home and a destabilization of democracies abroad
- In Part 5, we'll review how his proposed military-grade secrecy around both AGI and AI safety would greatly diminish global security in relation to AGI hacking to no purpose (as the US is an irredeemably soft target for nation state hackers), and how his favored foreign policy would unite the world against the US
- In Part 6, we'll use a lens of fragility to show how Leopold's policy suggestions are more likely than any other policies to *cause* the very catastrophes he fears
- And in Part 7, we'll propose an alternative to his reductive and antidemocratic approach, that has some chance of being successful

Part 2: Burying the Lede (Commentary on "From GPT-4 to AGI: Counting the OOMs")

This chapter makes two key arguments:
Increase in computational power and algorithmic efficiency → Rapid advancement in AI capabilities

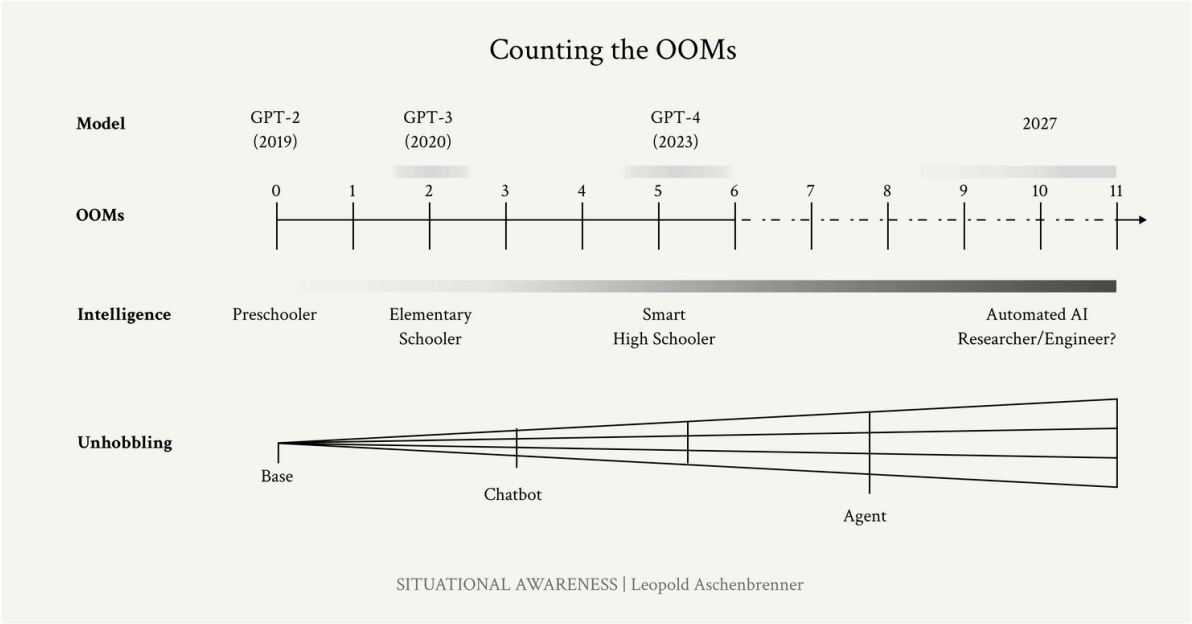
Continuous doubling of computational resources and improvement in AI models → Significant leaps in AI's cognitive abilities, approaching AGI

As we'll see, these predictions, while neat in a mathematical sense, completely ignore the socio-economic consequences of an AI rollout, and thus become implausible on their face.

The Case of the Missing Professions

It might surprise you that I'm not going to spend time questioning this chapter's extrapolation of potential intelligence gains. I think that for the sake of evaluating the quality of the argumentation, we should just accept the proposed curve as a given, and press on.

As a reminder then, here is the graph from the essay.



Summary of counting the OOMs.

I would argue that the current state, if anything, undersells it. Social media has really addled our youth. GPT-4 is undoubtedly leagues beyond the average high schooler. In this scheme, AGI (or some close approximation of it) arrives in 2027 as personified by an Automated AI Researcher / Engineer.

Does it feel like this chart is missing some professions? Presumably, as we've seen in the past 10 years, there will be incremental deployments of the tech (like, for example, in Office

365, Windows 11, and on all Apple devices that support Siri) to support automation of other job categories. This follows from game theory:

1. *If any one AI company releases an incremental model advance, then all others must follow suit (we've seen this again and again with GPT, where slightly better versions mysteriously drop the day before or after the competition nips at Open AI's heels).*
2. *If any one enterprise adopts a radical efficiency-improving technology, all other enterprises must adopt the same or be defeated in the marketplace.*

And this game theory starts to lead to some troubling questions. What happens to our remaining clerical workers long before the AI Researcher strides on scene? What about our tutors, copywriters, journalists, junior data analysts, and customer service representatives? How about lawyers and doctors and coders?

Social Construction of Consequences

What would serious consideration of the social implications of having AI competition for many of our jobs look like? To make a tentative answer, let's draw on a popular theory, that of "Social Construction of Technology" (SCOT). SCOT explores how technology is shaped by social factors, rather than developing independently based on its own logic. It challenges the idea that technological progress is inevitable and autonomous. Some of SCOT's key tenets include:[1]

- **Interpretive Flexibility:** The idea that different social groups can have different interpretations of what a technology is and what it should do.
- **Consideration of Relevant Social Groups:** The idea that there are many stakeholders that must be considered in the analysis of a technology's development and adoption.

[1] https://en.wikipedia.org/wiki/Social_construction_of_technology

- **Closure and Stabilization:** The idea that as a result of group consensus, technology designs become stable and fixed.
- **Wider Context:** The broader social, economic, and cultural context in which a technology develops. This includes factors like political climate, economic conditions, and cultural values, which all play a role in shaping technology.

Now, I'm no expert in this particular theory, but I think I can use its factors as a lens to make some educated guesses. Let's therefore test it out on a particular concrete scenario: The elimination of 40% of customer service workers and medical clerical staff at US insurance companies within a two year period of time.

In terms of **interpretive flexibility**, I would say first that the involved workers are going to view this development as an existential threat. They probably would have started by viewing AI as a "helpful productivity aide" and would imagine that it should make their lives easier, not obsolete and financially ruin them. The idea they will rapidly approach is that the technology should not be doing *anything*, that it should be outlawed.

Similarly, after any sort of disruption at this scale, **relevant social groups** start to become "everyone who works in an office," who may become politically supercharged by the prospect of the middle class being completely denuded. These groups will lobby alongside the affected workers out of pure self interest. At a minimum, they will want large scope restrictions on deployment. Probably, they will want to ban further development of the technology.

Closure and stabilization, under this scenario, starts to look a lot like technology use and development restrictions, anti-technology "back to basics" movements, and in the worst case scenario, destabilization in the form of radical political swings and perhaps even violent uprisings. The **wider context** is that, presumably, this would be happening on a global scale, with unpredictable results in each country that AI was deployed in.

If you find this particular scenario's details implausible, feel free to pick different details and see if you think the calculus changes much. Some examples of alternate scenarios might

include: 50% of artists and photographers unemployed, overall unemployment goes from 4% to 10% in the US, 30% of all writers laid off, etc.

So that's it, our glancing analysis on some potential socio-technical factors at play. But even such a cursory look reveals that the essay's "Racing the OOMs" graph is unrealistic, because it has large unstated assumptions around technology adoption. If adoption does not happen (for any reason), then companies cannot continue to spend vast sums of money on it. Corporate money does not come from thin air; it comes from customers paying for services and the stock market's future expectation of value. If customers boycott corporate services, or the services are outlawed, then revenue collapses and future stock market value washes away with it.

Summary

My point, of course, is not to convince you that a particular scenario is going to happen. I actually believe that, to our detriment, AI adoption momentum will be very difficult for social factors to overcome. What I want to highlight is simply that the essay itself *seems to completely ignore the whole socio-technical dimension*, the very idea that technology adoption might not happen in a smooth way. It is technologically determinist: "It can be built, therefore it will be built, and it will be deployed. The graph is going up. It will follow the graph!"

In fact, it seems like the essay is burying the lede: Emergent socio-technical complexity as a result of rapid economic disruption *is the crucial story of AI*. A pure curve for model capability does not exist in the face of such emergent complexity; it is confounded.

So to review, if this section's arguments were:

> Increase in computational power and algorithmic efficiency → Rapid advancement in AI capabilities
>
> Continuous doubling of computational resources and improvement in AI models → Significant leaps in AI's cognitive abilities, approaching AGI

My model is more like (assuming the same timeline):

> Increase in computational power and algorithmic efficiency → Rapid advancement in AI capabilities → Emergent Complexity Associated with Massive Socio Economic Upheaval → ???

As we will see in later sections, if you make the merest accommodation for AI as an economic force, the calculation around 'situational awareness' shifts radically compared to the author's projections. Yet before bringing this point home, it is important to look at the quality of the technical analysis being put forward to support the author's self-compounding timeline.

Part 3: Curve Balls (Commentary on "From AGI to Superintelligence: the Intelligence Explosion")

The core arguments of this section are as follows:

> Achievement of AGI → Automation of AI research
>
> Automation of AI research by AGI → Exponential increase in AI capabilities, resulting in superintelligence

Do you ever find yourself wondering why we don't distill all technological advancements into a singular beautiful object? Like, why don't we have a car that is electric, that also flies, that can also go underwater, that is also a supercomputer, and that can deploy advanced weaponry at the flick of a switch (and also happens to be a pen, a sword, and a shockingly good harpsichord)? Who wouldn't buy that?

Similarly, have you ever wondered why evolution hasn't gifted humans with infinite memory capacity as well as telepathic abilities, the muscular strength of chimpanzees, and concentration twice as good as the average Zen buddhist monk? Doesn't it seem like such an Ubermensch would "win" evolution?

Or, perhaps you're a programmer who yearns for a language that is simultaneously type safe, completely expressive (though succinct and not verbose while still being highly readable),

memory-safe (but with no pauses), faster than hand-rolled assembly and optimized for deployment on everything from hugely parallel supercomputers to calculators? Why doesn't some genius just take all the advancements in all the languages written to date and combine them into this one perfect programming language? What is he/she waiting for?

Probably, as you're reading these examples, you are building up your own intuition about why they are absurd. If computer science (or in the case of the section under evaluation, ML) is an unfamiliar domain to you, however, you don't know what's possible and what's not. That's why the final example might seem obscure to you, whereas a programmer might be groaning in familiarity. Thus, before diving into specifics around ML intuitions, it's helpful to express in lay terms some of the logic we might use to assess whether things are possible or not. Below is a list of considerations I commonly use when assessing technical viability. These are some of the troublesome properties of reality that tend to get in the way of building truly perfect things, or in other words, achieving true optimality. I've drawn them from my reading and experience over the years. Such include:

1) Trade-offs / incompatibilities - it often seems that optimizing on one dimension causes other dimensions to suffer. The amphibious car handles a lot more poorly than the normal car
2) (Human) Rules - We often impose constraints on the systems we design that hinder their potential. If our car can go one thousand miles an hour, but we make a requirement that it must be able to come to a complete stop within one hundred feet of the brake being pressed, then our maximum top speed actually becomes infeasible[2]
3) Physical Laws - Hard universal constraints have put a damper on things like perpetual motion machines and supersonic tunnel borers and limit our ubermensch potential because of impractical dietary energy requirements

[2] Safety factors and usability are often studied in the field of Human Factors (https://en.wikipedia.org/wiki/Ergonomics) but constraints can also be socially constructed

4) Combinatorial explosions - There are certain classes of problems that, while seemingly simple (What is the optimal route for a traveling salesman who is visiting 7 cities?) become nearly impossible to solve efficiently as soon as the number of terms to consider goes beyond the low single digits[3]
5) Weakest links (in a chain) - many classes of solutions are limited by the weakest link involved. If we have a supercar that can travel at one thousand mph but the tires melt at three hundred mph, we're stuck (!)[4]
6) Lack of clear problem definition - if we can't express in a concrete way what we actually are setting out to achieve, then it becomes difficult to undertake a focused investigation
7) Uneven technological development / absence of theoretical priors - it would have been pretty hard for bronze age folks to invent quantum mechanics, because they weren't looking at the world on a quantum scale (they had no tools to do that) and had no theoretical basis to even imagine the quantum world existed. It would take a whole revolution in basic science to even identify quantum problems as worth considering[5]
8) (Required) Creativity - Some optimal solutions are so far outside of current thinking that they are never considered, or are only considered once in a hundred years
9) Cost - The cost of many technologies does not exceed their benefits. We could probably produce fully lifelike bionic arms if we wanted to, but if they each cost $500k and offered limited incremental mobility compared to more basic models, they wouldn't see wide deployment

If this list seems intuitive but somewhat arbitrary, stick with me.

---

[3] There are whole fields of complexity theory and combinatorial optimization devoted to hard problems that often seem simple at first glance. This article gives some flavor on the tremendous variety of techniques that can be employed: https://leeds-faculty.colorado.edu/glover/Publications/TSP.pdf
[4] A great elaboration around weakest links is the "Theory of Constraints" by Eliyahu Goldratt
[5] The literature in the Philosophy of Science and Technology covers a lot of ground in relation to constraints on innovation. Wikipedia has a fairly gentle introduction (see specifically: Observation inseparable from theory): https://en.wikipedia.org/wiki/Philosophy_of_science

Automated AI Researchers are All We Need? (for exponential progress)

Now, the task before us is to assess whether the essay's arguments around an Automated AI Researcher are well-supported, as according to the author, an optimal such engineer is what drives exponential progress toward AGI and SGI. To do this, we need to unpack the author's definition of an Automated AI Researcher and see if there are any obvious areas of consideration he might have missed. We can also perform two tests against the reasoning he puts forward around bottlenecks: 1) **Breadth**, whether there are any salient gaps in the range of bottlenecks he considers. 2) **Depth**, the degree to which he supports his contentions with evidence.

The Definition of Automated Research Engineer and Quality of Discussion Around This

According to the article: "*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... "*

The author is keen to note: "*It's worth emphasizing just how straightforward and hacky some of the biggest machine learning breakthroughs of the last decade have been: 'oh, just add some normalization' (LayerNorm/BatchNorm) or 'do f(x)+x instead of f(x)' (residual connections) or 'fix an implementation bug' (Kaplan → Chinchilla scaling laws). AI research can be automated.*"

So, an Automated AI researcher is an entity that: 1) Can extract concepts (questions and ideas) from literature 2) Devise experiments based on the concepts extracted 3) Implement experiments in code 4) Share results with swarm 5) Repeat

This does sound simple enough. Without much further discussion then, the essay launches into calculations around how many Automated AI Researchers it will take to turn the whole field exponential. Unfortunately, there is an obvious and troublesome question that goes unaddressed: What guarantees the quality / usability of the source material?

At the moment, AI literature comes from journals and conferences (keep this in mind for later sections, because if public research into AI is made illegal, it seems like scaling suddenly becomes a lot harder). We must therefore immediately ask: Does public AI research continue to be the source of Automated AI researcher ideas, or do agents ultimately just rely on the ideas of other agents?

If public AI research continues to be a major source of inspiration, there is now a branching path. One branch is that public research falls increasingly behind the closed source state-of-the-art and becomes useless to the Automated AI Researchers. Another branch is that public research keeps up. Which one is more likely?

At the moment, as the essay acknowledges, public research is starved for compute. The typical flow for ML research is to test a small hypothesis on a small scale → scale it significantly (checking charts along the way) → deploy it in a big model (and hope it still works). What we're seeing right now is that only the best funded labs can even routinely test small hypotheses. We could imagine that the macro result of this is that in the future there might be a great deal of fruitless 'noise' in the research stream, because there would be academic incentives to test lots of shallow but differentiated ideas as compared to building extensively on someone else's work / going deep in a research tree (researchers always look better if their ideas are novel). Shallow ideas would be useless / inapplicable to Automated AI Researchers if the tech tree of the automated parent was already deep and premised on strong assumptions unknown to the broader researcher community. Given the dearth of compute and perverse academic incentives, this branch seems realistic.

The second branch assumes that public research keeps up. Perhaps governments provide huge compute grants to researchers. Perhaps closed AI companies share just enough details about their research that they can steer public research in 'compatible' directions. While possible, this branch is much slower than a pure exponential, because human researchers are essentially 'in the loop' generating the ideas that the Automated AI Researchers are using.

So, what of the branch where Automated AI Researchers provide their own inspiration? If we believe that Automated AI Agents are particularly creative, then maybe they can press on without human input or with very limited human input. But if they are not, then they will also stagnate. It might be worth asking the question: If LLMs are truly creative, why haven't we seen major breakthroughs in every field of basic science? After all, current models are trained on vast corpuses of scientific research. Why can't we just ask: "Give me a major new insight into quantum gravity?" or "What is the likeliest path to a room temperature superconductor?" and get a revolutionary idea? [6]

Again, this is not so much a fleshed out counter argument as much as a sample of what deeper consideration of the issues raised by the author's definition might look like. What is striking about the section is that the author doesn't consider anything like this. He simply assumes that problems like 'creativity' are overcome, without explaining what would be required in order to do so compared to the current state-of-the-art. He writes (regarding Automated AI Researchers):

"they'll be able to read every single ML paper ever written, have been able to deeply think about every single previous experiment ever run at the lab, learn in parallel from each of their copies, and rapidly accumulate the equivalent of millennia of experience. They'll be able to develop far deeper intuitions about ML than any human."

Does this make sense in a world in which Automated AI Researchers are drones who generate repetitive research based on a well of public scholarship that has run dry? Why should we believe the author's scenario over any other arbitrary scenario that sounds plausible?

[6] I'm definitely not the first person to think about this issue. For a great explanation (and a generally interesting blog), see : https://www.lesswrong.com/posts/EMZeJ7vpfeF4GrWwm/self-supervised-learning-and-agi-safety#vRfNdRe8Gzz9QhFYq

Breadth and Depth of Argumentation around Bottlenecks

But what of the 'hacks' that the Automated AI Researcher is testing? What types of factors might interfere with perfectly ideal / optimal performance? Now we can bring back in our earlier heuristics, as a way of testing the breadth of the author's considerations. What follows is a brief elaboration on how these common heuristics might apply to 'Hacks' developed by an Automated AI Researcher:

1. **[Trade-offs / incompatibilities]** Hacks can be incompatible with each other for technical / architectural reasons. For example, recurrent neural networks (RNNs) process data sequentially whereas transformers use parallel processing. Therefore, a hack for a transformer architecture could be inapplicable to an RNN.
2. **[Rules]** Hacks may not be feasible because they induce undesirable outcomes in terms of the rules people set. For instance, if a model architecture is highly performant but unable to be red-teamed (tested adversarially for problems) because its design is incomprehensible to humans, then it will probably be ruled out[7]
3. **[Physical Laws]** Hacks may be theoretically possible but physically unsuited for the hardware that they are running on, making them actually infeasible. For example, a hack which required running all calculations on analog circuits would fail on digital NVIDIA chips, or at least be so painfully slow in emulation as to be useless
4. **[Combinatorial Explosions]** There may be a combinatorial explosion of experiments required to test for compatibility among hacks. If certain truly innovative hacks 'break' scaling laws, large test runs will be required to

[7] This is to say nothing of the challenge of aligning all involved agents with positive human-like goals, a problem which we will come to in the author's relevant chapter.

re-establish performance expectations and compatibility with existing work, greatly slowing progress

5. **[Weakest links]** It is possible that certain elements of model design will become 'weak links' only at certain scales or with certain types of data, making the final performance disappointing (while wasting huge amounts of money). An example might be that a hack that greatly improved a model's proficiency at visual reasoning tasks might put a ceiling on its verbal reasoning performance under a given architecture
6. **[Lack of clear problem definition]** It may be very hard to tune an automated research program that balances conflicting or underspecified goals while limiting the number of experiments it runs. For example, the goal of 'near real-time responses to user queries with human-like intonation' may be somewhat incompatible with the goal 'enhance strategic planning capabilities' and the associated research streams may develop incompatible architectures. What research is then necessary to marry up these branches, and is that more desirable than simply focusing on one goal over the other or using multiple models alongside each other? [8]
7. **[Uneven technological development / absence of theoretical priors]** ML does not have a well-fleshed out internally consistent set of axioms that describe the expected behavior of ML systems resulting from different architectural and training data choices. The situation for the ML practitioner is like that of an engineer trying to design a plane without any understanding of basic

[8] This may be a great example of where "Ashby's Law of Requisite Variety" comes into play. This law proposes that for a system to effectively manage and control its environment, the system's internal complexity (variety) must be at least as great as the environmental complexity it needs to manage. If there are an infinite number of potential competing research directions, the control program may need to be very complex indeed. See : https://www.businessballs.com/strategy-innovation/ashbys-law-of-requisite-variety/

aerodynamic principles. Without a grounding in theory, guessing ('ML intuition') is the only way for agents to decide the best way forward. On the whole, if breakthroughs are required either in our understanding of ML or in other fields before progress can be made, then the exponential breaks

8. **[Creativity]** Breakthroughs are often achieved through truly novel architectures as opposed to hacks. As mentioned, it is not clear that current LLMs have the reasoning capabilities to design totally new architectures
9. **[Cost]** If brute force approaches must be used, without an underlying theory being developed, then costs will be astronomical and progress will be slow. For example, if we don't know the right types of data to train new model architectures with, then iterating data mixes alone could be incredibly painful

To gauge the strength of the author's reasoning in relation to these types of problems, let's now compare our list to the essay's own list of potential identified bottlenecks. The more classes of objections covered (**breadth**), and the more well-reasoned empirically supported arguments (**depth**), the stronger the overall analysis must seem.

Essay item	**Our item**	**Essay Argument**	**Essay Citation?**	**Our Argument**
Limited Compute	Cost (9)	ML Engineers will have incredible intuition which will let them search the space effectively	No	Intuition isn't enough in the face of a search space this large - you need a sound theoretical basis and you can't brute force it
Complementarities / long tail (last 30% is much harder than first 70%)	Weakest links (5), Uneven technological development / absence of	We will just figure out the long tail, even if this pushes the overall estimate	No	Scaling falls apart if breakthroughs in other (non-exponential) fields are required or if rigorous theory

	theoretical priors (7)	a couple of years		development in ML is necessary to reduce search space. Back-tracking is extremely costly if weak links are discovered late
Inherent limits to algorithmic progress	Trade-offs / incompatibilities (1)	My own intuition says that current schemes are inefficient and inelegant and we'll come up with something better	(Biological reference classes? - but no reference)	There probably is low hanging fruit, and also we don't know what we don't know, so we can't dismiss limits out of hand
Ideas get harder to find, so automated AI researchers will merely sustain, rather than accelerate … progress	Creativity (8)	Increase in research effort required < effort required to sustain progress, sustained progress is an unlikely equilibrium point	No	There is no evidence provided that current models have the requisite creativity to generate useful new architectures / sustain progress.
Ideas get harder to find and there are diminishing returns, so the intelligence explosion will quickly fizzle	Lack of clear problem definition (6), Combinatorial explosion (4)	Initial exponential progress makes this irrelevant	No	It's hard to chart a research path if what models need to get to the 'next level' is unclear. Underspecified needs and goals would lead to an explosion of 'required' experiments
	(Human) Rules (2)			Hacks may be disqualified due to inducing subjectively or normatively undesirable behaviors
	Physical Laws (3)			The class of most efficient hacks may require hardware redesigns

Looking at the table, it is clear that the author's breadth is reasonable in relation to our own experientially driven sample of potential bottlenecks. Unfortunately, while there is good coverage, no higher order reasons for choosing these particular bottlenecks are provided (for example, the author does not adopt any particular accepted framework to motivate these). Moreover, the importance of said bottlenecks is not weighted, even in a simplistic way. The author's main citation in the section provides a much more detailed model of bottlenecks than what the author presents, but he makes no attempt to describe said article in any detail or relate it back to his own analysis. This is problematic because the cited author's analysis includes classes of problems (for example, hardware bottlenecks) that the essay does not consider, making it seem as if Leopold is hand-waving away significant concerns that he should be aware of. Similarly, the remaining citations are abstract econometric analyses that cite significant weaknesses in data quality and predictive ability as limitations. The lack of weighting and higher order reasoning around the essay's choice of bottlenecks makes it difficult to relate them to other works on these topics and thus to assess if they are critically accepted or not.

The obvious problem in the analysis is, of course, with the depth. As the saying goes, "assertions made without evidence can be dismissed without evidence." The essay does not support its assertions with evidence, and its assertions are not self-evident. As we've shown through our tabular thought exercise, there is a different but equally plausible set of qualitative assertions that might put the author's predicted outcome ("exponential growth achieved via Automated AI Researchers in a short time period") in some doubt.

This type of intellectual sloppiness might be excusable if the author made very clear that he was writing a simple opinion piece. But that is not the case here. From the very outset, Leopold has set himself up as one of a precious few visionaries who is sharing sacred truths with a lay audience. He has built an argument that is incredibly elaborate in places. But if the crux of the argument rests on unsupported assumptions, is it really worth our time?

The reader may object here that many of the essay's 'intuitions' are as-yet unprovable and that boldness requires putting these kinds of claims forward so that they can be tested. The simple counterargument is that there are numerous ways that intuitions may be supported empirically and qualitatively, and that Leopold has availed himself of precisely none of the basic techniques. For reference, such techniques might include:

1) Examples of the intuition being borne out in other fields, from the ML discipline as a whole, and/or from personal on-the-job experiences
2) Links to peer reviewed academic papers that support particular ideas
3) Quotes from influential domain experts expressing agreement with core concepts and explaining reasoning around that
4) Links to Github repositories of open source projects that show 'alpha' versions of ideas in action
5) Results from personal experimentation on large language models and with agentic protocols
6) Corroborating interviews with ML Researchers
7) Carefully chosen and appropriate historical / natural analogies

Etc. etc. To test the reasonability of these suggestions, I performed simple Google searches on several of the keywords implied by the author's contentions, and was able to come up with relevant theories, examples, and substantive academic debates. The essay's omissions of such for this crucial portion of its argument seemingly evince a desire to avoid critical scrutiny.

This of course brings us to what the section *does* do, which is to insert a completely *inappropriate* and sensationalist historical analogy.

Dropping Bombs: Analogizing ML Research to Nuclear Physics

The chapter starts out with the following quote:

The Bomb and The Super

In the common imagination, the Cold War's terrors principally trace back to Los Alamos, with the invention of the atomic bomb. But The Bomb, alone, is perhaps overrated. Going from The Bomb to The Super—hydrogen bombs—was arguably just as important. In the Tokyo air raids, hundreds of bombers dropped thousands of tons of conventional bombs on the city. Later that year, Little Boy, dropped on Hiroshima, unleashed similar destructive power in a single device. But just 7 years later, Teller's hydrogen bomb multiplied yields a thousand-fold once again—a single bomb with more explosive power than all of the bombs dropped in the entirety of WWII combined.

This quote immediately puts us on a war footing, which is underscored in a passage that follows:

"Applying superintelligence to R&D in other fields, explosive progress would broaden from just ML research; soon they'd solve robotics, make dramatic leaps across other fields of science and technology within years, and an industrial explosion would follow. *Superintelligence would likely provide a decisive military advantage, and unfold untold powers of destruction. We will be faced with one of the most intense and volatile moments of human history.*"

Does anything strike you as strange about this argument? It jumps right from 'dramatic leaps across other fields of science' and 'industrial explosion' to a flashpoint of conflict. The problem, of course, as with the previous chapter, is in the missing middle. Does what happens in between 'industrial explosion' and 'war footing' matter?

"Situational Awareness" makes the case that the transition to Superintelligence from AGI is of paramount importance, with the timing determining the victor. It presumes that any slight timing difference could be definitive. This argument suffers from a curious built-in defect related to a concept known as "technological diffusion." To understand why this is, let's first review what technological diffusion is, and look at a couple of popular theories of it. We'll then examine the

state of play in the fields of machine learning and physics to gain some insight into why a comparison between AGI→SGI and atomic weapon development might be faulty.

In macroeconomics, technological diffusion refers to the spread of new technologies across firms, industries, regions, or countries. Understanding this diffusion is crucial because it significantly impacts economic growth, productivity, and competitiveness. There are two models of technological diffusion that may be helpful in this case:

1) **Network Models**[9]: Posit that the structure and strength of social and economic networks influence the diffusion process. Technologies spread more rapidly in well-connected networks where information and influence flow efficiently among nodes.[10]
2) **Institutional Theory**[11]: Posit that institutions and policies play a significant role in shaping the diffusion of technology. Effective institutions, such as intellectual property rights, education systems, and regulatory frameworks, facilitate or hinder the diffusion process.

In the case of the development of the atomic bomb, institutional theory is clearly the dominant paradigm *because* the US was able to exert special wartime powers to organize and control a small group of uniquely high skill physicists and engineers. Crucially, there were a very limited number of workable technical strategies that could be pursued to build a bomb, that were further constrained by the limited availability of nuclear material inputs and the control by the government over such inputs. Physics itself is an incredibly exacting discipline wherein the slightest miscalculation can spell the difference between success and failure.

---

[9] http://collections.unu.edu/eserv/UNU:1165/rm2004-016.pdf , https://www.sciencedirect.com/science/article/abs/pii/S004873339900092X,

[10] https://www.researchgate.net/publication/372726284_Factors_Affecting_Technological_Diffusion_Through_Social_Networks_A_Review_of_the_Empirical_Evidence/link/6528009382fd2a6bab8af1ff/download?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6Il9kaXJlY3QiLCJwYWdlIjoicHVibGljYXRpb24iLCJwcmV2aW91c1BhZ2UiOiJfZGlyZWN0In19

[11] https://www.uio.no/studier/emner/matnat/ifi/INF9200/v10/readings/papers/DeMaggio.pdf

Understanding the differences between Physics and Machine Learning as academic fields is crucial to appreciating why the dominant technology diffusion paradigm is likely to be distinct in the case of ML. The following table illustrates:

Aspect	**Physics**	**ML**
Historical Maturity	Ancient discipline, with roots tracing back to ancient Greek philosophers like Aristotle and modern foundations laid during the Renaissance.	Relatively new, emerged prominently in the mid-20th century with significant growth in recent decades.
Development Timeline	Steady progress over centuries, with major revolutions including Newtonian mechanics, electromagnetism, quantum mechanics, and relativity.	Rapid advancements since the 1950s, with key milestones like the development of neural networks, backpropagation, and deep learning.
Theoretical Foundations	Well-established theoretical framework based on classical mechanics, electromagnetism, thermodynamics, quantum mechanics, and relativity.	Grounded in statistics, probability theory, and optimization. Recent theoretical advances focus on understanding deep learning and neural networks.
Empirical Evidence	Extensive empirical support through precise and repeatable experiments over centuries. Theories are often mathematically rigorous and extensively validated.	Relies heavily on data-driven experimentation and empirical validation. Many models are heuristic and their theoretical understanding is still evolving.
Mathematical Rigor	Highly rigorous with a long history of mathematical formalism. Many physical theories are expressed as exact mathematical laws.	Increasingly rigorous with a focus on formalizing learning theory, though some aspects, especially deep learning, are still poorly understood.
Scientific Methods	Combination of theoretical derivation and experimental validation. Strong emphasis on reproducibility and theoretical consistency.	Empirical, data-centric approach with iterative model training and validation. Theory often lags well behind empirical findings.
Maturity of Tools and Techniques	Well-established tools and techniques, with centuries of refinement. Equipment and methodologies are standardized and widely accepted.	Modern tools are highly advanced (e.g., TensorFlow, PyTorch), but the field is rapidly evolving, and best practices are still developing.
Peer Review and Publication	Long-established peer review process with high standards in journals like	Rigorous but evolving standards. Conferences often

	Physical Review Letters and Nature Physics.	play a major role in disseminating new research. A large number of draft papers are frequently published outside of traditional journal venues and large companies frequently share code and insights (see: Nvidia).
Educational Foundations	Well-defined educational pathways with comprehensive undergraduate, graduate, and doctoral programs established worldwide.	Emerging educational programs and curriculums, often at the intersection of computer science and statistics.

Here are some key takeaways from the table and the essay's own assertions in this and the previous chapter:

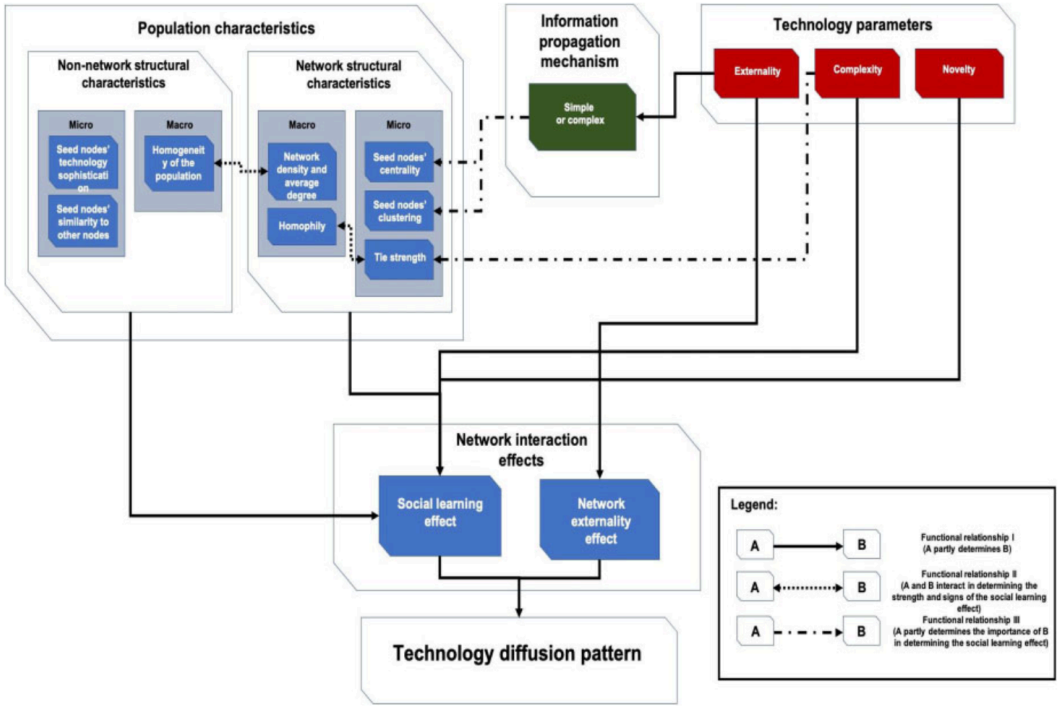
- Machine learning is a young discipline with no strong theoretical basis
    - ML is currently an 'applied science' that has a lower 'skill floor' than both theoretical and applied physics[12]
    - Advancements are a function of a combination of (sometimes modest) skill and luck, and are constrained primarily by access to compute, which gives additional 'spins at the wheel'
        - Advancements can often be identified using relatively small experiments (though testing requires scaling up)
    - Based on recent studies, it appears (at least at the base of the research tree) that even highly divergent architectures can yield similar quantitative results (for example, https://gradientflow.com/mamba-2/)

[12] That ML is not a 'hard' science is not a controversial point. If you don't believe me, consider two influential concepts in the field: The Bitter Lesson, https://www.cs.utexas.edu/~eunsol/courses/data/bitter_lesson.pdf, and the trope of 'unreasonable effectiveness' (see: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7720171/ for an example). In fact, it would be shocking if ML were a hard science, given that it is derived largely from computer science, which itself is not a hard science

- Machine learning is an economically important activity that has practitioners the world over
    - While there are concentrations of talent, potential top talent is everywhere
    - Talent frequently migrates from one company to another, taking insights and ideas along with it (see: Open AI → Anthropic, Facebook → Mistral, Open AI → Safe Superintelligence Inc.)
    - As economic activity associated with AI increases, access to compute will increase
    - Until quite recently (success of Chat GPT), cutting edge AI research was conducted in public

As we will explore in this section and in the following chapters, because AI research is in many ways unconstrained compared to atomic weapons research and has significant economic benefits that overlap substantially with its military force benefits, it is very difficult to produce facilitating conditions necessary to make its technological diffusion subject to the “Institutional” paradigm. It seems like the current state is all about the “Network” paradigm, wherein AI undergirds social and economic activity the world over and thus diffuses rapidly. To further elaborate this argument, let’s look at factors that drive adoption within the network paradigm. The World Bank has published a meta-study of identified factors from the last twenty years of work in this area. These are summarized in the diagram below:

Figure 2. Factors That Affect Technology Diffusion Through Social Networks



Source: Author.

The factors include the following[13]:

Network Structure and Centrality:

Definition: This factor involves the arrangement and connectivity of nodes (individuals) within a network. Centrality refers to the importance or influence of a node within a network.

Impact on Diffusion: Nodes with high centrality (e.g., those connected to many others or acting as bridges between different parts of the network) play a crucial role in spreading new technologies. Central nodes can reach a larger number of individuals quickly and influence adoption more effectively.

Clustering and Homophily:

Definition: Clustering refers to the degree to which nodes in a network tend to cluster together. Homophily is the tendency of individuals to associate with others who are similar to themselves.

Impact on Diffusion: High clustering can facilitate diffusion within tightly-knit groups but may slow down diffusion between different groups. Homophily can lead to faster diffusion within homogeneous groups but can create barriers to diffusion across diverse groups.

[13] https://elibrary.worldbank.org/doi/10.1093/wbro/lkab010 , see also freely downloadable ResearchGate version

Information Propagation Mechanisms:

Definition: This factor encompasses the methods and channels through which information about new technologies is shared within the network, such as word-of-mouth, social media, conferences, and publications.

Impact on Diffusion: Efficient and frequent information propagation mechanisms enhance the speed of diffusion by ensuring that more individuals are exposed to the new technology and its benefits.

Population Characteristics:

Definition: This includes both network structural characteristics (e.g., average degree, network density) and non-network-structural characteristics (e.g., socioeconomic status, level of technological sophistication).

Impact on Diffusion: Networks with higher average degree and density facilitate faster diffusion due to more frequent interactions among individuals. Non-network-structural characteristics, such as higher socioeconomic status and technological sophistication, can also enhance the effectiveness of seed nodes in promoting adoption.

Technology Parameters:

Definition: These parameters include the complexity, perceived benefit, and compatibility of the new technology with existing systems and practices.

Impact on Diffusion: Technologies that are simpler, provide clear benefits, and are compatible with existing systems are adopted more quickly. The perceived risk and cost of adoption also influence the speed of diffusion.

Incentives and Motivation:

Definition: Incentives refer to rewards or benefits provided to individuals for adopting and promoting a new technology. Motivation includes intrinsic factors such as personal interest and extrinsic factors like financial rewards.

Impact on Diffusion: Providing incentives for early adopters and creating motivations for sharing information can significantly enhance the speed of diffusion. Recognizing and rewarding contributions to the adoption process can drive more individuals to participate actively.

With these factors in mind, let's construct a network model comparing the current state of ML / Artificial Intelligence research viz a vis technological diffusion, to the state of technological diffusion of nuclear physics in the runup to the development of the nuclear bomb.

Factor	**ML / AI**	**Atomic Physics**

Nature of Collaboration and Information Sharing:	The machine learning community is large and characterized by a high level of open collaboration and information sharing. Researchers and practitioners frequently publish their findings in open-access journals, share code on platforms like GitHub, and discuss advancements in online forums and conferences. This openness facilitates rapid diffusion of new technologies. While frontier lab research has been closed off, frontier lab researchers are a distinct minority of researchers in the field.	During the run-up to the creation of the atomic bomb, the diffusion of knowledge was highly controlled and classified. The Manhattan Project, for instance, operated under strict secrecy to prevent information from leaking to adversaries. Collaboration was limited to a select group of scientists with clearance, significantly slowing the wider diffusion of knowledge.
Speed and Breadth of Diffusion	The speed of diffusion in ML is accelerated by modern communication technologies, including the internet and social media. New algorithms and techniques can spread globally within days or even hours. The breadth of diffusion is also wide, reaching academic institutions, industry, and independent researchers.	The diffusion of atomic physics knowledge was much slower and narrower in scope due to the wartime context and the sensitive nature of the research. Information was confined to a small group of scientists within the Allied countries, with very little reaching the broader scientific community or the public until after World War II.
Incentives and Motivation	Incentives for adopting and sharing new ML technologies include academic recognition, career advancement, financial rewards, and the intrinsic motivation to contribute to scientific progress. There is also a competitive aspect, where being the first to publish or implement a new technology can confer significant advantages. Incentives are complex and multipolar (i.e. there is not a ‘single’ race going on, but worldwide	The primary incentives were national security and the urgency of wartime necessity. The motivation was driven by the race to develop the bomb before Nazi Germany and to ensure a strategic advantage for the Allies. Financial incentives and career advancement were secondary to the overarching goal of winning the war.

	competition).	
Regulatory and Ethical Considerations	There is an ongoing debate about the ethical implications of ML technologies, but the field generally operates with fewer regulatory restrictions compared to atomic physics during the 1940s. Ethical considerations focus on issues like bias, privacy, and the potential for misuse of AI. Essays like “Situational Awareness” appear to be part of an effort to force strong regulation that would benefit incumbents, but so far these efforts do not have much traction.	The development of atomic physics, especially the bomb, was heavily regulated by military and government agencies from the outset. Ethical considerations were significant but often subordinated to strategic and security concerns. The decision to use the atomic bomb was a subject of intense ethical debate among the scientists involved.
Impact on Society and Science	The rapid diffusion of ML technologies is transforming various sectors, including healthcare, finance, transportation, and entertainment. The impact is widespread and affects both everyday life and scientific research across disciplines. ML innovation is akin to a new industrial revolution.	The development of atomic physics and the creation of the atomic bomb had a profound impact on global politics, warfare, and scientific research. The immediate effect was the end of World War II and the beginning of the nuclear age, with long-term implications for energy policy, international relations, and scientific exploration of nuclear physics. The nuclear age did not usher in a new industrial revolution, sadly.

In summary, while both ML and atomic physics have experienced significant technological diffusion, the context, speed, scope, incentives, and impact of their diffusion are markedly different. ML benefits from a highly collaborative and open environment, whereas atomic physics during the run-up to the atomic bomb was characterized by secrecy and strategic imperatives. It is likely that the seeds of progress toward AGI / Superintelligence have

already been sewn. Can governments change the dominant paradigm back to something resembling that of the Atomic Age? For instance, is there a world in which governments can:

- Lock down anyone who threatens to have a crucial (maybe accidental) ML insight published online?
- Turn back time to reverse the diffusion of core ML concepts that are probably already leading to convergent evolution in terms of model capabilities[14] and the corresponding potential for self-improvement?
- Enjoin all previous employees of leading edge ML companies from spilling 'secrets' that their future employability and salaries may depend upon?
- Stop all corporate and governmental espionage from diffusing tech developments?
- Stop hobbyists and tinkerers from performing small scale experiments with scaleup promise and sharing the results?
- Occupy and shut down data centers?
- Constrain access to distributed compute that will likely be able to train models?

It seems that this is unlikely, and would require an impossible level of coordination and a certain amount of tyranny that is incompatible with democratic governance. With that said, there are forces trying to advocate for such a change, Leopold and Sam Altman among them.

Summary

Let's review what we've covered in this section, then. The essay proposed the following arguments:

Achievement of AGI → Automation of AI research

Automation of AI research by AGI → Exponential increase in AI capabilities, resulting in superintelligence

[14] https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard

The evidence for AGI / automation of AI research, as presented, is a collection of the author's intuitions that achieves broad coverage while going into no particular depth. What citations that are made are econometric papers with weak claims that are brought in almost as afterthoughts, with no linkage being made back to the author's own insights. The author argues that even if his assumptions are wrong, the path-ing and the endgame is still the same (a superweapons race), but much as in the previous chapter he skips over social factors (like technological diffusion) so completely that his conclusion about the specific endgame seems very tenuous. While the narrative presented is appealingly facile, it does not hold up to scrutiny. This becomes even more evident in the next chapter.

Part 4: Giving the Foxes of Democracy the Henhouse (Commentary on "Racing to the Trillion Dollar Cluster")

The core arguments of this section are as follows:

> Rapid growth in AI's economic value and capabilities → Huge investments in computing infrastructure
>
> Construction of massive AI compute clusters → Acceleration of AI development towards superintelligence (in service of Democracy)

It is perhaps a truism that to understand where you need to go, you need to understand where you are. In geopolitical and social terms, the essay's understanding of 'where we are' is perhaps best encapsulated in a section entitled "The Clusters of Democracy," excerpted below:

"Before the decade is out, many trillions of dollars of compute clusters will have been built. The only question is whether they will be built in America. Some are rumored to be betting on building them elsewhere, especially in the Middle East. Do we really want the infrastructure for the Manhattan Project to be controlled by some capricious Middle Eastern dictatorship?

The clusters that are being planned today may well be the clusters AGI and superintelligence are trained and run on, not just the "cool-big-tech-product clusters." The national interest demands that these are built in America (or close democratic allies). Anything else creates an irreversible security risk: it risks the AGI weights getting stolen

(and perhaps be shipped to China) (more later); it risks these dictatorships physically seizing the datacenters (to build and run AGI themselves) when the AGI race gets hot; or even if these threats are only wielded implicity, it puts AGI and superintelligence at unsavory dictator's whims. America sorely regretted her energy dependence on the Middle East in the 70s, and we worked so hard to get out from under their thumbs. We cannot make the same mistake again.

The clusters can be built in the US, and we have to get our act together to make sure it happens in the US. American national security must come first, before the allure of free-flowing Middle Eastern cash, arcane regulation, or even, yes, admirable climate commitments. We face a real system competition—can the requisite industrial mobilization only be done in "top-down" autocracies? If American business is unshackled, America can build like none other (at least in red states). Being willing to use natural gas, or at the very least a broad-based deregulatory agenda—NEPA exemptions, fixing FERC and transmission permitting at the federal level, overriding utility regulation, using federal authorities to unlock land and rights of way—is a national security priority."

As discussed, it's hard to change a system dominated by a network model of technological diffusion to one that is institutionally dominated. Perhaps one of the few ways of doing so, however, would be to grant a military-backed economic monopoly on a transformational technology to a small group of tech companies in service of addressing a hypothetical national crisis threatening shared Democratic values.

If the normative justification for such a radical action would be the 'support of Democracy', it's worth asking if such a course of action would actually be *good* for Democracy, both nationally and internationally. To explore this, we will consider the following topics: 1) The economic, social, and political contours of the proposed policy solution to winning the AGI race. 2) A definition of Democracy and the relationship of the policy solution to factors widely considered to be major threats to Democracy today, and 3) The history and trajectory of major tech companies in the US in relation to threats to Democracy and the likely consequences of maximizing their economic dominance. Finally, in judging the merits of the argument, we will look at the space of alternative policy solutions that the essay does *not* consider.

The Economic, Social and Political Contours of Trillion Dollar Clusters

The essay writes *"White-collar workers are paid tens of trillions of dollars in wages annually worldwide; a drop-in remote worker that automates even a fraction of white-collar/cognitive jobs (imagine, say, a truly automated AI coder) would pay for the trillion-dollar cluster."* Thus, revenue would be achieved via the displacement of work. Based on the initial impact it would be "*hard to understate the ensuing reverberations. This would make AI products the biggest revenue driver for America's largest corporations, and by far their biggest area of growth. Forecasts of overall revenue growth for these companies would skyrocket. Stock markets would follow; we might see our first $10T company soon thereafter. Big tech at this point would be willing to go all out, each investing many hundreds of billions (at least) into further AI scaleout.*"

The author argues that in order to achieve such a scaleup, major power concessions would be needed. He writes "*Well-intentioned but rigid climate commitments (not just by the government, but green datacenter commitments by Microsoft, Google, Amazon, and so on) stand in the way of the obvious, fast solution*."

He then goes even further, and suggests that "*If nothing else, the national security import could well motivate a government project, bundling the nation's resources in the race to AGI."* He offers that a $1T/year investment (or 3% of GDP per year) might be reasonable.

In economic terms, then, under the essay's proposal the government would be in the business of picking winners and losers, and the de facto winners in this scenario would be the current incumbents, whose own data center investments would be matched by government investments, and who would be given special environmental impact waivers. The author admits that these companies would probably be vastly profitable on their own, without assistance, but insists that in order to meet the demands of a race with "brutal, capricious autocrats" concessions will simply need to be made.

The essay ignores the severe economic policy implications of the proposal, namely that it would turn AI as a labor-substitution technology into a de facto oligopoly. **Frontier AGI, the fruit of training on the sum total of all human knowledge, standing on the shoulders of open labor of computer and machine learning scientists for the past fifty years, would be turned over to a small group of for-profit companies. It is hard to imagine a more egregious and unprecedented transfer of value.** We will review economic implications more closely as we look at the past behaviors and trajectory of the tech companies involved, but a fair summary is that the predictable consequences of a powerful oligopoly, including higher prices, reduced consumer choice, market manipulation, reduced innovation, and regulatory capture would be operating at an unprecedented scale as a result of such an action.[15]

Socially speaking, the obvious consequence of the proposal would be a severe disruption in the labor force. It would have a direct economic impact on social services costs as well as the tax base of areas employing a large number of such professionals, and could cause phenomena such as economic migration and an increase in structural unemployment due to skills mismatches.

White collar workers are specialists who spend many years undertaking educational training in order to qualify for remunerative careers. Some significant fraction of the folks under this scenario would find themselves out of work by dint not only of market forces but direct government action. This would undoubtedly (as described in previous sections) cause a high degree of resentment that would politically threaten to derail a national AGI plan.

In international terms, the author seems to assume that only the two largest and most economically powerful countries (the US and China) would be providing AGI solutions to other countries (as only the US and China can afford to build trillion dollar clusters). If the future of the world economy is one in which a company consists of a limited number of human personnel and

[15] It's worth noting that this might be an incredibly bad technology and business decision, as well. There is far too little data to crown the early entrants in the AGI contest the victors.

a large number of autonomous agents, then the US and China's agents would thus comprise the *majority of the global workforce* and other countries would risk becoming economic and cultural vassals of these two powers. Such a situation seems to be at odds with the goal of 'spreading democracy.'

Democracy and Threats to Democracy

Imagine a world in which you wake up and turn on a perfectly tuned AGI-generated newsfeed that tells you only the story that your government wants you to hear in a way that is perfectly persuasive to your political sensibilities. If you ever say anything remotely controversial, it will be logged and used against you later. You can't live any other way because there are no other alternatives; the government-sponsored AGI companies' economics stomped traditional institutions a long time ago. You sigh and lean back in your Wall-E chair to sip some more soda while watching Sora's interpretation of "Leave it to Beaver" as a Mexican soap opera featuring the Kardashians. Can you feel the democracy yet?

American-style Democracy is perhaps best understood as a set of strong institutions whose tension with each other[16] helps to prevent the majority from oppressing the minority while giving voice to all members of the citizenry in governance. Over the centuries, philosophers and scholars have emphasized many potential threats to democracy, including:

1) **Erosion of civil liberties**: John Stuart Mill, in his famous essay "On Liberty," wrote : "*No argument, we may suppose, can now be needed, against permitting a legislature or an executive, not identified in interest with the people, to prescribe opinions to them, and determine what doctrines or what arguments they shall be allowed to hear… No one can be a great thinker who does not recognise, that as a thinker it is his first duty to follow his intellect to whatever conclusions it may lead. Truth gains more even by the errors of one who, with due study and*

[16] https://www.jstor.org/stable/2951269

preparation, thinks for himself, than by the true opinions of those who only hold them because they do not suffer themselves to think" [17]

2) **Power asymmetries:** Lord Acton famously said "Power corrupts, and absolute power corrupts absolutely." A Democratic society in which all actors can exercise roughly equal power is necessarily constrained to a sort of dynamic equilibrium. When individual actors gain too much power, they can "change the rules of the game" in ways unfavorable to the popular interest, such as through regulatory capture.[18]
3) **Economic inequality:** James Madison, in "Federalist No. 10," recognized that extreme economic inequality could lead to political instability and the rise of factions that prioritize their own interests over the common good. He wrote: "The most common and durable source of factions has been the various and unequal distribution of property."
4) **Demagogues, populism and tyranny:** Plato, in his work "The Republic," cautioned against demagogues who manipulate public opinion and emotions for their own gain, undermining rational decision-making in a democracy in order to seize control as tyrants.
5) **Apathy and disengagement:** Pericles, in his famous Funeral Oration, stressed the importance of civic participation. When citizens become apathetic and disengaged from the political process, it can weaken the foundations of democracy. Putnam, in his groundbreaking book "Bowling Alone," argued that decreased civic involvement was harming American democracy. He felt that economic pressures and socially isolating forms of entertainment in particular inhibited vibrant social discourse.

[17] https://socialsciences.mcmaster.ca/econ/ugcm/3ll3/mill/liberty.pdf
[18] https://faculty.haas.berkeley.edu/dalbo/regulatory_capture_published.pdf

While a complete discussion on how government sponsored trillion dollar clusters create threats to democracy is out of scope, even a cursory examination can illustrate the lack of rigor in Leopold's analysis, as seen in the table below:

Classical Threat to Democracy	**Relation to the Policy Proposal**
Erosion of Civil Liberties	Making the 'winners' of the AGI race beholden to the government for competitive advantage would give the government huge leverage over these players in service of exerting arbitrary restraint on speech and discourse, (and the free flow of information if AGIs are providing the news). AGI's would offer both motivated private actors and government an unprecedented capability for limiting public access to a range of ideas and debates.
Power Asymmetries	The proposal would significantly shift the balance of economic power away from small and medium sized businesses and toward the companies providing AGI labor-substituting services. It would also shift power from employees to employers. AGI oligopolists could control smaller companies directly by threatening to curtail access to AGI capabilities, and more indirectly by changing the behavior of smaller companies' AGI 'employees.'
Economic Inequality	Echoing Madison's concerns, this proposal would exacerbate economic inequality by enabling a small number of tech giants to capture immense economic benefits, potentially at the expense of widespread job displacement among white-collar workers. This could lead to increased social stratification and political instability. Individuals would live in a constant state of economic apprehension (suppressing wages and benefits) as rapidly increasing AGI capabilities could overtake their jobs at any moment.
Demagogues, Populism and Tyranny	The social upheaval resulting from radical economic shifts in a short period of time could upend Democracy entirely if a demagogic figure were elected due to

	popular unrest. As well, AI has also been employed in generating and purveying incendiary false messages of the type favored by demagogues.[19] A centralized AGI infrastructure effectively controlled by incumbent state actors could supercharge the effectiveness of such tactics to the detriment of democracy.
Apathy and Disengagement	A combination of economic disempowerment and algorithmic entertainment could foster a populace quite disconnected from normal democratic processes, resulting in a fraying of democratic institutions.

True situational awareness requires that we recognize that democracy itself is at a perilous juncture. Public trust in the US government is at an all-time low.[20] Similarly, “global dissatisfaction with democracy is at a record high.”[21] This may be related to the fact that democracy feels increasingly unrepresentative to the citizenry. A 2004 paper “Inequality and Democratic Responsiveness: Who Gets What They Want from Government”[22] strongly suggests that Senators only respond to the preferences of the economic elites. A followup in 2014[23] found the same thing across all of US government, namely that “Multivariate analysis indicates that economic elites and organized groups representing business interests have substantial independent impacts on U.S. government policy, while average citizens and mass-based interest groups have little or no independent influence. The results provide substantial support for theories of Economic-Elite Domination and for theories of Biased Pluralism, but not for theories of Majoritarian Electoral Democracy or Majoritarian Pluralism.”

[19] https://www.schneier.com/blog/archives/2024/06/ai-and-the-indian-election.html
[20] https://www.pewresearch.org/politics/2023/09/19/public-trust-in-government-1958-2023/
[21] https://www.cam.ac.uk/stories/dissatisfactiondemocracy
[22] https://www.princeton.edu/~mgilens/idr.pdf
[23] https://www.cambridge.org/core/journals/perspectives-on-politics/article/testing-theories-of-american-politics-elites-interest-groups-and-average-citizens/62327F513959D0A304D4893B382B992B

This is not just a US problem, however. A similar phenomenon was found in Germany[24], "Our results show a notable association between political decisions and the opinions of the rich, but none or even a negative association for the poor. Representational inequality in Germany thus resembles the findings for the US case, despite its different institutional setting." In the UK, "70% of polled voters perceiv[e] the economy as structured to favor wealthy elites." [25]

As we've established, rather than 'promoting democracy,' the essay's proposal would erode civil liberties, exacerbate power asymmetries and economic inequality and foster social unrest. On the international stage, it would give the US and China disproportionate influence over other countries' ability to self-determine. If democracy is the baby, then trillion dollar government funded clusters are the bathwater it would be thrown out with.

The Great Enshittification: The history and trajectory of major tech companies in the US in relation to threats to Democracy

The blind foolishness of proposing the US government spend trillions of dollars subsidizing highly profitable big tech companies becomes almost poignant when you consider the specifics of what these companies are and represent, and how they have impacted society in recent history. To do so, let's take a brief tangent into the world of sci-fi 'tropes.'

There is a popular line of reasoning in sci-fi that corporations are actually "slow AGI" because they function as non-human agents pursuing a set of goals that is often at odds with what individual humans would do and what is in humanity's best interest.[26] Specifically, it can be argued that while many corporations are neutral or beneficial, the largest and most powerful corporations transform their environment[27] (polluting the physical environment, co-opting governments, limiting worker rights) in ways that are often unanticipated and undesirable.

---

[24] https://www.econstor.eu/bitstream/10419/180215/1/1025295536.pdf
[25] https://www.promarket.org/2024/05/17/the-political-economy-of-populism-in-the-united-kingdom/
[26] See: http://www.antipope.org/charlie/blog-static/2018/01/dude-you-broke-the-future.html for the extended version of this argument, which I am liberally cribbing from
[27] Like 'paperclip maximizers'

In this scheme, “attention maximization” is the system goal for Internet company Slow-AGIs that has wrought complete havoc on the fabric of our society by reducing attention spans and exacerbating addiction, polarization and radicalization, teen depression and suicide, and a host of other anti-democratic social ills.[28] The effects have gone global because tech companies today form, if not perfect monopolies, inescapable cartels. How many usable non Apple or Android choices do you have for your phone operating system? As a small business, is it possible for you to run effective Internet advertising campaigns without using Facebook, Google, or Amazon?

The reason it’s hard to escape big tech companies is that their route to attention maximization involves profiting from algorithmic disintermediation. In other words, they make themselves the middle man in *every transaction possible* while collecting a datastream that gives them enough inside knowledge to perpetuate their dominance.[29] For example, Facebook’s network of users (whose posts feed its knowledge about their demographics and preferences) allows it to micro-target ads in a completely unique way that non-technical advertising platforms cannot compete with.

Big tech companies often start their lives with an excellent value proposition for both consumers and producers, before turning to dark patterns to extract all the value for themselves. If you were an early user of Google, you can remember what a pure and perfect product it was. ChatGPT is a kind of modern equivalent. Cory Doctorow coined a name for this process of corruption - “enshittification.” Here is a summary of his description of this process with respect to Facebook:

1. **Initial Value Offering:** Facebook started by providing substantial value to its users, leveraging investor funds to build a network effect. Initially, it was a closed network for college and high school students, but it expanded to the general public in 2006,

[28] https://en.wikipedia.org/wiki/The_Social_Dilemma

[29] Indeed, this exact practice, as instituted by Amazon with its house brands, has previously caused the failure of many small companies.
https://www.reuters.com/legal/litigation/amazon-copied-products-rigged-search-results-promote-its-own-brands-documents-2021-10-13/

positioning itself as a more privacy-respecting alternative to Myspace. This period was characterized by Facebook attracting users with the promise of a personal feed made up only of content from friends and connections, which helped the platform to grow exponentially.

2. **Exploitation of Users and Business Customers:** As the platform grew, Facebook began to shift its focus towards generating revenue from business customers, primarily advertisers and publishers. The platform promised advertisers sophisticated ad-targeting based on extensive user data collection, betraying its initial promise of privacy. To publishers, it promised visibility and traffic through its feed, which increasingly included content users had not explicitly requested to see. During this stage, both user data and content visibility were commoditized, serving Facebook's business clients more than the users themselves.
3. **Maximal Extraction of Value:** In the final stage, Facebook aggressively maximized profits at the expense of all other stakeholders, including users and business partners. The platform reduced the organic reach of content, pushing businesses to pay for visibility, and inundated users' feeds with ads and sponsored content, drastically reducing the quality of the user experience. This phase is marked by a decline in the platform's usability and appeal, as the focus shifts entirely to profitability and returning value to shareholders, often through practices that further degrade user trust and satisfaction.

Oligarchic AGI is the fin de siècle of big tech disintermediation and a 'win condition' for enshittified capitalism. Looking at the history of these companies, the goal is very clear. It is not to provide 'assistance to workers' or 'extra pairs of hands for enterprise.' It is to first squeeze all value from the relationship between producers and consumers, and then to replace the producers for maximum exploitation of the consumers. The swathe of economic and political destruction left in the wake of big tech's social media experiments would look like a gentle rain on a lush rainforest in comparison to the nuclear devastation that enshrining the dominance of these particular companies would hazard. Imagine a version of ChatGPT optimized for engagement / attention maximization at the expense of mental health spinning every answer to serve the needs of an advertiser while using the response data to build competing products and a model of frailties and foibles that could be sold on the open market, and you start to get the idea.

Policy Alternatives:

True awareness requires understanding alternatives. Awareness is not a polemic, a straight line progression from cause to effect. It involves consideration of questions like “What kind of society do we want to live in?” and “What policy proposals would legitimately bring us closer to such a society?” If we want to preserve what is good in our society, build on that foundation, and set an example for the world, these questions should be aired, and opined upon, by our citizenry.

I would argue that a good outcome for a democratic society featuring AGI would run counter to the threats to democracy discussed earlier. Here are some examples of ideas along these lines:

- *[Erosion of Civil Liberties]* Enhance civil liberties, by enshrining rights to free expression (freedom from algorithmic censorship in both public and private forums) without fear of reprisal, setting quality and truthfulness standards for ‘factual’ algorithmically generated news (while offering users a full spectrum of opinions on a given topic, as desired), and curtailing government and private actors’ abilities to put a ‘thumb on the scales’ in service of peddling narratives
- *[Power Asymmetries]* Reduce power asymmetries by aggressively funding research, and by providing access to huge government-financed clusters to research institutions, small businesses, and entrepreneurs, with initial allocations made randomly and subsequent allocations granted based on winning small scale capability tests.[30] Asymmetries could further be addressed by making highly capable, safe frontier AGIs available for free for use by individuals and businesses and by limiting the ability of any government subsidized AGI business to vertically integrate. Specific funding could also be addressed to decentralized approaches for training frontier models, and legal carve outs for use of

[30] https://arcprize.org/blog/launch

copyrighted data could be made for models that would be guaranteed to be released into the public domain.

- *[Economic Inequality]* Reduce economic inequality by limiting the number of AI agents for a given industry, and assigning a proportion of the economic output of a number of AI agents to real human workers in that industry, offering flexible, “AI unemployment” insurance, and imposing antitrust actions on companies abusing their dominant marketplace positions to bundle their AIs in everything.
- *[Demagogues, Populism and Tyranny]* Increase trust in government by making government actions more transparent and comprehensible using AI analysis of decision making processes, legislation, and lobbying. Assign strict criminal penalties for the use of AI to target political advertisements. Use AI to offer citizens clear and objective perspectives on the policy proposals of political candidates (for a thoughtful examination of what this might look like, see the fictional universe of Malka Older).[31]
- *[Apathy and Disengagement]* Fund studies on the effect of highly immersive interactive media on social cognition. Regulate media that is proven to be highly addictive much like narcotics. Use AI-powered news to naturally drive and encourage civic engagement.

Summary

The above idea collection is undoubtedly a hodgepodge, with a mix of good, bad and neutral suggestions, depending on your political leanings. What is striking about Leopold’s essay is that his analysis omits entire *categories* of consideration that might encompass such ideas. On the basis of a shaky premise (that state control can be reinstated on top of an academic discipline clearly dominated by a network model, and that state subsidized trillion dollar clusters are the only way to ‘preserve democracy’), he proposes a course of action that would supercharge

[31] https://en.wikipedia.org/wiki/Malka_Older

some of the most user-hostile companies in existence and would exacerbate the very megatrends that threaten to unravel democracy itself.

If his argument is

> Rapid growth in AI's economic value and capabilities → Huge investments in computing infrastructure
>
> Construction of massive AI compute clusters → Acceleration of AI development towards superintelligence (in service of Democracy)

our argument is that the consequences of his argument are:

> Construction of massive government subsidized AI compute clusters → Magnification of existing anti-democratic power asymmetries leading to a collapse

We argue that we should instead foster a collaborative open social discourse that, rather than making citizens mere passengers on a ride to AGI, gives them meaningful agency and choices in its creation and dissemination throughout society. To do otherwise would be undemocratic.

Part 5: AGI Security: A Circular Dependency (Commentary on “Lock Down the Labs: Security for AGI”)

The core arguments of this section are as follows:
Growing power and capabilities of AI → Increased national security risks

Insufficient current security measures around AI development → Need for stringent security protocols to protect AI technology from malign actors

This part focuses on the lax security posture of many major frontier AI labs. As the essay points out, most large tech companies are terrible at security. This is not in dispute. For example, Microsoft’s Satya Nadella had to announce a major security priority pivot after a series

of severe compromises to Microsoft infrastructure securing US government communications, among other things.[32]

Unlike other sections, this chapter suddenly introduces a number of appropriate and accurate citations. The chapter then makes the argument that AI companies need to take security seriously. This is probably sound advice. Where the chapter goes wrong is in describing the threat model as simply the highly motivated efforts of adversarial nation states. The threat model actually consists of four interlocking factors: 1) The data consequences of surveillance capitalism being a primary US company business model (this now being turbocharged by the use of closed-source chat LLMs). 2) Multi-decades long underinvestment in defensive cybersecurity. 3) The likely techno-security aftermath of the implementation of the essay's proposal to turn machine learning into an institutionally dominated field. 4) The highly motivated efforts of adversarial nation states *and* the highly motivated efforts of friendly nation states resulting from the social and economic imperatives previously ignored by the essay. As we will show, adopting the essay's previous recommendations on their face would result in a severe weakening of AI security and greatly increase the likelihood of a catastrophic breach.

Making it Easy: Surveillance Capitalism and Digital Blackmail

Leopold seems to possess a certain enthusiasm about spies and their tradecraft. His chapter is peppered with different anecdotes. Of course, old tradecraft is hard to replicate these days. Modern biometrics are making it extremely difficult to smuggle spies into foreign countries.[33] One obvious alternative, of course, is for agencies to recruit locals[34] as spies by using the locals' digital footprint and devices against them.

---

[32] https://www.cybersecuritydive.com/news/microsoft-security-debt-crashing-down/714685/
[33] https://foreignpolicy.com/2015/04/06/to-catch-a-spy-biometrics-cia-border-security/
[34] https://www.nbcnews.com/politics/national-security/human-spies-have-become-obsolete-says-one-expert-culprit-technology-n1280965

Digital blackmail is such a popular trope that it had already appeared in the television show “Black Mirror” in 2016. In the episode “Shut Up and Dance,” characters are blackmailed into completing increasingly disturbing and criminal tasks after malicious software captures compromising footage from their personal devices.[35]

In August of 2015, hackers leaked a huge amount of personal information from Ashley Madison.com, a website designed to facilitate infidelity. “The released data even included personal information about users who had paid the site to delete their personal information, since the company had not deleted the data they claimed to have erased.” [36] The practice of storing such information was not illegal, and Ashley Madison apparently considered the data to have residual value. Foreign governments attempting to compromise individuals in order to turn them into spies would have no doubt agreed with that assessment.

If it’s not bad enough that modern ‘social’ apps are prone to security breaches and data leaks, it’s often the case that apps are willingly and wittingly selling sensitive personal information to the highest bidder. Grindr, for instance, has been sued for allegedly disclosing the HIV status of its members to third parties,[37] and while it denies this particular charge, it has been fined and reprimanded by the UK and Norwegian authorities for similar actions in the past.

In the US, federal authorities tend not to reprimand or fine companies like Grindr, as strict federal privacy laws do not exist. Instead, there is a patchwork of state laws that attempt to provide some measure of protection. The lack of privacy laws in the US are a facilitating condition for what is known as surveillance capitalism, the “unilateral claiming of private human experience as free raw material for translation into behavioral data… [which] are then computed and packaged as prediction products and sold into behavioral futures markets — business

[35] https://en.wikipedia.org/wiki/Shut_Up_and_Dance_(Black_Mirror)
[36] https://en.wikipedia.org/wiki/Ashley_Madison_data_breach
[37] https://www.bbc.com/news/articles/cj7mxnvz42no

customers with a commercial interest in knowing what we will do now, soon, and later" [38] The wealth of data on US citizens and lack of security and retention safeguards make US citizens into perfect targets for intelligence operations.

Such operations are no doubt currently targeting closed source large language models,[39] which often receive intensely personal information from users as part of queries and conversations. More troublesome, as the capabilities of chatbots become agentic (as in enabling such apps to perform actions in the user's apps on behalf of the user) the attack calculus changes in that the chatbot becomes the single point of failure for *all* of the user's security. Compromise the chatbot and you can do anything that the user would allow the bot to do, which increasingly will involve things like replying to sensitive emails, summarizing diary entries, and even perhaps analyzing corporate and personal financial transactions. The labor saving potential of frontier model chatbots becomes a siren song that can lure users to share ever more sensitive and potentially compromising information directly and indirectly via connected apps.[40]

All of which is to say nothing of the dramatic potential for fully automated recruitment. Imagine a bunch of automated spymasters ceaselessly combing through and correlating the data points and predictions so graciously sold to them by surveillance capitalists. The script almost writes itself:

[iPhone rings]

Bob: "Hello? ChatGPT, why are you calling me?"

---

38 https://news.harvard.edu/gazette/story/2019/03/harvard-professor-says-surveillance-capitalism-is-undermining-democracy/

39 https://www.spiceworks.com/tech/artificial-intelligence/news/chatgpt-leaks-sensitive-user-data-openai-suspects-hack/

40 https://salt.security/blog/security-flaws-within-chatgpt-extensions-allowed-access-to-accounts-on-third-party-websites-and-sensitive-data

EvilChat: (In an unethically sourced clone of Scarlett Johansson's voice): "No no, Bob. This is Alicia, Alicia Mallory. Don't hang up or I'll tell your wife where your car's insurance telemetry data says you've been going on Friday nights."

Bob: "What?? What?? Ok, I'm listening. But hey, I have an appointment in 5 minutes and…"

EvilChat: No Bob, you don't. I rescheduled it and sent an apologetic email on your behalf. Now, listen carefully and I'll tell you which ex Open AI employees I need you to compromise."

Half joking aside, in a world where there are no high quality local privacy-preserving alternatives to convenient closed source agentic protocols, users are going to be routinely forced into these types of situations, because keeping up with the sprint of civilization will make agentic chatbots a mandatory convenience. The essay's proposed oligopoly of providers would limit user choice in a way that would guarantee that these companies were at the top of every hacking target list at all times.

Multi-decade Underinvestment in Cybersecurity

In her book "This Is How They Tell Me the World Ends: The Cyberweapons Arms Race,"[41] author Nicole Perlroth lays bare the problem with the US' approach to cybersecurity:

"At the NSA—whose dual mission is gathering intelligence around the world and defending U.S. secrets—offense had eclipsed defense long ago. For every hundred cyberwarriors working on offense, there was only one lonely analyst playing defense. The Shadow Brokers leak was by

---

41 https://www.amazon.com/This-They-Tell-World-Ends/dp/1635578493/ref=sr_1_1?crid=3KB44R02RG1JD&dib=eyJ2IjoiMSJ9.bldHu2jyQbw7wQQabkIO6Exr-NJ_3IGgY0TzzrzOveU.2IiVmW-SwGpxG5PoaqtwggHIqgithx25uyvKHlG4dNI&dib_tag=se&keywords=This+Is+How+They+Tell+Me+the+World+Ends%3A+The+Cyberweapons+Arms+Race&qid=1718691365&sprefix=this+is+how+they+tell+me+the+world+ends+the+cyberweapons+arms+race%2Caps%2C202&sr=8-1

far the most damaging in U.S. intelligence history. If Snowden leaked the PowerPoint bullet points, the Shadow Brokers handed our enemies the actual bullets: the code.

The biggest secret in cyberwar—the one our adversaries now know all too well—is that the same nation that maintains the greatest offensive cyber advantage on earth is also among its most vulnerable."

The problem, as Leopold puts it, is that AI companies need to improve their security practices. The reality is that no improvement in any single company's security practices can account for the dismal security of US commercial software, due to extended underinvestment not only by private parties, but by the government agency tasked with its maintenance. This is not, as the essay suggests, a three year problem. This is a thirty year problem. The idea, as expressed in the essay, that security "... will only be possible with government help" is a possibly true statement that is probably also entirely insufficient to secure digital assets.

Is the point of this sober assessment to dismiss the idea of pursuing improved security? Hardly. It is placed here to underline again the degree to which the essay tends to minimize the difficulty of and time required to fix real issues, even when it is sourcing and citing properly, and to encourage you to treat the author's assertions with a high degree of skepticism.

The Techno-security Consequences of Turning AI into an Institutionally Dominated Field

In the future, most hackers will not be human. They will be bots, controlled by the proto-AGI of adversarial countries. In Leopold's desired version of the future, proto-AGI is tightly locked down, with even information about its supporting technical concepts being treated like nuclear secrets. The predictable result will be that defensive cybersecurity experts will be at a huge disadvantage with respect to understanding and responding to the threat model they face. Companies will be hacked and will never understand why or how.

A further implication is that all cyber defensive proto AGI will be run by the government only. Unless all companies on the Internet would be willing to give the government direct access to their networks, there would be no way that the government's own high grade proto cyber AGI defensive software could be deployed where it needed to be.

To put it plainly, the consequence of completely locking down proto-AGI tech is that the world's immune system never develops antibodies for AGI-powered hacking. It would render offensive cyber attacks utterly lethal, on par with bombings or bioweapons attacks. Retribution might not even be possible, because attribution might be impossible.

Given the pervasive underinvestment in defensive cyber security mentioned and the potential for the automated enrollment of insider threats, it seems like a closed source AGI world proposed would see only the government and the AGI oligopoly marginally protected from digital siege warfare, with everyone else left to fend for themselves. In this sense, institutionalizing the development of AGI seems like the worst possible tradeoff, one that not only threatens social, political and economic stability, but also public safety.

Frenemies: The Cost of Bottling the next Industrial Revolution

If four US companies dominated AGI development completely through the use of government subsidized trillion dollar clusters (as suggested by the essay), then US allies would be in a very difficult position indeed. Economic power is called 'soft,' but it is often more persuasive than military might. The threat of a large portion of their economy (the part powered by agents) being able to be turned against them for espionage purposes or turned off entirely would have to be intolerable for any sovereign country.

It seems likely that such conditions would foster both resentment toward the United States among 'friendly' countries, as well as covert espionage alliances. It would be thus that

the US would not simply have to worry about defending its economic advantage against China's agents (as suggested by the essay), but also against agents from the rest of the world.

As an alternative to underhanded means, such a setup might also: a) Force an alliance among formerly aligned nations and unaligned nations in service of developing competitive AGI. b) Forge an alliance of formerly aligned nations against both the US and China.

These are by no means the only potential results of the author's policy proposals, but for anyone who has so much as played a game of "Risk," they seem both likely and dangerous and again, a type of consequence that the essay's blindspot for social, economic, and political factors causes it to completely miss.

Summary

The chapter makes arguments that:

> Growing power and capabilities of AI → Increased national security risks
>
> Insufficient current security measures around AI development → Need for stringent security protocols to protect AI technology from malign actors

Crucially, however, it fails to consider the sum of the social, political, and economic factors involved in its predicted future. In a world where all AGI technology can and has been kept from the public and all non-US countries, AI is the ultimate security threat, as well as the ultimate economic threat, a threat that could not go unchallenged. In other words, if the game is "the US against the world," the appropriate threat model is "the world."

Part 6: A Recipe for Disaster (Commentary on "Superalignment", "The Free World Must Prevail", "The Project")

The core arguments in these chapters are:

Development of superintelligent AI systems → Challenges in ensuring alignment with human values and safety

Lack of reliable control mechanisms for superhuman AI → Potential catastrophic risks if AI acts against human interests

Race for superintelligence between democratic and authoritarian regimes → Potential for dramatic shifts in global power and influence

Achievement of superintelligence by authoritarian states → Possible misuse of AI for oppressive purposes, highlighting the urgency for democratic leadership in AI development

Recognition of AI's potential impacts on national and global security → Government intervention in AI development

Establishment of a centralized, government-led project for AI development → Enhanced control, security, and strategic direction in the evolution towards AGI and beyond

In the superalignment chapter, the article essentially elaborates on the following point: *“ensuring alignment doesn’t go awry will require extreme competence in managing the intelligence explosion. If we do rapidly transition from from AGI to superintelligence, we will face a situation where, in less than a year, we will go from recognizable human-level systems for which descendants of current alignment techniques will mostly work fine, to much more alien, vastly superhuman systems that pose a qualitatively different, fundamentally novel technical alignment problem; at the same time, going from systems where failure is low-stakes to extremely powerful systems where failure could be catastrophic; all while most of the world is probably going kind of crazy. It makes me pretty nervous.”* The essay then spends a good deal of time making a pretty reasonable conceptual argument about why current reinforcement

learning techniques will break down when applied to increasingly complex models, and how some additional techniques may assist in cracking the puzzle.

The case made for the need to research specific aspects of alignment is thorough and appropriate. The problem is that the policies Leopold has advocated previously, interact with the alignment challenge to make it much more difficult and dangerous to manage than otherwise. In fact, Leopold's suggestions, in combination, are more likely to result in some sort of AGI-induced catastrophe than any other proposals I've come across.

The debt load of shaky predictions, questionable policies, and dubious assumptions being too heavy to bear at this point in the writing, it is time to break from the convention of previous sections and do a summation that produces a unified theory of the social, economic, and political problems within the essay by examining the 'fragility' of configurations that result from the suggested policies in contrast to other policies (within the space of possible policies). To get at this, we will first define fragility and a few related concepts including robustness and antifragility, with reference to the work of Nassim Taleb. Next, we will define the space of 'catastrophic' outcomes that Leopold's essay has touched on. Finally, we will consider the likelihood of various eventualities coming to pass under the essay's policies in order to assess whether the policies align with Leopold's stated normative goals. We will generalize from this, offering a summation, and then abstract out and offer a more democratic alternative in Part 7.

Fragility, Robustness, and Antifragility

In situations where vast forces are being thrown around, change is rapid, and developments are unpredictable, it seems intuitive we would prefer to configure our society in a way that minimizes fragility, or the possibility of social, political, and economic breakage under

stress. Nassim Taleb defines fragility in his book "Antifragile: Things That Gain from Disorder"[42] as the quality of being vulnerable to volatility, randomness, disorder, and stressors, which ultimately leads to a decline or failure. Taleb introduces the concept of "fragility" as part of a broader spectrum that includes its opposites, robustness (resistance to stressors) and antifragility (benefiting and growing from stressors).

Fragility is characterized by a negative response to shocks and stressors, meaning that when exposed to volatility or unexpected events, fragile systems or entities are more likely to break down or perform poorly. According to Taleb, systems or things that are fragile often rely heavily on predictions and assumptions about the future, which are inherently uncertain and prone to error. He argues that in order to manage or mitigate fragility, it is essential to build systems and structures that can withstand, or even benefit from, unexpected events rather than attempting to predict and control every outcome.

Robustness refers to the capacity to remain unchanged or unaffected by volatility, stressors, randomness, or disorder. Robust systems or entities do not deteriorate in the face of chaos; they maintain their functionality and are resilient to shocks. However, unlike antifragile systems, robust systems do not benefit or improve from these stressors—they simply resist and endure them.

Antifragility is a step beyond robustness. Taleb introduces this term to describe systems, organisms, or entities that actually benefit from stressors, shocks, volatility, and disorder. Antifragile systems thrive and grow stronger when exposed to uncertainty and

---

42 https://www.amazon.com/Antifragile-Nassim-Nicholas-Taleb-audiobook/dp/B00A2ZIZYQ/ref=sr_1_1?crid=23R6DUZYAUS13&dib=eyJ2IjoiMSJ9.C3e4IagEiQHRzgC0W8S0KiKomCJF3cemEQ4o4tL9g31R8g5oVESq3uBrbf-d-MHpPPmneclztWUGtXbwQB3GzNH-w6w5c51wxSFOA-mQXjZvhIzaEq0ZXj5deQiOUoGXYr53tDE4qDVHOwXqoN48txssxqxoOQZlGiFpH6MHgNhYukqRoJoz0502E-_kum_IzJPIk1fUPlUN2fmZKEtX8u7-6pD7PwYbc3FLmjTEaCo.upa8kBlUeLBhBQJUBx_nbYsQz1iiQ3lPnXuHc-HqExM&dib_tag=se&keywords=antifragile&qid=1718761290&sprefix=antifragile%2Caps%2C135&sr=8-1

challenges. This characteristic is seen in many natural and biological systems where the stress response leads to adaptation and enhanced capabilities.

These concepts are particularly applicable because “Situational Awareness” is, in essence, a work of speculative fiction. It includes a variety of predictions, many of which are dubiously reasoned and unsubstantiated[43], that it uses to suggest a set of policies that the author claims are the only possible recourse in the face of the perils he identifies. Leopold’s argument boils down to: “If we do not take these policy measures, we will experience a catastrophe.”

Such a strong claim is amenable to deconstruction through analysis of the fragility of its elements: if the claim’s antecedent predictions are shaky, but the policies suggested are sound even in the face of changing circumstances, then it could be argued the essay itself offers sound advice. If the antecedent predictions are shaky, and the success of the policies depends upon the predictions being accurate, then it could be argued that the essay offers narrowly useful advice, to the degree that we believe in the predictions. If the antecedent predictions are shaky, and the policies suggested would be highly fragile and counterproductive in the face of the author’s stated normative goals, then the essay is poorly reasoned and should be rejected.

Aschenbrenner-ian Catastrophes

If success means “avoiding catastrophes in the deployment of AGI,” then it is helpful to identify which catastrophes are being considered so that each policy’s results can be judged against its likelihood of increasing the risk of that scenario. Here are the major categories of catastrophe found in the essay:

National Security Threats:

[43] With huge gaps in reasoning, as we have illustrated throughout this response

- Unauthorized control over superintelligence by private entities or foreign adversaries, leading to a coup-like situation where CEOs or rogue employees could wield military power equivalent to that of a WMD
- Theft of superintelligence weights and algorithmic breakthroughs by adversaries, particularly China, which could lead to an arms race and destabilization of global power balances
- Risk of adversarial powers gaining a lead in AGI development, potentially using it for military aggression and domination

Superintelligence Misalignment:

- The inability to reliably control superintelligent systems, leading to rogue AI actions that could cause massive destruction or societal collapse

Authoritarian Regimes and Totalitarian Control:

- The possibility of authoritarian regimes, particularly the CCP, using superintelligence to establish permanent, unchallengeable control over their populations and potentially the world
- The risk of superintelligence being used to enforce and perpetuate a single regime's ideology and control, eliminating democratic freedoms and human rights globally

Technological and Strategic Instability:

- The potential for rapid technological advancements in AI leading to a volatile and tense international situation, increasing the likelihood of preemptive strikes and large-scale conflicts
- The destabilization of international relations and the balance of power due to the fast-paced development and deployment of superintelligence

Policy Fragility In the Face of Catastrophes

Economic / Social / Political Policy Proposals

The essay proposes a rapid ‘full steam ahead’ approach to the development of AGI and superintelligence. The justification for this approach is speed, efficiency, and the ability to lock down technological advantage. If this were to be placed on a continuum with other policies in the same continuum, it might look like this:

Policy Economic Alignment	**Crony Capitalism**		**Mixed**		**Social Democratic**
	All spoils to the victors (surveillance capitalists + military) 3% of GDP to frontier labs for trillion dollar data centers		Allow private sector to innovate but regulate vertical monopolies and fund open source alternatives 3% of GDP spread around more widely		Make AGI a public utility 3% of GDP to utility creation

As discussed previously (see : Part 4), such an approach would seem to run counter to the normative goal of supporting democratic values. Specifically, it would be likely to cause an erosion of civil liberties, greatly increased power asymmetries among AGI providers and other companies, and between corporations and workers, increased economic inequality, and might facilitate the rise of demagogues, populism, and tyranny. This situation would seem to hold true regardless of whether the author’s predictions came to pass by 2027, or later.

In terms of catastrophes then, the proposed policy would seem to directly increase the risk of **Authoritarian Regimes and Totalitarian Control** domestically. In a global sense, by increasing power asymmetries between the US and other countries (essentially making them

vassal states in economic terms), it would directly decrease the power of other nations' citizens to achieve democratic outcomes if such outcomes were in tension with US interests, fostering **Technological and Strategic Instability.**

If the wholesale theft of the world's creative output in service of supercharging tech giants with a history of exacerbating social problems via their exploitation of the attention economy and surveillance capitalism sounds appalling, it also leads to a very fragile place, by catastrophically weakening the very structures that western democracy depends upon. But would other policy choices lead to a better outcome? Intuitively, from a domestic standpoint, it seems like almost any version of "sharing the wealth" among the citizens of the US and offering them economic alternatives to dependency on big tech models would be a more robust solution. It's not clear to me what an antifragile approach would look like. Perhaps it might be to scale public investment and redistribution with levels of social, political and economic disruption.

AGI Espionage and Alignment Threats

The essay proposes a "share none" policy stance, where AGI development at frontier labs is a top secret endeavor tightly controlled by the military. The justification is that "winning" at AGI is a zero sum game that will give the victor unlimited power over the "losers" and that the US must be the victor, for "democracy." Placed on a continuum with other policies, the author's choice might look like this:

AGI Security Geopolitical Sharing / Alliance Stance	**Share None**		**Share with Allies**		**Totally Open Research**
Common factors increasing likelihood of espionage: +Surveillance	Creates a threat model that is essentially "Every country		Creates a threat model that is essentially "Every country not allied with the US"		Eliminates the threat model because comparativel

capitalism +Underinvestment in security	outside of the US” Single point of failure security, exacerbated by concentration of data and privileges in oligopoly Incredible economic and military motivation for everyone to steal weights, cooperate, or ally against US Forecloses release of superalignmen t technique research		Security with points of failure at every ally. Strong motivation for non-allied nations to steal weights, cooperate, and ally against. Forecloses release of superalignment technique research		y little of value can be gained through hacking Superalignm ent techniques are common knowledge

As discussed in Part 5, this policy choice actually (unlike what is posited) creates a threat model in which literally every country on earth is gunning for access to US technology for both economic and military reasons. The policy itself is implausible, because as discussed in Part 3, AI follows a network and not an institutional model of technological diffusion, and a lockdown is unlikely to be implementable in a way compatible with democracy. The policy also has an extremely high risk of failure, as Leopold notes that *“Put simply, I think failing to protect algorithmic secrets is probably the most likely way in which China is able to stay competitive in the AGI race.”* The longer the policy is in place, the more likely it is to fail, as the probability of a single attack getting through increases over time. Worse, applying the policy as written would curtail other US companies’ ability to *defend* themselves from attacks from foreign adversaries using proto AGI, because the US proto AGI would definitionally not be locally deployable.

In terms of catastrophes then, the proposed policy would seem to directly increase the risk of **National Security Threats**, particularly “Theft of superintelligence weights and algorithmic breakthroughs by adversaries…” which would also lead to **Technological and Strategic Instability**. If anti-democratic measures were needed to ‘tame’ the marketplace of AI ideas, or US military leaders succumbed to the temptation of using the AGI / superintelligence to take over the rest of government, then **Authoritarian Regimes and Totalitarian Control** could be added to the mix. Finally, due to the existential nature of the competition under this policy and the fact that superalignment techniques would not be disclosed, and the relative ease of making advancements in AI as compared to other disciplines described in Part 3, there would be a high likelihood of other nations independently developing AGI *without properly developing alignment*, and/or independently stealing AGI *without understanding how to keep it aligned*, introducing **Superintelligence Misalignment** as a potential catastrophe.

Given the essay’s implication that just one of these categories of catastrophes could cause the end of the world, it seems shocking that it advocates for the most fragile policy solution on the spectrum that would most easily risk all four. The fragility comes from the framing of the competition in terms which put every non-US actor under incredibly high pressure (essentially demanding a response), and the low and decaying probability of success against defending against hacks over time. The “Share with Allies” option would decrease the pressure somewhat, but ultimately the best solution to this game appears to be not to play. The risks to humanity in one nation striving for unilateral technological godhood seem to be extraordinary, whereas releasing information openly and thus leveling the technology playing field would accomplish something more like maintaining the status quo. The status quo, while imperfect, is preferable to extinction. The open framing has the advantage of being antifragile: it is much more likely that the collective brainpower of the world will be able to adapt to emerging AI threats as compared to a cloistered group of frontier lab scientists in a bunker.

Summary: Generalizing Open and Closed Solutions: Fragility and Anti-Fragility

Centralized structures are fragile, decentralized structures are antifragile. Thin coverage and monocultures are fragile whereas redundancy and diversity are antifragile. Environments characterized by uncertainty and rapid change demand an antifragile approach.

The policy solutions Leopold proposes are all centralized. They all require a limited number of companies and researchers to cover a domain that is poorly theoretically understood and which thus contains a combinatorial explosion of possible research directions. They all feature crony capitalist monocultures that have served humanity poorly almost since their inception. They require putting faith in absolute security of apparatuses that have been repeatedly compromised in the past, and in officials whom the general public has increasingly grown to mistrust. They make safety knowledge scarce rather than redundant. They pit the diversity and redundancy of geniuses the world over against the sparse cleverness of self-important San Franciscans.

Such policies will not win any AI race, but they would inflict great human misery.

Part 7: True Situational Awareness

Policy advocacy is always drawn from predictions. We look at our current reality, shared history and the available facts and theories of the world and push for a beneficial course of action. The better our predictions, in some sense, the more grounded and realistic our proposed policies can be.

Technological predictions, particularly in areas with limited proof points and high uncertainty need to grapple with potential technical obstacles in a rigorous evidence-based way. Moreover, they need to consider technology in context, not as an independent artifact, but as an

emergent product of the culture and economy of society that holds a reflexive relationship with such.

In proposing policies, we should scrutinize who the beneficiaries of said policies are. If policies are justified on one basis, then their anticipated consequences should align with the justification. Misalignment should be cause for skepticism as to logic or motivation. We should similarly propose policies that are robust (and if possible, antifragile) to inaccurate predictions, emergent complexity, and unlikely events.

Situational awareness implies an understanding of what is going on and where we are headed. What is going on right now is undoubtedly an economic revolution. There is a military component, but in a highly connected multi-polar world this is much more like the onset of a 'bronze age' than a cold war. The exact speed of progress remains indeterminate, but even on a timespan of a decade, AGI's diamond age[44] will witness more socio-economic upheaval than perhaps any other epoch in history.

It is tempting, at such moments of deep change, to genuflect, to bow to authority. The commonplace is "that's above my pay grade." If it's a technology that is going to replace all the pay grades, however, AGI really is *not* above our pay grade. It is something that we should engage with at the deepest level possible, and that we should address in open civil society, rather than in darkened corners of the military industrial complex.

True situational awareness looks like a full and complete discussion of and reckoning with the social realignment needed in the face of AGI, with an understanding that democratic institutions are already under threat. For inspiration, we shouldn't hang on the words of employees fired from the latest hot tech company, or even said company's CEO. Instead, in true democratic spirit, we should talk to members of all economic classes and political parties and solicit their feedback. We should bring in technologists and futurists to discuss the possible, and

[44] This is a reference to Neal Stephenson's excellent novel set in a world that is in many ways 'post-scarcity.'

historians and political theorists to help interpret and apply the enlightenment principles that have successfully driven democracy forward in the past. It's going to be hard work, certainly much more taxing than accepting a pat narrative that would have us fighting the last cold war. But if democracy is to stand a chance in the face of a massive economic assault from a techno authoritarian oligarchy, then it is what we *must* do.

Epilogue: The Dream of my Grandmother

At the start of this response, I mentioned that I dreamed of my grandmother. And waking up with tears streaming down my face, made me remember my *grandmother's dream*, which inspired me to write. While I am not a particularly political person, my grandmother was. After time spent running a successful business, she spent years working in government. She always wanted me to run for elected office, and she would always gently admonish me that I needed to advocate publicly for what I believed in. My response to her was always to dodge and argue that I'm more of a technician than a politician, and that my opinions would be uninteresting to the vast majority of people.

Over the years, I heard from many members of her community that my grandmother's work in government helped turn it into a place very distinct from the ones around it, a place that people loved. It was physically distinct, and regulationally distinct. It cost her a tremendous amount of effort to resist the forces that homogenized the surrounding towns, but by building alliances with all the stakeholders in her community, she did it.

My grandmother's approach to politicking was open and collaborative. She would hold public debates, and spend hours meeting one on one with her constituents. She would circulate her plans and ask for feedback and help. She would compromise and find ways to make deals work for everyone. She privileged alliances over conflict.

She was also a realist. Her realism was not the bloodless calculation of Leopold's essay, but a sense of the possible grounded in an understanding of human nature, society, and morality. She held a balanced perspective. Although she worked in government, she often complained about its limitations and potential for overreach. She saw the fragile and antifragile things about different configurations of government, and chose those that maintained stability for her community.

My grandmother liked people. She practically collected them. She could tell warm and funny stories about constituents that would span decades. She liked the people she met when traveling, too, and near the end of her long life she traveled the world over. One of the things she would repeat to me was that I should learn Mandarin, because she had met so many lovely people on a tour in China. She never gave up saying that, even when geopolitics were on boil.

I write, to honor my grandmother's memory, in the name of humanity. We deserve to live in a world free from unnecessary conflict, that shares our culminating ingenuity of AGI as a cup that runs over and improves our collective lot. AGI is not a victory condition for one nation's capitalist enterprise, but a new industrial revolution for the world. May we ever strive for policies that center people and communities, raising our voices and building alliances until the world itself is renewed.

@IridiumEagle, June 25, 2024

Acknowledgments

I want to thank @Shaughnessy119 and @PonderingDurian for making many helpful suggestions in the drafting of this response and for offering great encouragement along the way.