Close the Gates: *How we can keep the future human by choosing not to develop superhuman general-purpose artificial intelligence*

Anthony Aguirre

September 14, 2024

Introduction

Something has begun in the past ten years that is unique in the history of our species. Its consequences will largely determine the future of humanity. Starting around 2015, researchers have succeeded in developing *narrow* artificial intelligence (AI) – systems that can win at Go, play computer games, recognize images and speech, and so on, better than any human.[1] The field has succeeded to the degree that if one can precisely specify a task, say by creating a calculable metric of success, a machine system can probably be trained to do that task, generally better than people can.

This is amazing success, and is yielding extremely useful systems and products that will empower humanity. But narrow artificial intelligence has never been the true goal of the field. Rather, the aim has been to create *general* purpose AI systems (GPAIs), particularly ones that are simultaneously as good or better than humans across nearly *all* tasks, just as AI is now superhuman at Go, chess, poker, drone racing, etc. This is sometimes called "artificial general intelligence" and we will here call it "superhuman general purpose AI" (SGPAI).[2] This is the stated goal of a number of efforts, including those of several major companies.[3]

Here, too, *these efforts are succeeding.* General purpose AI systems like GPT-4, Gemini, and Claude, based on massive computations and mountains of data, have reached parity with typical humans across a wide variety of tasks. Now AI engineers at some of the largest of our technology companies are racing to push these giant experiments in machine intelligence to the next levels, at which they can match, and then exceed, human experts. **We should not do so. Not now, perhaps not ever.**

Why? *Because soon after these machine intelligence systems compete with human intelligence, we are likely to progressively lose control of them, and possible even lose control* **to** *them.* As Alan Turing put it already 72 years ago, "once the machine thinking method had started, it would not take long to outstrip our feeble powers. There would be no question of the machines dying, and they would be able to converse with each other to sharpen their wits. At some stage therefore we should

[1] This chart shows a set of tasks; many similar curves could be added to this graph.

[2] This naming is used to emphasis that generality and capability are distinct. General-purpose AI is *here*, and likely to simply get more powerful; different adjectives like "human-competitive" and "superhuman" in this essay will indicate levels of capability we can expect to move through. We should not necessarily expect some new breakthrough or step-change to something fundamentally different and worth calling "AGI." This approach is similar to that taken in a recently proposed framework for AGI classification.

[3] Deepmind, OpenAI, Anthropic, and X.ai were all founded with the specific goal of developing artificial general intelligence, but Meta, Microsoft, and others are now pursuing substantially similar paths.

have to expect the machines to take control..."[4]

This essay is an extended argument for why we should not, in the next few years, irrevocably open this gate: we should not train neural networks better than nearly everyone at a wide range of intellectual tasks, let alone ones better than the very best human experts or even all of human civilization. Instead, we should set an indefinite hard limit on the total computation employed in training an individual complete neural network, a bound on how fast a such a neural network runs, and likely other limits.

Although *don't build machines smarter than us* probably sounds like a pretty sane idea to most people,[5] it is a rare position in the AI community and even in policy discussions of AI, where the inexorable advance of AI capability is often taken as a given. But there are many technologies humanity could have pursued but has either chosen not to develop, or chosen to cease advancing: examples include human cloning, human germ-line engineering, eugenics, advanced bioweapons, and others.[6] The argument that we should choose similarly with AI systems "outside the Gates," i.e., significantly past the capability of today's frontier systems, goes as follows:

1. We are at the threshold of creating expert-competitive and superhuman GPAI systems in a time that could be as short as a few years.

2. Such "outside the Gates" systems pose profound risks to humanity, including at minimum a massive disruption of social, political, and economic systems that takes place much faster than we can manage.

3. Among the capabilities of these systems would be self-improvement, leading almost inevitably very powerful and uncontrollable non-human intelligences.

4. AI has enormous potential benefits. However, humanity can reap nearly all of the benefits we really want from AI with systems inside the Gates, and we can do so with safer and more transparent architectures.

5. Many of the purported benefits of superhuman GPAI are also double-edged technologies with large risk. If there are benefits that can *only* be realized with superhuman systems, we can always choose to development deliberately, carefully, and as a species.

6. Systems inside the Gates will still be very disruptive and pose a large array of risks. But these risks are potentially manageable with good governance.

[4] David Leavitt. *The man who knew too much: Alan Turing and the invention of the computer (great discoveries)*. WW Norton & Company, 2006

[5] Indeed polls show that a strong majority of people are against doing so.

[6] See this piece for more historical discussion.

7. Finally, we not only should but *can* implement a "Gate closure": although the required effort and global coordination will be difficult, there are dynamics and technical solutions that make this much more viable than it might seem.

The following sections develop these seven points in detail.

At the threshold

We now know about how much computation is sufficient to create performance across a significant span of basic intellectual tasks, including reasoning and problem solving. It is about one hundred trillion trillion, or 10^{25}, floating-point operations (FLOP).[7] This is the estimated level of computation employed in the training of OpenAI's GPT-4 and comparable deep-learning neural networks.[8]

These neural networks are able to perform well across a wide range of text-based intelligence tests that include mathematics, common-sense reasoning, various scientific disciplines, and code-writing. This includes tests, such as the Winograd schema,[9] that were specifically designed to test for general and human-level intelligence. There are many things these neural networks *cannot* do. Currently most are disembodied – existing only on servers – and process at most text and still images.[10] And current systems cannot out-perform top human *experts* in the tasks at which they are expert.

One can debate whether these systems "think" or "understand" or "reason" in the senses humans do; probably they do not. But whatever we call it, they do it at least as *well* as many people, across tasks that range from writing poetry to proofreading code.[11] And with known techniques it takes about 10^{25} FLOP, along with an appropriate training dataset, to create from scratch.[12] Human-*competitive* general purpose AI is here. Using the scientific prefix "yotta" to represent 10^{24}, we can say that it takes about 10 yottaFLOP. This number is very important: it may signify the threshold between human-competitive and super-human capability. Thresholds can be very important: 10^{24} is also roughy the number of uranium atoms in one kg of Uranium, and there's a very big difference between being below and above this number.

We also know (very roughly) how much computation speed, in operations per second, is sufficient for such a system to match the speed of human text processing. It is about $10^{15} - 10^{16}$ FLOP per second.[13] Again using scientific prefixes, this is 1-10 petaFLOP/s.

What happens if we increase these numbers of 10 yottaFLOP and 10 petaFLOP/s, or make the algorithms more efficient? Well, the AI systems will become faster and/or "smarter," i.e. more competent.

[7] Note that AI hardware performance can vary by a factor of ten more depending upon the precision of the arithmetic and the architecture of the computer. Counting logic-gate operations (ANDS, ORS, AND NOTS) would be fundamental but these are not commonly available or benchmarked. For present purposes it is useful to standardize on 16-bit operations (FP16), though appropriate conversion factors should be established.

[8] A collection of estimates and hard data is available from Epoch AI and indicates about 2×10^{25} 16-bit FLOP for GPT-4; this roughly matches numbers that were leaked for GPT-4. Estimates for other mid-2025 models are all within a factor of a few of GPT-4.

[9] Hector Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*, 2012

[10] A previous version of this paper stated "None so far can by themselves carry out careful and reliable symbolic manipulation or formal reasoning through many steps." But since that version, GPT-O1-preview has demonstrated strong ability to do this.

[11] A recent study by METR shows that current systems can match human performance in a suite of well-defined online tasks taking educated and competent humans up to about 30 minutes, succeeding more often than humans for shorter tasks and less often for longer tasks.

[12] It can take considerably less starting with some pre-existing AI – either by further training that model, or by using that model to train the new one.

[13] The above leaked information quotes 560 TFLOP per token generated. Around 7 tokens/s is needed to keep up with human thought, so this gives $\approx 3 \times 10^{15}$. Numbers quoted by NVIDIA for inference in Llama 3.1 405B give comparable results. Note, though, that inference speed depends on a lot of factors including context length. It is also interesting that $10^{15} - 10^{16}$ FLOP/s matches well with (very rough) estimates of the computational capacity of the human brain, and also that 10^{16} FLOP/s times 30 years of training yields $\sim 10^{25}$ FLOP.

That faster hardware allows faster AI run-speed is quite clear. As for "smarter," greater amounts of training computation have reliably yielded increases in AI effectiveness in the training metric.[14] Somewhat surprisingly, simple training metrics (such as word prediction accuracy) have emergently translated into competence at both related and seemingly-unrelated tasks (i.e. all sorts of text processing). For clearly-defined metrics and a given AI architecture, increases in training computation can be reliably translated into improvements in those metrics.[15] For less crisply defined general capabilities (such as those discussed below), the translation is less clear and predictive, but it is near-certain that larger models with more training computation will have new and better capabilities, even if it is hard to predict what those will be. So far those advances have been quite significant when significantly greater computation is employed, and as a rule nearly all capabilities improve at least somewhat with increased computation. Simply extrapolating performance on various tests with computation, expected computation used for models with time, leads to expectation of very high (i.e. expert level) performance on a suite of current performance metrics in a few years.[16]

Moreover, many of the limitations of current GPAI models can be remedied by other known techniques in AI. AI systems already exist that can take and process sensory data like sound and imagery, that can generate new media, that can solve puzzles in embodied form in (simulated) environments, that can do formal mathematical reasoning, and that can plan and pursue goals.[17] No major obstacles have arisen in combining these capabilities (nor in combining different training modalities), and there is no reason to think human capability at them is any sort of barrier that cannot be breached. Researchers are also learning how to build software "scaffolding" around AI systems, and give them tooling,[18] to make them work significantly better, and to integrate them together. There are well-known AI researchers (see for exampleMarcus and Chollet) who argue strongly that simply further scaling the techniques that go into contemporary GPAIs won't lead to SGPAI. But those are not arguments against approaches that combine current techniques with others.[19] Researchers do *not* have any real idea how to build human qualities such as phenomenal consciousness, sentience, and felt emotion into AI systems. It is in principle possible that this could be key to fully unlocking human-level cognition,[20] but so far it has not been necessary to have any consideration of these in building AI systems that can perform an extensive variety of tasks at a very high level.

Thus while there is no clear consensus on exactly what is missing from today's GPAI models to get to something worth calling superhuman GPAI, it is quite possible that no *fundamental* obstacles

[14] Jordan Hoffmann et al. Training compute-optimal large language models, 2022; and OpenAI. Gpt-4 technical report, 2023

[15] Yangjun Ruan, Chris J Maddison, and Tatsunori Hashimoto. Observational scaling laws and the predictability of language model performance. *arXiv preprint arXiv:2405.10938*, 2024

[16] For a full exposition of this method see this paper. For a more recent, compelling, trend-based argument for a short timeline to very powerful general AI see Section one of this one.

[17] e.g., Deepmind's AlphaStar and other game-playing systems that require long-term planning and strategy in a game environment.

[18] An example of tool use would be allowing a language model to search the web, or use a calculator; a scaffold would be for example a program that iteratively calls on language models, feeding outputs from one into prompts for another.

[19] For example the neurosybolic approaches that Deepmind recently used to attain silver-medal level performance in the International Math Olympiad.

[20] Antonio R Damasio. *The feeling of what happens: Body and emotion in the making of consciousness*. Houghton Mifflin Harcourt, 1999

Table 1: *

Capability	Description of capability	Status/prognosis	P_{scl}	P_{knwn}	P_{new}
Reasoning	People can do accurate, multistep reasoning, following rules and checking accuracy.	Strong reasoning in recent scaffolded LLMs; formal reasoning systems easily integrable; strong progress in neurosymbolic systems.	10	85	5
Agency	People can take actions in order to pursue goals, based on planning/prediction.	Many ML systems are agentic; LLMs can be made agents via wrappers.	5	90	5
Planning	People exhibit long-term and hierarchical planning.	Improving with scale; can be strongly aided using scaffolding and better training techniques.	30	65	5
Truth-grounding	GPAIs confabulate ungrounded information to satisfy queries.	Improving with scale; calibration data available within model; can be checked/improved via scaffolding.	30	65	5
Multi-sense processing	People integrate and real-time process visual, audio, and other sensory streams.	Training in multiple modalities appears to "just work," and improve with scale. Realtime video processing is difficult but e.g. self-driving systems are rapidly improving.	30	60	10
Embodied intelligence	People understand and actively interact with their real-world environment.	Reinforcement learning works well in simulated and real-world (robotic) environments and can be integrated into multimodal transformers.	5	85	10
Flexible problem-solving	Humans can recognize new patterns and invent new solutions to complex problems; current ML models struggle.	Improves with scale but weakly; may be solvable with neurosymbolic or generalized "search" techniques.	20	70	10
Learning & memory	People have working, short-term, and long-term memory, all of which are dynamic and inter-related.	All models learn during training; GPAIs learn within context window and during fine-tuning; "continual learning" and other techniques exist but not yet integrated into large GPAIs.	5	75	20
Originality	Current ML models are creative in transforming and combining existing ideas/works, but people can build new frameworks and structures, sometimes tied to their identity.	Can be hard to discern from "creativity," which may scale into it; may emerge from creativity plus self-awareness.	50	30	20
Self-direction	People develop and pursue their own goals, with internally-generated motivation and drive.	Largely composed of agency plus originality; likely to emerge in complex agential systems with abstract goals.	40	40	20
Self-reference	People understand and reason about themselves as situated within an environment/context.	Improving with scale and could be augmented with training reward.	70	10	20
World model	People have and continually update a predictive world model.	Improving with scale; updating tied to learning; GPAIS weak in real-world prediction.	20	50	30
Self-awareness	People have knowledge of and can reason regarding their own thoughts and mental states.	Exists in some sense in GPAIs, which can arguably pass the classic "mirror test" for self-awareness. Can be improved with scaffolding; but unclear if this is enough.	20	50	30
Abstraction & recursion	People can map relation sets into more abstract ones for reasoning and manipulation, including recursive "meta" reasoning.	Weakly improving with scale; could emerge in neurosymbolic systems.	30	40	30
Sentience	People experience qualia; these can be positive, negative or neutral valence; it is "like something" to be a person.	Very difficult and philosophically fraught to determine whether a given system has this.	5	10	85

Key capabilities present in human cognition that have been discussed as significantly sub-human in current (transformer-based, unscaffolded) language or multimodal GPAI systems ("GPAIS" in this table.) These capabilities should not be considered independent, as increase in any one is likely linked with increase in multiple others. The third column gives very terse relevant considerations. The last three columns represent the author's personal prediction of whether these capabilities will match (high) human level (a) through just scaling currently-used GPAI techniques, or (b) require combination of current techniques with other known techniques, in known ways, or (c) require new or poorly-known techniques to be developed and made scalable and combinable with those in (a) and (b). Note that not all of these capabilities (in particular "sentience") are necessary to make an AI that could successfully do AI research and engineering, which would dramatically speed the improvement cycle.

remain. Table 1 lists a number of capabilities that have been put forward as presently weak or largely absent. For some of these, it is likely that simply building larger systems using the same techniques could bridge the gap; in others folding other known techniques may be sufficient. In a few (particularly "sentience") major new techniques may be necessary; but then sentience is probably not necessary to have an extremely capable AI.[21] It is worth emphasizing that these are not independent capabilities, and often strengthening one will significantly strengthen others; indeed *all* of these (save perhaps "sentience") are present at some level in current GPAI systems despite approximately none of them having been specifically designed into those systems.

[21] This is not to say sentience, or phenomenal consciousness, is unimportant – it's arguable the most important thing there is! But the flip side of the "hard problem" of consciousness is that it is difficult to point to a functional capability that *requires* it – if we could do so then we could use that capability to determine which systems have phenomenal consciousness and which don't.

This analysis suggests that simply scaling present-day systems and incorporating known techniques is likely to bring many of these capabilities to human or superhuman level. Given that these systems are already human-competitive at many tasks, this would correspond to AI systems that are *better* than typical humans across a wide variety of tasks, and very likely better than human *experts* across a some subset of them (*expert competitive* GPAI). It is very plausible that this would soon thereafter bring systems that are generally superhuman (SGPAI), even if there may be some niche or abstract qualities missing.

Exactly how far away are each of these? We don't know.[22] Computation used in the largest AI training run has increased by 100,000,000 times in the past ten years, bringing it to around 10 yottaFLOP. The capability difference between AIs trained with 1 versus 10 yottaFLOP appears to be significant. In the biological world, a 10x difference in neuron count is the difference between bears and humans; and a 10x training difference could be that between a 3-year-old and a 30-year-old. Meanwhile, scaffolding and tooling techniques are growing very rapidly better, and researchers are pushing hard to combine different known AI techniques into upcoming systems. And underlying it all, tens of billions of research dollars per year are being poured in, and a vast level of technical talent with it.[23]

[22] As of writing, the technology forecasting platform Metaculus predicts a timeframe to a "weak" form of general intelligence between 2025 and 2031 (50% confidence interval), a "strong" form between 2026 and 2038, and an 83% probability of AI-human intelligence parity by 2040.

[23] Per a report by Goldman Sachs, companies are slated to spend more than a trillion USD on AI infrastructure in the next few years. This greatly exceeds the Apollo project at around $300B in current dollars, and dwarfs the mere $20B spend on the Manhattan project.

Given how capable current 10-100 yottaFLOP systems are, if we scale them further, we can confidently predict that AI systems will initially be like those now, but much better. Past that, we will soon start entering *terra incognita* in terms of intelligent systems and the risks they may pose to humanity.

Systems outside the Gates are profoundly disruptive and risky to humanity

Today's AI generally feels like an empowering tool. It largely does what you ask, in service of your larger project or agenda; it does little autonomously, and there are few significant tasks for which it provides a complete solution. Although AI can generate images, write essays, and do math, it requires humans to request the images, create the essay topics, or pose the math problems – and then a human to review, judge, and improve the results, which are generally not up to the standard of top human output. This is good. This is what we want.

But going through the Gate, so that most and then all tasks now done by humans can be done by AI, two major things go wrong. First, insofar as AI remains a tool, it can replace people — who have responsibility, ethics, social ties, and other aspects of humanity — with AI tools that can be misused. Second, rather than acting as a tool, powerful AI can be an *agent* that replaces people in their work, decisions, judgement, and everything else – cutting humanity out of the loop. And of course both types of replacement[24] can happen at once, and at very large scale.

This leads to a host of risks of systems outside the Gates. Here are ten examples.

- They will significantly disrupt labor, leading to dramatically higher income inequality and potentially to large-scale underemployment or unemployment, on a timescale far too short for society to adjust.
- They can dramatically increase the ability of terrorists, bad actors, and rogue states to cause harm via biological, chemical, cyber, autonomous, or other weapons, without AI providing a counterbalancing ability to prevent such harm.
- They would likely lead to the concentration of vast economic, social, and political power – potentially more than that of nation states – into a small number of massive private interests unaccountable to the public.
- They could enable effective mass surveillance and manipulation systems usable by governments or private interests to control a populace and pursue objectives in conflict with the public interest.
- They could be, or create, advanced self-replicating intelligent software viruses and worms that could proliferate evolve, massively disrupting global information systems.

[24] Note that the *generality* of GPAI is a big part of the problem here. We are used to tools, but a tool that can do *everything* leaves little relevance for the tool-wielder.

- They could flood society's information gathering, processing, and communication systems with completely realistic yet false, or spammy, or overly-targeted, or manipulative media so thoroughly that it generally becomes almost impossible to tell what is physically real or not, human or not, factual or not, and trustworthy or not.
- By undermining human discourse, debate, and election systems, they could reduce the credibility of democratic institutions to the point where they are effectively (or explicitly) replaced by others, ending democracy in current democratic states.
- They could lead to rapid large-scale runaway hyper-capitalism, with effectively AI-run companies competing in a largely electronic financial, sales, and services space. All of the failure modes and negative externalities of current capitalist economies could be exacerbated and sped far beyond human control, governance, or regulatory capability.
- They could fuel/power an arms race between nations in AI-powered weaponry, command-and-control systems, cyberweapons, etc., leading to very rapid buildup of extremely destructive capabilities.
- They could effectively end human culture when nearly all cultural objects (text, music, visual art, film, etc.) consumed by most people are created, mediated, or curated by nonhuman minds.

These risks are not speculative. Many of them are being realized as we speak, via today's AI systems! But consider, *really* consider, what each would look like with AI that is far more effective than today's, and without its weaknesses.

That is, consider labor displacement when most workers simply can't provide any significant economic value beyond what AI can, in their field of expertise or experience. Consider mass surveillance if everyone is being individually surveilled by something faster and more clever than themselves. How does democracy look when the most convincing public voices aren't even human, and have no stake in the outcome? What becomes of warfare when generals have to defer to AI (or simply put it in charge), lest they grant a decisive advantage to the enemy? Any one of the above ten risks represents a catastrophe for civilization if fully realized.

What is our plan for managing them? As it stands there are two on the table.

The first is to build safeguards into the systems to prevent them from doing things they "shouldn't." That's being done now: com-

mercial AI systems will, for example, refuse to help build a bomb or write hate speech.

This plan is not up to the task for systems outside the Gate.[25] It may help decrease risk of AI providing manifestly dangerous assistance to bad actors. But it will do nothing to prevent labor disruption, concentration of power, runaway hyper-capitalism, or replacement of human culture: these are just results of using the systems in ways that are allowed and will make their providers money! And governments will surely obtain access to systems for military or surveillance use.

The second plan is even worse. That plan is to just openly release very powerful AI systems for anyone to use as they like,[26] and hope for the best.

Implicit in both plans is that someone else, e.g. governments, will help to solve the problems through soft or hard law, standards, regulations, norms, and other mechanisms we generally use to manage technologies. But putting aside that the companies developing the models fight tooth-and-nail against any substantial regulation or externally imposed limitations at all, it's quite hard to see what regulation would even really help. Would it prevent companies from wholesale replacing workers with AI? Would it forbid people from letting AI run their companies for them? Would it prevent governments from using very high-powered AI in surveillance and weaponry? The issues are fundamental, and feel more and more intractable as AI becomes as or more capable than the people trying to manage it.

Still, humans and society are very adaptable. While none of these risks to humanity[27] are something society seems remotely prepared to deal with on a timescale of years, it is possible that in time we could adapt to the proliferation of human-level AI capabilities. If only we had the time, and if only AI would stay at roughly human level.

These systems would be capable of self-improvement, leading almost inevitably to loss of control of, or to, very powerful non-human "super"intelligences.

This brings us to an *additional* risk that is global and pervasive, because it is not an accident, or side-effect, but rather the natural and almost inevitable destination of unrestricted AI development.

This is that we (as a species) are currently in a process, the endpoint of which is one or more *highly* super-human general-purpose AIs, often called "superintelligences."[28] Systems like this are almost

[25] Technical solutions in this field of AI "alignment" are unlikely to be up to the task either. In present systems they work at some level, but are shallow and can generally be circumvented without significant effort; and as discussed below we have no real idea how to do this for much more advanced systems.

[26] Such AI systems may come with some built-in safeguards. But for any model with anything like current architecture, if full access to its weights are available, safety measures can be stripped away via additional training or other techniques. So it is virtually guaranteed that for each system with guardrails there will also be a widely available system without them. Indeed Meta's Llama 3.1 405B model was openly released with safeguards. But *even before that* a "base" model, with no safeguards, was leaked.

[27] Also worth adding is that there is "moral" risk that we might create digital beings that can suffer. As we currently do not have a reliable theory of consciousness that would allow us to distinguish physical systems that can and cannot suffer, we cannot rule this out theoretically. Moreover, AI systems' reports of their sentience are likely unreliable with respect to their actual experience (or non-experience) of sentience.

[28] A bright line between SGPAI and "highly" SGPAI probably does not exist, but there is a threshold in acronym awkwardness that also should not be crossed.

certainly not controllable by human organizations and institutions. Therefore we are currently in a process, the endpoint of which is the high probability of global loss of control by humanity over AI. If we lose control, things might go well for humanity, or very badly. Nobody really knows, and it would not be up to us.

This is a strong statement, so let's look at this argument in more detail.

We are currently in a process, the endpoint of which is one or more highly superhuman general-purpose AI system, i.e. AI systems that are more capable at nearly all intellectual tasks than human experts or even entire human institutions – e.g. more capable at doing theoretical physics than the human community of theoretical physicists and their institutions.

The process we are in is the competitive development of more capable general-purpose AI systems, using abilities and resources generated by one generation of AI capability to build the next. That the largest technology companies in the world are racing to try to build ever-more-capable general-purpose AI systems is clear, and the explicit goal of at least several of them is either "artificial general intelligence" or "superintelligence." But what is important is not just that they are trying but that they are succeeding, and how. Success – both in demonstrated systems and in deployed products – brings additional investment, talent, and competition. That would be true of any technology being successfully developed, and is very important. AI development goes further, in that AI systems can help develop new and better ones. This is already happening, with extensive AI coding support for humans,[29] AI-enabled chip design,[30] AI improvement of algorithms relevant to AI training,[31] and AI systems training other AI systems.[32] Currently, the "improvement loop" from one generation of AI systems to the next is mostly human-driven.[33] But with each successive improvement, AI systems can take over more of the human tasks and do them better and/or faster.

Whether AI itself, or merely the human mental, fiscal, and corporate/institutional resources it unlocks turn out to be most relevant, the key is that there is a positive feedback cycle on general AI capability. No significant damping terms are present: resources are abundant, talent is limited but being produced, new institutions are being rewarded, and there is currently near-zero regulatory process. And there is a very strong financial and competitive pressure driving companies forward.

So our expectation should be a continual progression of capabilities until parity with human experts is achieved. The first expert-competitive GPAIs may or may not do every single thing as well as some very particular trained and talented people. But they would do

[29] OpenAI's Codex, Github's Copilot, GPT-4, and Deepmind's alphacode are all examples.

[30] Anthony Agnesina et al. Autodmp: Automated dreamplace-based macro placement. In *Proceedings of the 2023 International Symposium on Physical Design*, pages 149–157, 2023a

[31] Daniel J Mankowitz et al. Faster sorting algorithms discovered using deep reinforcement learning. *Nature*, 618(7964):257–263, 2023b

[32] See this article for one flavor of this work.

[33] There are already examples of AI more directly improving itself; some are compiled here.

many things far better, just as chess- or go-playing systems do now, and far faster, just as current language models compose quality text far faster than any human. And they would have foreseeably have abilities no humans do, such as direct and instant access to formal systems and simulations for planning and engineering, ability to read and write complex code almost instantly, a much greater potential working memory, and direct apprehension of far more knowledge and understanding. And that is just to start.

Once there is a system that can perform broadly at human expert level, it is very likely a short step to a *highly* superhuman general-purpose AI, a.k.a. superintelligence. At minimum, simply running many highly capable systems in parallel and at high speed, which are able to communicate and share understanding far more efficiently than humans,[34] would probably constitute a superintelligence. More generally, the same self-reinforcing feedback loop that led to expert-level systems would likely just continue from there. There are no established theoretical limits to the capability of a runaway superintelligence like this, other than some proposed on the basis of complexity of the world and those imposed on computation by physics, both of which exceed human limitations by orders of magnitude.[35]

A superhuman GPAI system is not controllable by human systems and institutions.

For AI systems to be under control, it must be constituted so as to *do what we want them to do.* That sounds straightforward but hides an incredible level of difficulty with advanced systems. To do what we want, AI systems will have to (a) be competent at accomplishing requested tasks; (b) "understand" what "we" want; (c) actually do it. Since we're discussing a very competent superhuman GPAI system they key parts are (b) and (c). The difficulty of these is what is often termed the "alignment problem."[36]

The systems in question are smart, so they won't be dumb about what we want.[37] Nonetheless it's a very hard problem. One key issue worth focusing on is: who are "we"? On one, perhaps simpler, hand, an AI could be *loyal* to a particular person or organization, reliably taking on that human system's goals and interests as its own. Alternatively, the AI system could be more *sovereign*, pursuing internal objectives while constrained by a set of norms, rules, or built-in ethics so that what it does generally accords with what "we" collectively want.

The glaring and really unfortunate problem with either "loyalty" or "sovereign" alignment is: as systems grow in power, we really don't know how to do it. This is more-or-less universally agreed amongst researchers studying the problem. Why not? Well, "understanding" what "we" want is itself very difficult. Even considering individual

[34] Hinton has pointed out that (as in "federated learning") AI systems can directly share updates to their neural network weights, potentially "brain dumping" a large amount of learning from one AI to another in a way impossible for biological brains; Turing's admonition that they would "be able to converse with each other to sharpen their wits" was more true than he probably could have guessed.

[35] The Landauer limit and related thermodynamic limits are significant for irreversible computations near room temperatures. If these assumptions are relaxed, the limits on the computation that can be done in a given region of space-time with a given amount of energy are extraordinarily generous.

[36] Brian Christian. *The alignment problem: Machine learning and human values*. WW Norton & Company, 2020; Stuart Russell. *Human compatible: Artificial intelligence and the problem of control*. Penguin, 2019; Eliezer Yudkowsky. The ai alignment problem: why it is hard, and where to start. *Symbolic Systems Distinguished Speaker*, 4, 2016; and Richard Ngo, Lawrence Chan, and Sören Mindermann. The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*, 2022

[37] For example it's clear that current AI systems already have "common sense" at the level that would preclude them from, say, mistakenly cooking the family cat for dinner.

humans we often don't understand what others want, or even what we want ourselves; it can be extremely subtle and even ill-defined. And if "we" are many humans with conflicting wants, what it is that "we" want?

As even harder problem is: how do we guarantee that an AI system will "care" about what we want? We can train AI systems to say and not say things by providing feedback; and they can learn and reason about what humans want and care about just like they reason about other things. But we have no method – even theoretically – to cause them to reliably "care" about what people care about. There are high-functioning human psychopaths who know what is considered right and wrong, and how they are supposed to behave. They simply don't *care*. But they can *act* as if they do, if it suits their purpose. Just as we don't know how to change a psychopath (or anyone else) into someone genuinely, completely loyal or aligned with someone or something else, we have *no idea* how to solve the alignment problem in systems advanced enough to model themselves as agents in the world, and potentially deceive. If this proves unachievable or impossible, then SGPAIs won't be under control, period.[38] Also of crucial importance: alignment or any other safety features only matter if they are actually used in an AI system. Systems that are openly released (i.e. where model weights and architecture are available) can be transformed relatively easily into systems of comparable power *without* those safety measures. Open-releasing a smarter-than-human AI systems would be astonishingly reckless, and it is hard to imagine how human control or even relevance would be maintained in this scenario.[39]

This bring us to the second reason to expect control loss. Suppose now that somehow alignment *succeeds*: via some breakthrough we are able to design powerful intelligences that reliably understand and do what "we" want, for different definitions of "we"; this would encompass both the "loyalty" flavor of obedience to a particular person or organization, and the more "sovereign" form of "doing things that are good for humans" and "not doing things that harm humans." Now consider the latter. That's *not* really human control, since it would almost certainly require disobeying human directives to do harmful things. That is, even if we had complete mastery over AI system design, there is the fundamental issue that *we can't have both total obedience and total benevolence because humans are not totally benevolent!* We see this already: language models will, by design, refuse to comply with certain requests, such as to create toxic or dangerous content. But now extrapolate: either a given superhuman GPAI will be absolutely obedient and loyal to some human command system, or it won't be. If not, *it will do things it believes to be good for us, but*

[38] This is the "rogue AI" scenario. In principle the risk could be relatively minor if the system can still be controlled by shutting it down; but the scenario could also include AI deception, self-exfiltration and reproduction, aggregation of power, and other steps that would make it difficult or impossible to do so.

[39] There would be every motivation, for example, to let loose powerful self-reproducing and self-sustaining AI agents with the goal to make money and send it to some cyptocurrency wallet. Or to win an election. Or overthrow a government. Could "good" AI help contain this? Perhaps – but only by delegating huge authority to it, leading to control loss as described below.

contrary to our wishes. That isn't something that is under control. A world full of powerful sovereign AI *might* end up being a good world for humans to be in; but as they grow ever more powerful, it won't be our world.

So let's imagine that we are able to produce a highly obedient AI, subject perhaps to some solid core ethics that most people would be comfortable with. Then we come to the third way in which we would lose control: a combination of *overdelegation* and *incommensurability.* A sufficiently advanced AI system could autonomously operate at many times human speed, sophistication, complexity, and data-processing capability, pursuing complex goals in complicated ways. Someone in charge of such a system may see it accomplish what they want; but would they understand even a small part of *how* it was accomplished? No, they would have to just accept it. What's more, much of what the system may do is *advise* its putative boss on what to do. That advice will be good! Over and over again.[40] At what point will the role of the human be reduced to clicking "yes, go ahead"? The power of delegation seems very likely to turn inexorably into the handover of all the important decisions. So that's "overdelegation"; what about "incommensurability"? Imagine you are CEO of a large company. There's no way you can track everything that's going on, but with the right setup of personnel, you can still meaningfully understand the big picture, and make decisions. But suppose just one thing: everyone else in the company operates at 100 times your speed. Can you still keep up? With strongly superhuman GPAI, you'd be "commanding" something that operates much faster, processing vastly more data than you can, and in ways you cannot comprehend. In what sense can there be meaningful *control* by humans of such a system? This can be put on a formal level. Ashby's law of requisite variety (and see the related "good regulator theorem"[41]) states, roughly, that a control system must have as many knobs and dials as the system being controlled has degrees of freedom.[42] A person controlling a strongly superintelligent AI system would be like a fern controlling General Motors: even if "do what the fern wants" were written into the corporate bylaws, the systems are so different in speed and range of action that "control" does not apply. (And how long until that pesky bylaw gets rewritten?)

Finally, let's come to a fourth reason to expect control loss. Suppose a superhuman GPAI system is, *somehow*, made both perfectly loyal to, and a perfect delegate for, some operator. That is, it understands perfectly well, and does, what its operators want and intend, while leaving them meaningfully in control. And it is powerful enough to be game-changing, while still commensurate enough to humans to be meaningfully controllable by them. While we're at it,

[40] This is especially acute in a competitive context, e.g. in a market economy running companies, or in a geopolitically adversarial situation. If AI is making good, fast decisions, there will be a powerful motivation to delegate more and moreto it.

[41] William Ross Ashby. An introduction to cybernetics. 1956; and Roger C Conant and W Ross Ashby. Every good regulator of a system must be a model of that system. *International journal of systems science*, 1(2):89–97, 1970

[42] In fact what he wrote was a bit more alarming in this context: "When the variety or complexity of the environment exceeds the capacity of a system (natural or artificial) the environment will dominate and ultimately destroy that system."

further suppose that there's just one, significantly more advanced than competitors, so that there isn't the competitive pressure to let its capabilities run away further, or to delegate too much. (Or perhaps its capabilities can be used to slow or stop competitors.) This is the dream situation for the AI developer and operator.[43]

But what do the other power structures in the world think of this? What does the Chinese government think when the US has (or appears to be obtaining) this? Or vice-versa? What does the US government (or Chinese, or Russian) think when OpenAI or Deepmind or Anthropic has (or appears to be obtaining) this? What does the US general public think if they believe Microsoft or Meta or a Chinese company has (or appears to be obtaining) this? What they will correctly think is that this is an existential threat to their existence as a power structure – and perhaps even their existence period (given that they may not be assured that the system *is* in fact under control.) All of these very powerful agents – including governments of fully equipped nations that surely have the means to do so – would be highly motivated to either obtain or destroy such a capability,[44] whether by force or subterfuge. The current world simply does not have any institutions that could be entrusted to house an AI of this capability without inviting immediate attack.[45]

And after all such attacks either:

1. The world is a smoking ruin, or
2. obedient SGPAI has proliferated to multiple mutually-hostile actors, or they are back in competition to get one, or
3. one group has, with the help of its SGPAI, taken over the others.[46]

The fundamental problem[47] is that there is no stable situation in which there is both competition between countries, companies, etc. *and* human control – the competition inevitably leads to delegating away the control. And eliminating the AI competition requires either one set of humans to seize power (which seems exceedingly unlikely to succeed rather than lead to war), or for the human groups to voluntarily create a long-term, stable, shared nexus of power to "hold" the SGPAI and prevent others from competing with it. No such nexus exists or appears to be on the horizon.

Therefore we are currently in a process, the endpoint of which is the high probability of global loss of control of AI by humanity. This includes the case where AI is "out of control" in the sense that AI has caused *everything* to be out of control.[48]

There is also a significant probability of loss of control *to* sovereign AI or to a single private interest controlling a loyal AI. We will note, although it is obvious enough, that loss of control by humanity to AI

[43] This appears to be the hope of the author of the "Situational Awareness" essay. This essay makes a compelling case for short timelines toward SGPAI, and the momentousness of its arrival. It makes a much less compelling case for why this AI would stay under control, or – as described below – how a destructive arms race can be avoided by leaning in to "winning" it.

[44] Such agents presumably would prefer "obtaining," with destruction a fallback; but securing models against both destruction *and* theft by powerful nations is difficult to say the least, especially for private entities. Note that the author by no means endorses either action – this is merely to point out that it is likely.

[45] In discussions of this topic it often appears to be implicitly assumed that achieving superhuman AI first would grants its operator the ability to prevent others from also developing them. But how, *exactly*, would that happen without leading to war?

[46] Of course, this possibility is what would more realistically lead to the smoking ruin outcome.

[47] Or, to put it succinctly, given sufficiently advanced AI, either it controls the world, or we somehow control it, or we all ruin the world through a power struggle. And the desirable middle option is open only if *both* we solve the (possibly unsolvable) technical control problem *and* the (presently unsolved) social coordination problem of agreeing who "we" are that do the controlling. So at present the only way to win is not to play.

[48] For example in a major war, or total ineffectiveness of current major institutions to actually make and implement significant decisions.

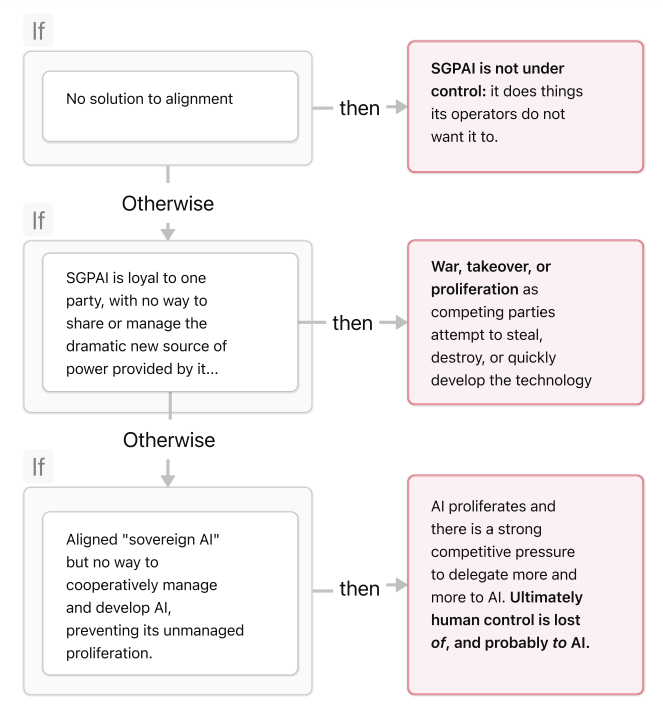

Figure 1: A simplified flowchart of how control loss could occur

also entails loss of control of the United States by the United States government; it means loss of control of China by the Chinese Communist party, and the loss of every other country by their own government. Thus AI companies are, even if this is not their intention, currently participating in the potential overthrow of world governments, including their own. This could happen in a matter of years. It should be evident on its face that: *if we lose control to AI, things might go well for humanity, or very badly.* Nobody really knows, and it won't be up to us. We will have replaced ourselves as the dominant species on Earth, that has agency over its own fate and that of others.

This, along with facing the pile of other large-scale risk listed above, is what happens if we don't stop somewhere. But why stop *here*? Why is the present level of advancement, at the yottaFLOP and petaFLOP/s level, the crucial one?

If humanity wants to retain control of its future, we don't really know how far we can let AI advance before the control loss becomes unstoppable. It may be that the next generation of AI is the real threshold, or the one after that. What we do know is that the currently-underway runaway process will only get harder to interrupt later, when the systems are more powerful and the race dynamics even more blatant. We also know that once a system is deployed, and especially once that level of system capability proliferates, it is exceedingly difficult to roll back. And if a system is *developed* (espe-

cially at great cost and effort), there will be enormous pressure to use or deploy it, and temptation for it to be leaked or stolen. This is to say that simply developing systems and *then* deciding whether they are deeply unsafe is a dangerous road.[49] Without a way of understanding and predicting the capabilities of new AI systems, each successive scaling of their capability is the opening of an un-closeable ever-larger Pandora's box of risk.

[49] There are also systems that would be *intrinsically* risky to develop, i.e. pose major risks even before deployment. These could include generalized hacking systems (trained to penetrate a wide variety of computer systems, or to escape confinement or boxing), virulent systems (designed to replicate themselves by utilizing difficult-to-obtain computation or memory resources, and to evolve progressively improved ability to do so), and recursively self-improving systems (that can undergo very large gains in capability without humans "in the loop," and/or in ways unanticipated by their designers and operators).

We can reap AI's benefits inside the Gates

Intelligence, whether biological or machine, can be broadly considered the ability to bring about futures more in line with some set of goals. As such, intelligence is of enormous benefit when used in pursuit of wisely chosen goals. Artificial intelligence is attracting huge investments of time and effort largely because of its promised benefits. So we should ask: to what degree would we still garner the benefits of AI if we contain its runaway progress? More precisely, we can ask: how much would "closing the Gates" by imposing limits on neural network systems (as proposed in detail in a section below) really curtail what we actually want to do with AI in the foreseeable future?[50] It may be surprisingly little, and with the added benefit of making our AI systems much more understandable and better integrated into human society.

[50] What *do* we actually want to do? The Sustainable Development Goals are an interesting place to start, representing the closest thing to a consensus on what the bulk of humanity is looking for from new technological and economic development.

Consider first that systems of GPT-4's generation are already very powerful, and we have really only scratched the surface of what can be done with them.[51] They are reasonably capable of "running the show" in terms of "understanding" a question or task presented to them, and what it would take to answer that question or do that task. Algorithmic improvements, new training regimes, advances in prompt crafting, and better dataset curation could almost certainly make considerably more capable systems using the same level of training computation.

[51] As economist and former Deepmind researcher Michael Webb put it, "I think if we stopped all development of bigger language models today, so GPT-4 and Claude and whatever, and they're the last things that we train of that size – so we're allowing lots more iteration on things of that size and all kinds of fine-tuning, but nothing bigger than that, no bigger advancements – just what we have today I think is enough to power 20 or 30 years of incredible economic growth."

Next, much of the excitement about modern AI systems is due to their generality; but some of the most capable AI systems – such as ones that generate or recognize speech or images, do scientific prediction and modeling, play games, etc. – are much narrower and are well "inside the Gates" in terms of computation.[52] These systems are superhuman at the particular tasks they do. They may have edge-case[53] (or exploitable[54]) weaknesses due to their narrowness; however *totally* narrow or *fully* general are not the only options available: given a computation budget, we'd likely see GPAI models pre-trained at (say) half that budget, and the other half used to train up very high capability in a more narrow range of tasks. This would give superhuman narrow capability backstopped by near-human general

[52] For example, Deepmind's alphafold system used only 100,000th of GPT-4's FLOP number.

[53] The difficulty of self-driving cars is important to note here: while nominally a narrow task, and achievable with fair reliability with relatively small AI systems, extensive real-world knowledge and understanding is necessary to get reliability to the level needed in such a safety-critical task.

[54] Tony T. Wang et al. Adversarial policies beat superhuman go ais, 2023c

intelligence.[55]

Now, these systems of varying levels of generality can be combined into composite systems. Current GPAIs are perfectly capable of using tools that are presented to them, and this could include *other* AI systems of varying generality.[56] We've also seen increasingly sophisticated scaffolds that auto-generate prompts, then process the output and generate new prompts.[57]

Crucially, the communication between these elements of a composite system can be human legible.[58] That is, rather than a powerful AI system being a giant inscrutable black box, it can instead be an amalgam of multiple components, each of which has an understandable function, and where humans (or AI systems) can audit how and why each is being called. Some of these components might be black boxes, and others could be clearly interpretable (e.g. if they are simply code.) But the idea would be that none of the black boxes would be *both* very general *and* superhumanly capable – only the composite could be both. Likewise, in terms of speed, many components (like most modern programs) could be extremely fast. But if a run-speed limit is implemented, the most powerful and general parts would not dramatically outstrip humans. This is crucial, because humans cannot stay meaningfully in-the-loop in a system operating dramatically faster than a human can.

Of course, such composite systems closely resemble what humans do in order to create more capability: we operate together in groups. This has significant downsides, but it is what has allowed us to create technology, civilization, and everything else of interest we've done. AI systems can and likely will operate in groups and collectives also; the question will be whether we will be meaningfully involved, or whether we will allow the elements of those systems to become powerful enough that this is not possible.

This idea of more interpretable and safer AI through modularity has been developed in some detail; see e.g. the "Comprehensive AI Services" model developed by Drexler,[59] the "Open Agency Architecture" of Dalrymple and the "Cognitive Emulation" model of Leahy & Alfours. All of these authors see the proposed architectures as both having much better properties in terms of safety and control than unitary agents (especially black-box ones), while supporting an aggregate AI system that is transformative in its capability. The primary "weakness" of these architectures is that they may be more work to create, and may be less powerful than a system created by "brute force" training of a single agent with huge amounts of computation across a very wide variety of tasks. That is precisely where a computation-based Gate closure helps, by eliminating these potentially unsafe or uncontrollable alternatives.

[55] This is just an extreme version of the "fine tuning" often applied to foundation models. Many other ways of combining general models with specialized training and tooling are also possible.

[56] For example, ChatGPT can write code for, and call, Mathematica, enabling it to do sophisticated computations and symbolic manipulation. And ChatGPT's code interpreter can write and execute code in other common languages.

[57] Yao Yao, Zuchao Li, and Hai Zhao. Beyond chain-of-thought, effective graph-of-thought reasoning in large language models, 2023; and Mehmet FIRAT and Saniye Kuleli. What if gpt4 became autonomous: The auto-gpt project and use cases. *Journal of Emerging Computer Technologies*, 3(1):1–6

[58] AI systems could communicate in other ways with different properties (e.g. more precision a la "code" or more efficiently as compressed data over even model weight updates) than human language. But these should be discouraged (especially the latter) if they come at the cost of much less intelligibility to us.

[59] Eric Drexler. Reframing superintelligence. *Future of Humanity Institute*, 2019

How could Gate closure of this type affect the companies building AI hardware and software? In terms of hardware, training and inference in yottaFLOP level systems would still require huge amounts of specialized hardware. On the software side, defusing the explosion in AI model and computation size should lead to companies redirecting resources toward making yottaFLOP-level systems better, more diverse, and specialized, rather than making bigger new ones. This may decrease the market advantage of the largest AI firms, which presently hold a monopoly on the ability to perform such huge computations, and help combat the current drive toward concentration of power into a few giant companies.

The upside to rushing to superhuman capability is modest relative to the downside risk: we can always choose to pursue it later and under better control.

Why are people and companies trying to build superhuman general-purpose AI? When asked, some responses would be that companies are "merely" building human-level AI. This is disingenuous: for many online tasks we already have "human level" AI, depending on the humans. Companies are seeking to build AI better than the most expert humans at the things those experts are best at; and this can hardly help but quickly lead to systems that go beyond human capability. Other responses often are to list, somewhat vaguely, problems that AI could help with: new medicines, new materials, new coordination mechanisms, and in general improving things for people. A more precise list of worthy goals is present in the UN Sustainable Development Goals. These are, in a sense, the closest we have to a set of global consensus goals for what we'd like to see improved in the world. And AI could help (see Vinuesa et al.[60]). But it is important to note that all of these goals – and indeed the applications often listed as what AI is being developed for – are those for which in-Gate AI probably suffices, or at the very least where we're very far from tapping out its potential.

[60] Ricardo Vinuesa et al. The role of artificial intelligence in achieving the sustainable development goals. *Nature communications*, 11(1):1–10, 2020

So let us be clear about what is actually motivating the quest for strongly superhuman AI. Whether or not these are actually achievable, these are things like:

1. Cures for many or all diseases;
2. Stopping and reversal of aging;
3. New sustainable energy sources like fusion;
4. Human upgrades, or designer organisms via genetic engineering;

5. Nanotechnology and molecular manufacturing;
6. Mind uploads;
7. Exotic physics or space technologies;
8. Superhuman advice and decision-support;
9. Superhuman planning and coordination.

As discussed above, many of these may be achievable with in-Gate systems (and humans), albeit likely more slowly.[61] Putting this aside, it is worth categorizing these motivations for SGPAI a bit, as follows.

The first three are largely "single-edge" technologies – i.e. likely to be quite strongly net positive. It's hard to argue against curing diseases or living much longer if one chooses. And we've already reaped the negative side of fusion (in the form of nuclear weapons); it would be lovely to also get the positive side. So the question here is whether getting these technologies sooner compensates for the risk.

The next four are pretty clearly double-edged: transformative technologies with potentially huge upside and immense risks, much like AI. All of these, if sprung out of a black-box tomorrow and actually deployed, would be incredibly difficult to manage well: we have none of the rules, norms, or institutions in place to do so.[62]

The final two concern the superhuman AI actually doing things rather than just inventing technology. More precisely, putting euphemisms aside, these involve powerful AI systems telling people what to do. Calling this "advice" is disingenuous if the system doing the advising is far more powerful than the advisee, who cannot really understand the basis of decision (or even if this is provided, trust that the advisor would not provide a similarly compelling rationale for a different decision.) Similarly with "coordination": coordination is a strong human capability, and we are actually excellent at it when we choose to be. There are undoubtedly systems we could (and should) devise or implement to coordinate better, more widely, and more intelligently. But we don't need superhuman AI for this; the way in which AI would be able to make it happen would either be through superhuman persuasion, politics, or force.

This points to a key member left off the above list:

10. Power.

It is abundantly clear that much of what is underlying the current race for superhuman AI is the idea that *intelligence = power*. Each racer is banking on being the best holder of that power, and that they will be able to wield it for ostensibly benevolent reasons without it leaving or being taken from their control.

[61] It is also worth noting that translating new ideas and inventions – even very good ones – into actual deployed technologies often takes far longer than hoped, so super-inventions by SGPAI may take quite some time to actually provide benefit unless we also let the AI take charge of many other processes.

[62] Thus we'd likely have to leave management of these technologies to the SGPAI, again exacerbating control loss.

Even in the extremely unlikely event that superhuman AI were a sort of controllable and power-granting genie, that these people and organizations want enormous power does not mean the rest of us should allow them take it and try to hold onto it.

Does this mean superhuman GPAI should never be developed? Let's suppose there is, in fact, some enormous upside to superhuman GPAI that cannot be obtained by humanity using in-Gate GPAI. In weighing the risks and rewards, there is an enormous asymmetric benefit in waiting versus rushing: we can wait until it can be done safely and beneficially (preferably provably[63] so), and almost everyone will still get to reap the rewards; if we rush, it could be – in the words of the CEO of OpenAI – lights out.

[63] Stuart Russell. *Human compatible: Artificial intelligence and the problem of control*. Penguin, 2019; and Max Tegmark and Steve Omohundro. Provably safe systems: the only path to controllable agi. *arXiv preprint arXiv:2309.01933*, 2023

The risks of systems inside the Gates are in principle manageable.

But it will not be easy. Current cutting-edge AI systems are quite powerful, and can significantly empower people and institutions in achieving their goals. This is, in general, a good thing! However, there are natural dynamics of having such systems at our disposal – suddenly and without much time for society to adapt – that offer serious risks that need to be managed. It is worth discussing a few major classes of such risks, and how they may be diminished, assuming a Gate closure.

One class of risks is of GPAIs allowing access to knowledge or capability that had previously been tied to a person or organization, making a combination of high capability plus high loyalty available to a very broad array of actors. Today, with enough money a person of ill intent could hire a team of chemists to design and produce new chemical weapons – but it isn't so very easy to have that money or to find/assemble the team and convince them to do something pretty clearly illegal, unethical, and dangerous. To prevent AI systems from playing such a role, improvements on current methods may well suffice,[64] as long as all those systems and access to them are responsibly managed. On the other hand, if powerful systems are released for general use and modification, any built-in safety measures are likely removable. So to avoid risks in this class, strong restrictions as to what can be publicly released – analogous to restrictions on details of nuclear, explosive, and other dangerous technologies – will be required.[65]

A second class of risks stems from the scaling up of machines that act like or impersonate people. At the level of harm to individual people, these risks include much more effective scams, spam, and phishing, and the proliferation of non-consensual deepfakes.[66] At a collective level, they include disruption of core social processes like

[64] The current dominant alignment technique is "reinforcement learning by human feedback" (RLHF) and uses human feedback to create a reward/punishment signal for reinforcement leaning of the AI model. This and related techniques like constitutional AI are working surprisingly well (though they lack robustness and can be circumvented with modest effort.) In addition, current language models are generally competent enough at common-sense reasoning that they will not make foolish moral mistakes. This is something of a sweet spot: smart enough to understand what people want (to the degree it can be defined), but not smart enough to plan elaborate deceptions or cause huge harm when they get it wrong.

[65] In the long run, any level of AI capability that gets developed is likely to proliferate, since ultimately it is software, and useful. We'll need to have robust mechanisms to defend against the risks such systems posed. But we *do not have that now* so we must be very measured in how much powerful AI models are allowed to proliferate.

[66] The vast majority of these are non-consensual pornographic deepfakes, including of minors.

public discussion and debate, our societal information and knowledge gathering, processing, and dissemination systems, and our political choice systems. Mitigating this risk is likely to involve (a) laws restricting the impersonation of people by AI systems, and holding liable AI developers that create systems that generate such impersonations, (b) watermarking and provenance systems that identify and classify (responsibly) generated AI content, and (c) new sociotechnical epistemic systems that can create a trusted chain from data (e.g. cameras and recordings) through facts, understanding, and good world-models.[67] All of this is possible, and AI can help with some parts of it.

A third general risk is that to the degree some tasks are automated, the humans presently doing those tasks can have less financial value as labor. Historically, automating tasks has made things enabled by those tasks cheaper and more abundant, while sorting the people previously doing those tasks into those still involved in the automated version (generally at higher skill/pay), and those whose labor is worth less or little. On net it is difficult to predict in which sectors more versus less human labor will be required in the resulting larger but more efficient sector. In parallel, the automation dynamic tends to increase inequality and general productivity, decrease the cost of certain goods and services (via efficiency increases), and increase the cost of others (via cost disease). For for those on the disfavored side of the inequality increase, it is deeply unclear whether the cost decrease in those certain goods and services outweighs the increase in others, and leads to overall greater well-being. So how will this go for AI? Because of the relative ease with which human intellectual labor can be replaced by GPAI, we can expect a rapid version of this with human-competitive GPAI.[68] If we close the Gate to SGPAI, at least some people will still represent valuable intellectual labor (and physical labor will be much more slowly automated), but huge labor displacement is still possible over a period of years. To avoid widespread economic suffering, it will likely be necessary to implement both some form of universal basic assets or income, and also engineer a cultural shift toward valuing and rewarding human-centric labor that is harder to automate (rather than seeing labor prices to drop due to the rise in available labor pushed out of other parts of the economy.) Other constructs, such as that of "data dignity" (in which the human producers of training data are auto-accorded royalties for the value created by that data in AI) may help. Automation by GPAIs also has a second potential adverse effect, which is of *inappropriate* automation. Along with applications where AI simply does a worse job, this would include those where AI systems are likely to violate moral, ethical, or legal precepts – for

[67] Many ingredients for such solutions exist, in the form of "bot-or-not" laws (in the EU AI act among other places), industry provenance-tracking technologies, innovative news aggregators, prediction aggregators and markets, etc.

[68] The automation wave may not follow previous patterns, in that relatively *high*-skill tasks such as quality writing, interpreting law, or giving medical advice, may be as much or even more vulnerable to automation than lower-skill tasks.

example in life and death decisions, and in judicial matters. These must be treated by applying and extending our current legal frameworks.

Finally, a significant threat of in-gate GPAI is its use in personalized persuasion, attention capture, and manipulation. We have seen in social media and other online platforms the growth of a deeply entrenched attention economy (where online services battle fiercely for user attention) and "surveillance capitalism"[69] systems (in which user information and profiling is added to the commodification of attention.) It is all but certain that more AI will be put into the service of both. AI is already heavily used in addictive feed algorithms, but this will evolve into addictive AI-generated content, customized to be compulsively consumed by a single person. And that person's input, responses, and data, will be fed into the attention/advertising machine to continue the vicious cycle. As well, as AI agents provided by tech companies become the interface for more online life, they will likely replace search engines and feeds as the mechanism by which persuasion and monetization of customers occurs. Our society's failure to control these dynamics so far does not bode well. Some of this dynamic may be lessened via regulations concerning privacy, data rights, and manipulation. Getting more to the problem's root may require different perspectives, such as that of AI loyalty:[70] requiring a standard of loyalty to the user (and not just the AI-providing company), or creating genuinely loyal AI assistants that can represent the fiduciary interests of their users and counteract the power imbalance when an individual user constantly interacts with a massive corporate/AI system.

[69] Shoshana Zuboff. The age of surveillance capitalism. In *Social Theory Re-Wired*, pages 203–213. Routledge, 2023

[70] Anthony Aguirre, Gaia Dempsey, Harry Surden, and Peter B Reiner. Ai loyalty: a new paradigm for aligning stakeholder interests. *IEEE Transactions on Technology and Society*, 1(3):128–137, 2020

The upshot of this discussion is that of hope: in-Gate systems – at least as long as they stay comparable in power and capability to today's most cutting-edge systems – are probably manageable if there is will and coordination to do so. Decent human institutions, potentially empowered by AI tools, can do it.

There are two ways this could fail. First, we could simply fail to create the governance mechanisms and institutions needed to manage in-Gate systems. But it's hard to see how allowing more powerful systems would help (other than by putting them in charge and hoping for the best). Second, it could turn out that in-Gate AI, when developed and networked together enough, is already powerful enough to create quite superhuman GPAI and many of its attendant risks. In this case, the resulting systems are likely at least to be much more manageable than monolithic giant neural network systems.[71] If humanity decides it does want to retain control of AI, and this requires additional limits, at least Gate closure would help enable this.

[71] This manageability is not automatic, and depends on deliberately designing the aggregate system in an interpretable and monitorable way.

We can *close the Gate now.*

What would it look like to choose not to develop superhuman general-purpose AI? At present we know of only one way to *make* such AI, which is via truly massive computations of deep neural networks. So while in the longer-term it may take additional measures, for now all we have to do is to *not* do those incredibly difficult and expensive computations. However, since companies are currently racing each other to perform them under heavy competitive and financial pressure from investors and otherwise, it will require regulation from the outside to place this limit.

Computation limits as Gate closure

To prevent the risks of superhuman GPAI while reaping the benefits of AI in general, there are various forms of limits one might imagine, depending for example upon the capabilities and/or risks of the systems; and risk-based limits will be required even for non-SGPAI systems. However, limiting SGPAI this way will be complex in terms of definitions, in line-drawing, and in implementation.[72] Computation limits are thus an extremely useful first and foundational step, for three key reasons.[73]

First, total training computation has been shown to be a good proxy for capability in GPAI systems. Most experts agree that the major progress in AI has been enabled by (and arguably largely resulted from) application of far more computation and data to techniques that have existed for decades. There is a reason AI companies are buying or leasing enormous numbers of high-end AI-specialized chips. Second, computation can be easily quantified, accounted, and audited, with relatively little ambiguity once good rules for doing so are developed. Third, large amounts of computation are, like enriched uranium, a very scarce, expensive and hard-to-produce resource. Although computer chips are ubiquitous, the hardware required for SGPAI is expensive and enormously difficult to manufacture.[74]

So a relatively straightforward but enormously impactful step would be a licensing system for AI development and deployment above some threshold of computation, which includes a global cap on the total amount of computation, in yottaFLOP, that goes into a neural network,[75] and a cap on the petaFLOP/s used in doing inference on a given neural network.[76]

A computation limit on neural networks will not indefinitely suffice to prevent SGPAI by itself: in principle SGPAI could be written entirely in code with no neural network. Likely sooner, as discussed

[72] Indeed the sort of governance we *should* have is one in which we can understand the general risk profile of AI systems before they are deployed and even before they are developed, and require that they satisfy a reasonable quantitative cost/benefit threshold before being granted a license for development or deployment. It might turn out that SGPAI systems may *never* satisfy such an analysis because their effects are too unpredictable or uncontrollable. Or we may develop currently-absent techniques by which it is possible, and the Gates could be safely opened in a controlled manner. For reviews of work along these lines see Dalrymple et al. 2024 and Tegmark & Omohundro 2023.

[73] Lennart Heim and Leonie Koessler. Training compute thresholds: Features and functions in ai regulation, 2024. URL `https://arxiv.org/abs/2405.10799`

[74] For example, the machines required to etch AI-relevant chips are made by only one firm, ASML (despite many other attempts to do so), the relevant chips themselves essentially all made by one firm, TSMC (despite others attempting to compete), and the design and construction of hardware from those chips done by just a few including NVIDIA, AMD, Google, etc.

[75] This would include all computation done in preparing the data or in creating any existing neural-network artifacts that are used as ingredients; see the proposal by Dalrymple for more detail on how this might work.

[76] At first, this could be done by simply limiting the number of GPUs that can communicate with each other at high bandwidth. For a more precise limit, this could be implemented as a required minimal time interval between each successive computation in an AI system if the total computation to produce the precursors to that next computation exceed a threshold.

above, enough carefully-crafted neural networks correctly connected together might create SGPAI; if poorly architected this could have all of the risks of a giant neural network. And as discussed below, as time goes on these limits would intrude more on other types of computation. But the precedent, processes, and systems put in place for a computation cap would be invaluable in creating a more durable limit should we choose to place one.

Implementing computation limits

It may seem that placing hard global limits on AI computation would require huge levels of international coordination and intrusive, privacy-shattering surveillance. Fortunately, it would not. The hardware powering advanced AI is very highly specialized and expensive, and flows though a very tight and bottle-necked supply chain.[77] So once a limit is set legally (whether by law or executive order), verification of compliance to that limit would only require involvement and cooperation of a handful of large companies. With a couple of notable exceptions (in particular NVIDIA) the AI-specialized hardware is a relatively small part of these companies' overall business and revenue model. Moreover, the gap between hardware used in advanced AI and "consumer grade" hardware is significant, so most consumers of computer hardware would be largely unaffected.[78]

[77]

[78] A current top-end consumer graphics card, the RTX4090, runs at 330 teraFLOP/s, a factor of thirty below an inference-speed cutoff at 10 petaFLOP/s; and a 10 yottaFLOP training cutoff would take over 1000 years to achieve on one. Non-AI industries most affected would probably be cryptocurrency mining and scientific high-performance computation, but both of these seem addressable.

Here is one *example* of how a gate closure could work, given a limit of 100 yottaFLOP (10^26 FLOP) for training and 10 petaFLOP/s for inference (running the AI). Many variations are possible, and this one targets simplicity and economy rather than other factors such as, e.g., political expediency.

1. Pause: for reasons of national security, the US Executive branch asks all companies based in the US, doing business in the US, or using chips manufactured in the US, to cease and desist indefinitely from any new AI training runs exceeding the N yottaFLOP limit. (Even if this does not have a clear enforcement mechanism, it can be expected that most if not all companies will comply, even if they complain.) The US should commence discussions with other countries hosting AI development, strongly encouraging them to take similar steps and indicating that the US pause may be lifted if they do not. Development of smaller models, and research, can continue unimpeded.

2. US oversight and licensing: By executive order or action of an existing regulatory agency, the US requires that within (say) one year:

   (a) All AI training runs estimated above 10 yottaFLOP done by

companies operating in the US be registered in a database maintained by a US regulatory agency; these registrations would include details of the training run, data and tools used, etc.[79]

(b) All AI-relevant hardware (GPU, TPU, and neuromorphic computing) manufacturers operating in the US or doing business with the USG adhere to a set of requirements on their specialized hardware and the software driving it.[80] Among these is a requirement that if the hardware is part of a high-speed-interconnected cluster capable of executing 10 petaFLOP/s of computation, a higher level of verification is required. This includes regular permission by a remote "custodian" who receives both telemetry and requests to perform additional computation, and grants a license to do this computation.[81]

(c) The custodian (which either is or includes the hardware manufacturer) reports the total computation performed on its hardware to the agency maintaining the US database; this can be compared to the registry to ensure that only registered large-scale computations are taking place.

(d) Stronger requirements are phased in to allow both more secure and more flexible oversight and permissioning,[82] as well as verification schemes to ensure that when a neural network is done training, the inputs and methods by which it was trained can be checked.

3. International oversight: at this point the US has requested but not strongly enforced any computation limits. But it has created the infrastructure to verify adherence to limits. This may then be extended internationally.

(a) The US, China, and any other countries hosting advanced chip manufacturing capability negotiate an international agreement.

(b) This agreement creates a new UN agency charged with overseeing AI training and execution just like the US federal agency in step 1.

(c) Signatory countries must require their domestic AI hardware manufacturers to comply with a set of requirements at least as strong as those imposed in the US, including registry of computations above the 10 yottaFLOP limit.

(d) Custodians are now required to report AI computation numbers to both agencies in their home countries as well as a new office within the UN agency.

(e) Additional countries are strongly encouraged to join the existing international agreement: export controls by signatory

[79] A slightly weaker version of this has in fact already been included in the 2023 US executive order on AI, requiring registration for models above 100 yottaFLOPs.

[80] Many of these requirements could be built into software and firmware updates to existing specialized hardware, but longterm and robust solutions would require changes to later generations of hardware.

[81] See this proof-of-concept project for how such system could be implemented. Modern hardware-based cryptographic security measures allow a wide range of capabilities that make governance of computational hardware far more interesting and flexible than other "dual use" materials such as uranium, while being much more difficult to circumvent than limits on software.
Gabriel Kulp, Daniel Gonzales, Everett Smith, Lennart Heim, Prateek Puri, Michael J. D. Vermeer, and Zev Winkelman. *Hardware-Enabled Governance Mechanisms: Developing Technical Solutions to Exempt Items Otherwise Classified Under Export Control Classification Numbers 3A090 and 4A090*. RAND Corporation, Santa Monica, CA, 2024. DOI: 10.7249/WRA3056-1

[82] For example, "multi-signature" permissions could be set up so that N of M custodians are sufficient to maintain computation, and thus $M - N + 1$ custodians in agreement could "pull the plug."

counties restrict access to high-end hardware by non-signatories (backed up by hardware mechanisms in AI chips that disallow their unsanctioned use.[83]) while signatories can receive technical support in managing their AI systems.

4. International verification and enforcement of agreed-upon Gate closure.

   (a) The hardware verification system is updated so that it reports computation usage to both the original custodian and also directly to the UN office. Similarly, license to run blocks of computation must be obtained from the agency (or both the agency and the custodian).

   (b) The agency, via discussion with the signatories of the international agreement, agrees on computation limitations. These then take legal force in the signatory countries. Verification of adherence can be checked by the telemetry provided to the agency and comparison with the registry. And if necessary, the limit can be enforced by refusing to grant permissions (i.e. licenses) to systems requesting more than the allowed amount of computation.

   (c) In parallel, a set of international standards may be developed so that training and running of AIs above a threshold of computation (but below the limit) are required to adhere to those standards. These would be minimal, safety-focused, standards, supplemented by regulation in particular jurisdictions such as the EU, China, and US.

   (d) The agency can, if necessary to compensate for better algorithms etc., lower the computation limit. Or, if it is deemed safe and advisable (at say the level of provable safety guarantees), raise the computation limit. In either case it can and should retain the ability to monitor, and if necessary shut down, advanced AI systems.

[83] James Petrie. Near-term enforcement of ai chip export controls using a minimal firmware-based design for offline licensing. *arXiv preprint arXiv:2404.18308*, 2024

Strengths and weaknesses of a computation-limit-based approach

A plan like this has a number of highly desirable features. It is minimally invasive in the sense that only a few major companies have requirements placed on them, and only fairly significant clusters of computation would be governed. The relevant chips already contain the hardware capabilities needed for a first version.[84] Both implementation and enforcement rely on standard legal restrictions. But these are backed up by terms of use of the hardware and by hardware controls, vastly simplifying enforcement and forestalling

[84] For more detailed analysis, see the recent reports from RAND and CNAS. These focus on technical feasibility, especially in the context of US export controls seeking to constrain other countries' capacity in high-end computation; but this has obvious overlap with the global constraint envisaged here.

cheating by companies, private groups, or even countries. There is ample precedent for hardware companies placing remote restrictions on their hardware usage, and locking/unlocking particular capabilities externally,[85] including even in high-powered CPUs in data centers.[86] Even for the rather small fraction of hardware and organizations affected, the oversight would be limited to telemetry, with no direct access to data or models themselves;[87] and the software for this could be open to inspection to exhibit that no additional data is being recorded. The schema is international and cooperative, and quite flexible and extensible. Because the limit chiefly is on hardware rather than software, it is relatively agnostic as to how AI software development and deployment occurs, and is compatible with variety of paradigms including more "decenteralized" or "public" AI aimed combating AI-driven combating concentration of power.

A computation-based Gate closure does have drawbacks as well. First, it is far from a full solution to the problem of AI governance in general. Second, as computer hardware gets faster, the system would "catch" more and more hardware in smaller and smaller clusters (or even individual GPUs).[88] It is also possible that due to algorithmic improvements an even lower computation limit would in time be necessary,[89] or that computation amount becomes largely irrelevant and closing the Gate would instead necessitate a more detailed risk-based or capability-base governance regime for AI. Third, no matter the guarantees and the small number of entities affected, such a system is bound to create push-back regarding privacy and surveillance, among other concerns.[90]

Of course, developing and implementing a governance scheme in a short time period will be quite challenging, and is outside the current window of discourse. But it absolutely is doable: it accords with what the public actually wants, and the required hardware technologies and software techniques exist already (though may need improving and enhanced security). What could supply the requisite *political will* would be widespread understanding that if decisive action is not taken soon, we could be approaching the end of the human era.

[85] Apple devices, for example, are remotely and securely locked when reported lost or stolen, and can be re-activated remotely. This relies on the same hardware security features discussed here.

[86] See e.g. IBM's capacity on demand offering, Intel's Intel on demand., and Apple's private cloud compute.

[87] Cryptography affords various methods of verifying properties such as model provenance that require no direct access to the model itself, which may even be encrypted.

[88] This study shows that historically the same performance have been achieved using about 30% less dollars per year. If this trend continues, in a decade or so a top-end consumer GPU could hit a 10 petaFLOP computation limit.

[89] Per the same study, given performance on image recognition has taken 2.5x less computation each year. If this were to also hold for highly-capable GPAI systems as well, a computation limit would not be a useful one for very long.

[90] In particular, at the country level this looks a lot like a nationalization of computation, in that the government would have a lot of control how computational power gets used. This, however, seems far safer than and preferable to SGPAI *itself* being nationalized via some merger between major AI companies and national governments, as some are starting to advocate for.

The choice before us

The last time humanity shared the Earth with other minds that spoke, thought, built technology, and did general-purpose problem solving was 40,000 years ago in ice-age Europe. Those other minds went extinct, probably wholly or in part due to the efforts of ours.

We are now re-entering such a time. The most advanced products of our culture and technology – datasets built from our entire

internet information commons, and 100-billion-element chips that are the most complex technologies we have ever crafted – are being combined to bring advanced general-purpose AI systems into being.

The developers of these systems are keen to portray them as tools for human empowerment. And indeed they could be. But make no mistake: our present trajectory is to build ever-more powerful, goal-directed, decision-making, and generally capable digital agents. They already perform as well as many humans at a broad range of intellectual tasks, are rapidly improving, and are contributing to their own improvement.

Unless this trajectory changes or hits an unexpected roadblock, we will soon – in years, not decades – have digital intelligences that are dangerously powerful. Even in the *best* of outcomes, these would bring great economic benefits (at least to some of us) but at the cost of a profound disruption in our society, and replacement of humans in most of the most important things we do: these machines would think for us, plan for us, decide for us, and create for us. We would be spoiled, but spoiled children. Much more likely, these systems would replace humans in both the positive *and* negative things we do, including exploitation, manipulation, violence, and war. Can we survive AI-hypercharged versions of these? Finally, it is more than plausible that things would not go well at all: that relatively soon we would be replaced not just in what we do, but in what we *are*, as architects of civilization and the future. Ask the neanderthals. Perhaps we provided them with extra trinkets for a while as well.

We don't have to do this. We have human-competitive AI, and there's no need to build AI with which we *can't* compete. It is in no way inevitable. By imposing some hard and global limits, we can keep AI's general capability to approximately human level while still reaping the benefits of computers' ability to process data in ways we cannot, and automate tasks we actually don't want to do. These would still pose many risks, but if designed and managed well, be an enormous boon to humanity, from medicine to research to consumer products. Imposing limits would require international cooperation, but less than one might think, and they'd still leave plenty of room for an enormous AI and AI hardware industry, focused on applications that enhance human wellbeing, rather than on the raw pursuit of power. And if, with strong safety guarantees and after a meaningful global dialogue, we decide to go further, that option is always open to us.

Humanity should choose to *not* develop superhuman general-purpose AI. Not now, perhaps not ever.

Acknowledgements

This work reflect the opinions of the author and should not be taken as an official position of the Future of Life Institute, or any other organization with which the author is affiliated. I'm grateful to Mark Brakel, Ben Eisenpress, Carlos Guiterrez, Emilia Javorsky, Richard Mallah, Jordan Scharnhorst, Max Tegmark, and Jaan Tallinn for comments on the manuscript, and to David Dalrymple for thoughts on computation limits. This work did not employ the use of generative AI models in its creation.