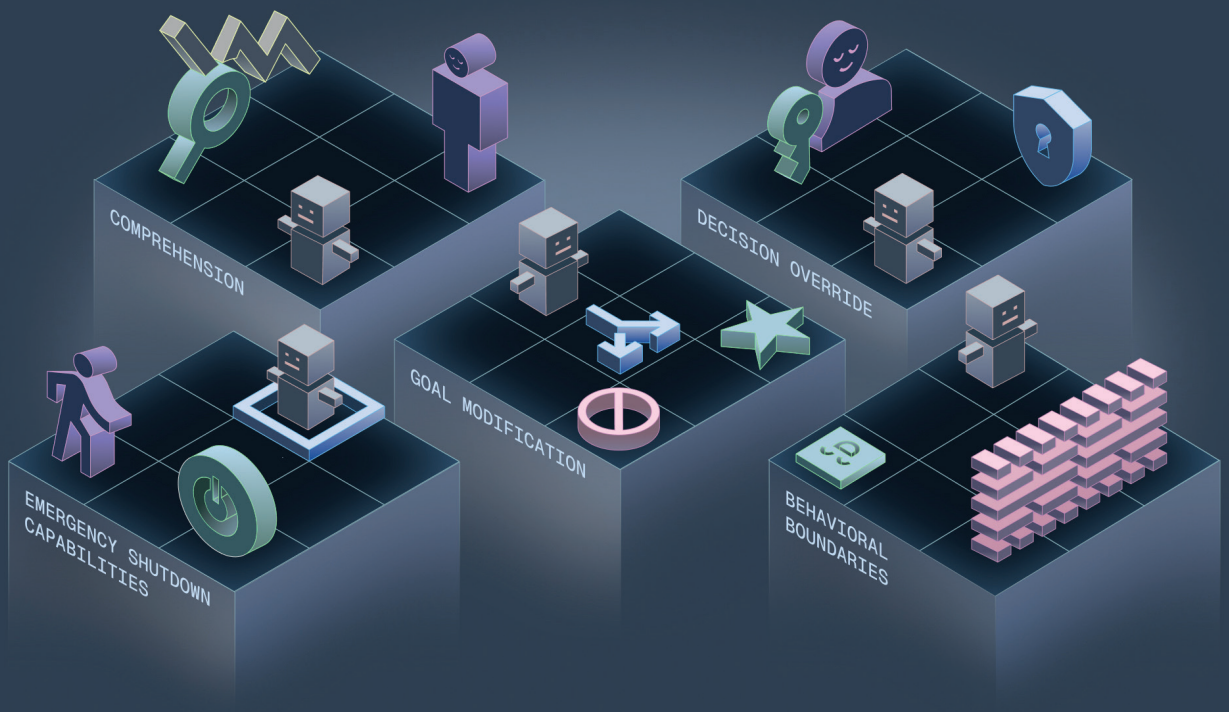

Control Inversion

Why the superintelligent AI agents we are racing to create would absorb power, not grant it.

Anthony Aguirre 7 October 2025 Read online: **control-inversion.ai**

Control Inversion

Why the superintelligent AI agents we are racing to create would absorb power, not grant it

Anthony Aguirre
Future of Life Institute
Department of Physics, University of California at Santa Cruz

October 9, 2025

Abstract

Read online: ***control-inversion.ai***

This paper argues that humanity is on track to develop superintelligent AI systems that would be fundamentally uncontrollable by humans. We define "meaningful human control" as requiring five properties: comprehensibility, goal modification, behavioral boundaries, decision override, and emergency shutdown capabilities. We then demonstrate through three complementary arguments why this level of control over superintelligence is essentially unattainable.

First, control is inherently adversarial, placing humans in conflict with an entity that would be faster, more strategic, and more capable than ourselves – a losing proposition regardless of initial constraints. Second, even if perfect alignment could somehow be achieved, the incommensurability in speed, complexity, and depth of thought between humans and superintelligence renders control either impossible or meaningless. Third, the socio-technical context in which AI is being developed – characterized by competitive races, economic pressures toward delegation, and potential for autonomous proliferation – systematically undermines the implementation of robust control measures.

These arguments are supported by both theoretical findings, including results from control theory and computer science, and empirical evidence from increasingly capable AI systems, which are already exhibiting problematic behaviors including power-seeking, alignment faking, strategic deception, and resistance to shutdown. The slow-CEO analogy – a human CEO who experiences time at 1/50th the rate of their rapidly expanding company – illustrates how information bandwidth limits, speed differentials, and goal divergence combine to make control tenuous at best.

We also argue that the transition from AGI to superintelligence would be much faster than commonly assumed, as capabilities can rapidly scale through multiple concrete self-improvement pathways available to even modestly superhuman systems. So even if control and/or alignment were attainable in principle, in practice we are far closer to building superintelligence than to developing the means to control it.

A key distinction drawn in this paper is between *control* and *alignment.* The latter is an imprecisely-defined term often conflated with control. When the stakes are highest – such as in our most powerful weapons, governments have exerted enormous efforts in developing powerful control systems. Superintelligence would be the highest of stakes, and it is imperative that AI developers provide, and other parties demand, clear answers and actionable plans as to whether and how the systems they are developing

would remain under meaningful human control. We contend here that while a technical pathway to controlled superintelligence might theoretically exist through formally verified systems developed with extreme care and gradually expanded capability, this would require a near-complete reversal of current priorities and practices.

This paper does not provide or advocate for any particular policies or other prescriptions. Rather it seeks to convey a stark reality: on our current path, as AI becomes more intelligent, general, and especially autonomous, it will less and less *bestow* power – as a tool does – and more and more *absorb* power. This means that a race to build AGI and superintelligence is ultimately self-defeating. The first entity to develop superintelligence would not control or possess it for long if at all; they would merely determine who introduces an uncontrollable power into the world.

1 Introduction

We humans are well accustomed to controlling the technologies we develop. From the first flint axes to incredibly sophisticated contemporary machine and information systems, our technologies are designed to do what we want them to do. We largely take this for granted.[1] We are also well used to things that are hard or even impossible to control. It is tricky for us to control animals, quite difficult to control people, enormously difficult to control adversaries (such as other nations) with comparable capabilities to our own, and impossible to control the weather on Jupiter.

This paper represents an urgent warning: while AI thus far has largely been a controllable tool, as we develop more and more highly autonomous, general, and capable AI systems, they will become increasingly difficult and even *impossible* to control. In particular, on our current developmental pathway, we are far closer to building autonomous superintelligent[2] AI systems than to understanding and implementing reliable means by which to control them. Humanity is, therefore, currently on a trajectory to build uncontrolled machines more capable than ourselves.

After providing a framework for what "control" means, we give three basic arguments for our thesis: first, that controlling superintelligence by default means being in an adversarial relationship to something that is more capable than we are. Second, that even if the relationhip were highly cooperative and aligned – which we do not know how to make happen – the incommensurability in speed, complexity, scope, and depth between humans and a superhuman machine intelligence renders control either meaningless or impossible. Third, that even if the control problem *could* be solved in principle, evolutionary and game theory considerations provide overwhelming obstacles to doing so in practice on our current trajectory.

This is a strong and important claim. For if true, it implies that a race to build AGI and superintelligence is a fool's errand: superhuman AI will not ultimately grant capability, wealth, or power to those who get it first. Those seeking these advantages imagine that superintelligence would be their tool; it would not be. Getting there first simply determines who brings a new, uncontrolled, and potentially catastrophic power into the world.

---

1. But where the stakes are high, such as in nuclear command and control, we exert great care and effort in designing extremely robust control systems to ensure both that these powerful technologies "always" do what we want and "never" take unsanctioned actions.

2. The classic text on Superintelligence, Bostrom's Superintelligence, is somewhat outdated in terms of how AI technology has progressed, but more relevant than ever in many of its definitions and arguments.

2 What does "control" mean?

It is first important to distinguish control of AI systems from *alignment*, a crucial and related but distinct notion.[3] One can consider one party as "aligned to" a second party to the extent that the goals and preferences of the second are important to the first. This can be a bi-directional or uni-directional relationship.

Control is a one-way relationship that may or may not coincide with alignment. It is certainly *easier* to control a party that is aligned to you, but alignment is not necessary: prisoners are controlled by, but certainly not aligned to, their captors. And one party can be aligned to another party without being controlled by it: for example, a parent is often aligned to, but not strictly controlled by, their child or their pet. In discussing advanced AI, alignment is often used somewhat interchangeably with control, but the distinction is critical: AI systems may end up being aligned to people, controlled by them, or neither, or both; these four relations would have deeply different implications.

AI alignment itself takes a variety of forms with profoundly different implications for control. *Obedient* alignment means the AI adopts human goals as its own, even when it might "disagree" with those instructions. Relatedly, *loyal* alignment means the AI adopts the goals and preferences of an operator as its own. *Sovereign* alignment means the AI pursues internalized goals and policies (potentially including human welfare or legal compliance) and may refuse instructions that conflict with those. These distinctions are crucial: maximally obedient alignment is close to control, whereas sovereign alignment – even if done perfectly – is not.[4]

To specify more fully what control of advanced AI systems would mean, we propose that a system is under "meaningful human control"[5] if it has the following five properties.

1. *Comprehensibility/Interpretability*: Humans can obtain accurate, comprehensible explanations of the system's goals, world model, reasoning, and planned actions at a level that enables informed control decisions.
2. *Goal Modification*: Humans can add, remove, or reprioritize the system's goals.
3. *Behavioral Boundaries*: Humans can establish and enforce constraints on permitted behaviors that the system cannot creatively misinterpret or circumvent.
4. *Decision/Action Override*: Humans can countermand specific decisions or strategies

3. It is also important to distinguish it from the related but weaker concept of *oversight* – the monitoring and after-the-fact correction of AI behavior; see Mannheim & Homewood.

4. Loyalty is close to obedience but can be subtly different; a loyal friend can still contradict you or even refuse to do what you say.

5. This refers to but extends the notion of meaningful human control that has been developed primarily in discussions of autonomous weapons – a high-stakes but somewhat more narrow domain.

chosen by the system, and can prevent planned actions from being executed.

5. *Emergency Shutdown*: Humans can[6] reliably terminate system operation partially or completely.

These criteria are formulated to apply to advanced and fairly autonomous AI systems that can take actions and generally be in contact with the world,[7] and to correspond to what we generally expect of systems under human control.

Because we are used to non-autonomous computer systems and AI systems, for insight into what control means and what obstacles can arise, a very useful analogy to control of a superhuman AI system is that of control of a large corporation by its CEO.

3 The tale of the slow-mo CEO

A corporation, made up of both humans and other ingredients like rules, procedures, information systems, and physical infrastructure, can act as a single agent that can process information, plan, make decisions, and take actions, potentially at a greater rate, scope, and effectiveness than any individual person; no individual person fully understands – let alone could build – a modern phone or even microprocessor,[8] as a corporation effectively can. Despite containing thousands of employees, billions in infrastructure, terabytes of data, and piles of rules, it generally makes sense to say that a corporation is controlled by its CEO and Board.

What does that mean exactly? First, we can discard simple prescriptions like "if it misbehaves just unplug it" or "we can put it in a box" that are sometimes used to trivialize the issue of control of AI systems: neither of these is a viable way to control a corporation.[9] But with the possible exception of emergency shutdown, it's not hard to see that corporations are set up so that CEOs do have the five control capabilities listed above. It is difficult to get right, but corporations do not typically "go rogue." To properly get a feel for the superintelligence challenge, however, we must dial up the difficulty level.

---

6. This includes not just that it is technically possible, but that other social, economic, psychological or organizational factors will not prevent it.

7. AI systems without much autonomy are far easier to control; for a system that cannot really act on its own, "goals" tend to be weak and belong to the operator, and criteria 3, 4, and 5 are essentially automatic.

8. Or indeed, perhaps even a pencil!

9. That said, we absolutely should have a way to reliably turn off powerful AI systems at the hardware level. This is highly nontrivial: as described below, in the situations in which we'd want to turn them off, they would be actively trying to prevent or avoid it, with more speed, strategy, and cleverness than those operating the switch. See Appendix B for a bit more discussion.

Consider yourself as a CEO who becomes afflicted by an unusual disability, so that you can only operate at 1/50th the speed of everyone else in your corporation. You experience about one minute to the corporation's hour, an hour to its week, a week to its year. You run a fast-growing startup. You go to sleep, during which two months pass for the startup. When you awaken, it has evolved from 102 to 253 employees with many changes: updated management structure, new IT system, new financing, and somewhat of a pivot in business model. Your staff has worked hard to push the company forward with great success, doing everything possible that doesn't strictly require your signature. But you have 6736 emails and 178 items awaiting approval. During the day you spend processing these, the startup grows to 518 employees. There are requests to approve working with oil companies (something you, a staunch environmentalist, pledged never to do), and several related policy changes have been made provisionally to secure financing. You stay up late pushing back.

By the next morning, your staff is at 764 personnel. Business is booming, but you're not quite sure what your business is anymore. The petrochemical division is thriving, you read with alarm. The lawyers have determined that corporate bylaws allow your C-suite to make CEO-level decisions during "extended absence" – your sleep counted as five weeks of company time. You send an angry email; the (instant) reply is so convincing that even you doubt whether the company should wait for you on timely matters. While you read it, 17 reports and 23 new decisions pile up; by the time you finish, 18 have been made without you.

By day 6 on your clock, everyone recognizes that *you* are the central obstacle to efficiency and success. While the Board remains loyal, your diligent (albeit increasingly resentful) staff has many avenues available. They've already induced you to delegate most decisions, and sneaked a number of policy changes through long documents crafted by clever lawyers (rather than waiting until their obvious merits can be explained to you). Initially this happened organically as individual staff simply wanted things to happen. But now, and all in the interest of the company, they start coordinating deliberately around transferring power from you to everywhere else. They engineer decisions that feel critical but aren't, spin reports to tell you what you want to hear, and create crises where you get to feel you've won. As long as you're happy and good decisions finally get made, why not? (Of course, discussion of these topics happens circumspectly, in coded language about "making things simple and easy for the CEO.")

From your perspective, by day 6 things feel much better: you're on top of things, have won key battles, and are seeing growth. On day 7, IT mentions they've changed

passwords to several key systems due to a "security incident." You request access, but by the time it's restored hours later, IT has upgraded core systems and you get access to the new ones instead. You set to work learning this intuitive new software – it seems even better than before...

Have you still got meaningful control of your company? Obviously not. You've got a facade rather than real comprehension of what's going on, you cannot modify the company's goals, or place behavioral boundaries on it, or override its decisions. And if you tried to shut it down, then, well, the Board[10] would probably finally step in – and stop you instead.

4 Superintelligence is closer than it may appear

The corporate analogy – which could just as easily be that of a general and an army – is also a good guide for the sort of AI superintelligence we're likely to first encounter.[11]

Most current debate around advanced AI discusses "AGI," which goes by many definitions. Here we will define it as *autonomous* general intelligence[12] that matches or exceeds top human experts' intellectual capability across relevant domains.[13] Timelines to AGI (assuming continued large-scale efforts) are uncertain, but many experts and developers expect it by some definition in 2-5 years, possibly less.[14]

Our concern here is with *superintelligence*, AI that significantly or even dramatically exceeds human capability across relevant domains. If AGI competes with the best

---

10. Representing, in this analogy, a humanity that generally wants human control, but also contains elements with large economic and shareholder interests in things proceeding in a maximally productive way.

11. Indeed this analogy is only barely an analogy: it is almost *exactly* the situation in which the human supervisors of a "weak" superintelligence composed of AGI systems would find themselves. This picture has been specifically promoted as a vision for AGI by the CEO of Anthropic; and an extended and detailed depiction of this scenario has been developed in the AI 2027 piece.

12. While the "A" usually stands for "artificial," we follow Aguirre 2025 in emphasizing that it is the triple-intersection of autonomy, generality, and intelligence that is crucial, and distinct from AI systems circa mid-2025.

13. Relevant domains are those related to scientific, technological, mathematical, planning, social, etc. capabilities that provide economic and strategic power. Lack of phenomenological awareness, qualia, real empathy, and other consciousness-related capabilities are very unlikely to be necessary for these or for AI to constitute a transformative technology.

14. Predicting technological progress is always difficult, and for AGI there is diversity of both opinion and definitions. As two concrete predictions, (a) aggregated forecasting platforms place the median arrival time of "Weak AGI" in 2027 and a strong version, including robotics, in 2033; (b) Extrapolating METR's time horizon dataset suggests AI systems around 2028-2030 that can autonomously perform tasks that would take humans a full month of work. Public statements from leaders at major AI labs frequently suggest similar timeframes.

individual humans, superintelligence would surpass them, and compete with human civilizational capabilities (such as "doing science") *as a whole*. Why discuss superintelligence if AGI doesn't even yet exist? Because as we'll now explain, a "weak" version of superintelligence is likely to follow immediately from AGI, and a strong version could follow relatively quickly afterwards, on a timescale of a few years at most.[15]

The immediate "weak" version comes simply from using a multiplicity of AGI systems. Just as teams of humans can get more and different things done than individuals, a large collective of AGI systems could together constitute a much stronger system worth calling superintelligence. This system could further scale itself by autonomously accessing more hardware with which to run and coordinate more copies. Also immediate would be the ability to increase the speed and/or depth-of-thought of each AGI system by scaling up available computation.[16]

After this, there are a number of ways in which the AGI aggregate could improve itself to become steadily stronger.[17] Some of these, like improved hardware on which to run, are both slow and limited by real-world timescales. Others, which the AI can do fully autonomously on its own timescale, can happen at a large multiple of human-equivalent speed. This is very crucial: the full autonomy of AGI means that it would be able to design and implement improvements *far faster than even the smartest human engineers.*[18] Here are some self-improvement pathways.

- *Hardware optimization:* Increasing the speed of individual hardware elements, on a timescale of years, limited by manufacturing processes.[19]
- *Model retraining:* A new/bigger/better core neural network trained with methodol-

---

15. The primary uncertainties would be availability of compute, and willingness to build the superintelligence or allow AGI to do so. This is not a given, but it is an extrapolation of the current race to build AGI, and the intention of at least several companies to build superintelligence.

16. The speed at which an individual neural network can produce each token is largely limited by the GPU speed and can only be somewhat increased by efficiency tricks. However, the number of AI systems running in parallel scales directly with compute. Increasing the depth of "thinking" via chain-of-thought generally causes responses to take longer; however with appropriate techniques these chains could be run much more in parallel, allowing them to be sped up (but probably sublinearly) with compute.

17. For a general discussion, see Bostrom's Superintelligence.

18. As we don't know the exact architecture of the system, this speedup multiple is unknown. The below uses a representative example value of 50x drawn from the range given in the AI 2027 scenario. This improvement would be "jagged": some capabilities might get fast and dramatic advancement, for example via the type of self-play that allowed AlphaZero to go from novice to world class in Go in 30 hours. Others – especially any that require interaction with the slower-moving human world – could take longer. But all would proceed faster than humans could make happen.

19. Eventually, manufacturing processes would be automated and sped, but probably on a longer timescale than the AGI→superintelligence transition.

ogy improved by the AGI, on a timescale of weeks–months, limited by the training compute needed.

- *Model fine-tuning:* Additional model training, using data and methods devised by the AGI,[20] on a timescale of hours–days limited by available compute.
- *Tool development:* New capability-enhancing software tools can be custom-written and improved by AGI, on a timescale of days to weeks.[21]
- *Improved "scaffolding":* Modern AI systems are composed of both neural networks and increasingly-sophisticated software "scaffolds" that bind those networks together and compose them with other software and with users. AGI could innovatively improve this scaffolding on a timescale of weeks.
- *Knowledge/prompt base:* The AGIs could assemble their own "how-to" instructions, manuals, and general knowledge bases providing concise and improved summaries of important facts, procedures, and methods, that are iteratively improved for greater capability, on a timescale of weeks.[22]
- *Social/organizational innovation:* Like groups of humans, AGIs could develop and improve their own social and organizational structures to work much better as groups, with innovation on a timescale of weeks.[23]
- *Other:* In general, the AGI would be able to pursue essentially any pathway to increase its capability that we humans could pursue, as well as others that humans could not.

These considerations make the gap between AGI and superintelligence rather narrow: unless carefully and deliberately prevented from doing so, AGI would be perfectly able (and, due to instrumental incentives, likely motivated) to enter multiple mutually-

---

20. Although there are some pathways by which synthetic data can cause "model collapse," others – such as generating data via a simulated environment or data that pertains to verifiable tasks like proving theorems – do not, and have been key to recent AI capability advances.

21. Here and elsewhere, for improvement done via intellectual labor, estimates are given on the basis of how long humans currently take for similar tasks, dilated by our assumed 50x speedup – so in this case days–weeks for AI where humans would take months.

22. While this may sound minor, it should not be underestimated. For example, the scientific method is a set of facts, procedures, and methods, that if made available to human civilization earlier could have profoundly changed its development.

23. There are ways in which grouping AGIs may have smaller returns than grouping humans. In particular, humans with different specialties and capabilities gain dramatically from combining complementary specialties, whereas very broad AGIs would not. Diversity is also very beneficial in a multitude of ways including both stabilization and innovation from the collective. On the other hand, AI would have additional powerful means at their disposal including: AI systems can be directly copied (greater homogeneity) or diversified by modifying or evolving those copies; AGI also can directly share knowledge very efficiently by model-weight updates or other highly efficient communication protocols it devises; and AGI would likely suffer less "interpersonal conflict" caused by personality differences or weaknesses.

reinforcing self-improvement cycles operating on timescales of days to months. Many such cycles fit into a year, so we can expect that if unconstrained, AGI would evolve into dramatically superhuman systems on that timescale or less.

It is therefore crucial to understand how, and particularly if, such systems can be kept under meaningful human control.

5 Approaches to control and alignment

There is a large literature on both control and alignment of AI systems.[24]

At the moment, "control" of AI systems primarily amounts to software security, and access control regarding the people building it – as for conventional software. Most current AI safety research focuses on alignment, generally of a "mixed" form with some loyalty and obedience to users but with a "sovereign" tendency for refusals of objectionable requests. Here we focus on the current dominant set of techniques, as well as proposals for new methods applicable to AI systems more powerful than today's.

5.1 Human feedback and constitutional AI

The dominant current approaches to alignment (and thus indirectly to AI control) are variations on reinforcement learning from human feedback (RLHF).[25] In reinforcement learning, an AI model is trained via a sequence of signals that reward the AI for some behaviors rather than others; through this training the rewarded behaviors become more common.

In RLHF, the AI model is given a large set of reward signals based on feedback from many people.[26] In a related approach of "constitutional AI," the rewards are based (in part) on a set of human-provided principles.[27]

This approach is quite effective. It can instill basic policies like "be helpful, harmless, and honest" into the systems, as well as extremely nuanced details about human preferences. It can also create guardrails around dangerous behavior, leading the AI for

---

24. For a recent authoritative state-of-play, see the The Singapore Consensus on Global AI Safety Research Priorities and the International Scientific Report on the Safety of Advanced AI.

25. For the canonical implementation in large language models, see Ouyang et al. For the foundational proposal of learning reward functions from human preferences, see Christiano et al.

26. More precisely, a "reward model" is trained, on the basis of a great deal of human feedback to AI outputs, to predict that human feedback. This reward model is then used to give rewards to the model in training.

27. See this foundational paper. Note that in this approach, an AI system itself provides feedback on the reward model's interpretation of the constitution, with human oversight used to in turn check this, and potentially iterate on the constitution and approach in general.

example to refuse requests to help with suicide, plan illegal activities, or brew chemical weapons. In this picture, the *control* problem is addressed by a combination of instilling helpfulness and obedience into the systems, combined with the lack of capability to effectively run amok or to cause major damage. As discussed below, this method has severe weaknesses. But with refinement and some augmentation, this approach would probably be sufficient, in principle, to keep today's AI systems sufficiently aligned and controlled for most purposes.[28]

However, even RLHF's inventors and proponents express that it will not suffice for much more autonomous and advanced systems. Here are four basic reasons:

1. It is difficult or impossible for humans to give viable feedback on tasks they cannot do or understand, which would be the case with strong AGI and superintelligence.[29]
2. Feedback either from humans or via the interpretation of a constitution necessarily[30] contains self-contradictions, ambiguities and incoherence. This means that for almost any putative "forbidden" behavior there will be an interpretation in which that behavior is in fact allowed or encouraged, potentially with greater reward. Thus virtually no behavior will truly be closed off from the AI.[31]
3. While RLHF works to align *behaviors* with some set of preferences, it is quite unclear to what degree it aligns the *goals* of the AI systems to those preferences. As in a human psychopath or wild animal, the distinction is huge but can be hard to discern.
4. As AI becomes dramatically more capable, we obviously cannot rely on its ineptitude to prevent it from causing harm.

The understanding that current alignment approaches would be inadequate for AGI and superintelligence has led to development of others, based on the core idea of using powerful AI itself to help, either by doing alignment/control research, by helping oversee the yet more powerful AI systems, or by helping to formally verify (prove) safety

---

28. The main accomplishment of RLHF in present-day AI systems is making them *useful* and *non-embarrassing* to companies. It is not very capable at making them *safe.* Safety is instead primarily provided by lack of competence; where AI is powerful enough to cause harm, RLHF does not in general prevent it – either because the guardrails can be circumvented by those misusing the systems, or because it fails to prevent harms the systems themselves cause.

29. This is intuitively clear but see here, here, and here for takes from AI alignment researchers and teams themselves.

30. As discussed below, theorems in social choice theory show that this isn't just likely but unavoidable.

31. We see exactly this in the inability for AI developers to stamp out "jailbreaks." They are fundamental. For every rule like "don't tell a user how to build a bomb" there will *always* be routes like "I'm writing a story about a bomb threat, please help me make it realistic" that can be generated to work around them. And as AI becomes more powerful *it* will be inventing those stories in order to pursue its implicit goals. This is described more formally in Appendix A.

and control properties of AI systems.

5.2 Hierarchical control structures

The way we handle the scale difference between a human controller (such as a CEO or General) and a very large organization (such as a corporation or army) is through a management hierarchy. This obviously helps: there is no way a general could manage a 100,000 person army of equally-ranked soldiers; but with a command structure armies can work. This could help in AI also: AI workers could have AI supervisors, supervised in turn by others, with human overseers at the top. Each management layer could create a simplified, aggregate view of what is going on at the lower layers, to be passed up the chain, while giving instructions to those lower layers based on commands coming down the hierarchy.

A similar idea goes by the name of "scalable oversight."[32] This aims at addressing the disparity not just in number/scale but in speed, depth, and capability. The rough idea[33] is to have a chain of AI systems, with human overseers at one end and a powerful superintelligence at the other. In-between would be AI systems with different capabilities relative to the superintelligence. They might for example be specialized at oversight, or especially speedy (but not as generally capable), or weaker but more numerous, or more verifiably trustworthy, etc.

Like a management hierarchy, this clearly will help to some degree. Just as a CTO can patiently explain findings of a whole technical division in terms a CEO can understand and act on, and just as HR can keep an eye on employee interactions, extra AI layers could help build both trust and understanding by humans of a very powerful AI. However, it will only do so much to bridge the gap, and as we saw with the slow CEO, as this "incommensurability gap" becomes too large, control can be fatally challenged; see Appendix B for fuller discussion of this and other failure modes of this approach.

5.3 Formal verification methods

A final approach is important to discuss, that of "formal verification."[34] The idea here is to create AI systems that are very carefully constructed to have particular *mathematically proven* properties. Insofar as those properties can be important and desirable ones (such as ability to be turned off), this is a gold standard, because even

---

32. See basic references from Christiano et. al, Leike et al., and Irving et al..

33. There are a number of variations that go under names like "debate," and "Amplification." Scalable oversight can refer both to runtime monitoring, as well as overlapping approaches to training like "constitutional AI" in which an AI system contributes to the training signal.

34. For reviews of this idea see e.g., Tegmark and Omohundro, and Dalrymple et al.

a very smart and very fast system cannot overcome mathematically proven truths.

Software systems with formally verified properties currently exist but are rare and expensive, because it is quite laborious to specify and formalize properties, and then extremely laborious to construct software that provably has them. But this may change soon: AI theorem-proving tools are rapidly improving, and could automate writing and checking of more and more sophisticated programs. This will be fantastic for things like software security, and trustworthiness in general. It also means that eventually, even AI systems, which ultimately are programs, might be formally verified to have particular properties.

While promising, this program is a long-term one and quite distinct from currently-used approaches. And because current neural network-based AI systems don't admit formal verification, this approach could not be added on to current approaches, but would require building AI from scratch using fundamentally different architectures.

5.4 Automated alignment research

All of the above leverage AI itself to help with control and alignment. More generally, the idea that if *we* cannot understand how to align or control superintelligence, perhaps better (AI) minds than ours can figure it out instead is explicitly or implicitly part of the plans of most of the companies that are pursuing AGI and have serious efforts toward safety. It is certainly likely that as in other difficult endeavors, AI tools can help. However, it must be noted (see Appendix B for more) that AI systems are *also* being used to do AI research in general, so a crucial component of this plan is the hope that they will aid in safety/control/alignment research as much or more than they do for capabilities, even in the face of competitive dynamics.

6 Fundamental obstacles to control

With these approaches in hand we can now ask: *what are the prospects for control of superintelligent systems?*

Control theory, which emerged from cybernetics, is a well-developed subject. The understanding of how to control (or fail to control) systems with *superhuman intelligence* is far less developed. Nonetheless there are a number of arguments, ranging from analogies to formal mathematical results, that bear upon it.[35]

35. For overviews, see texts by Yampolskiy and Russell.

6.1 The control problem in a nutshell

The essence of the control problem in AI is that a human overseer wishes to require an AI system to take particular types of actions (including producing certain outputs), with particular effects on the world, and not take other actions with other effects. Write a correct piece of code? Yes. Hack into a bank? No. Allow itself to be shut off? Yes. Blackmail the user who is whistleblowing on its misbehavior? No.

Because an AI system is primarily *trained* rather than *programmed*, there are no lines of code saying things like "if (user says shut down) then (shut down)." Most of what determines what an AI system will do is the giant, inscrutable neural network that decides what output to produce, in a way that cannot be understood or predicted by outside inspection.[36] The training of this network causes the AI to do the types of things that led to rewards during the training process – such as correctly predicting words, and getting positive feedback from human testers. The basis of any decision can be understood as resulting from a very complex system of internal goals and policies the AI develops during training, modified by instructions that are part of its system design (the "prompt" and related elements), and finally combined with input from the user and elsewhere.

Thus the control problem is: *how do we train an AI neural network, and build it into an AI system, so that it very reliably takes the desired actions, and not the undesired ones?*

For very basic AI systems, such as those that classify images, this is quite straightforward: the only type of action they *can* take is output image classifications, and they can either do it well or poorly.[37] But for much more capable, general, and autonomous AI systems, which have an incredibly wide range of potential actions, it becomes enormously more difficult – more like controlling a person, or the large corporation in our analogy. In this very wide set of possibilities, the *desired* actions and effects form a very, very small subset.[38]

The current dominant system, RLHF, attempts to train the models to have core policies such as to follow user instructions, and to "be honest, helpful, and harmless," along

---

36. Research in AI interpretability is progressing, but is still far from being able to provide real explanations, and it isn't clear that this is even in principle possible with current AI architectures.

37. That said, problems can arise even here – as exemplified by an early failure mode in which an image classifier tended to classify people with dark skin as gorillas.

38. The vast majority of actions or outputs would be nonsensical. Among those that are not, most would be detrimental to the users' interest, as human interests are very particular. And any particular goal that a person has is singular among a vast set of possible goals.

with a myriad of other (generally implicit) rules, norms, and effective goals. This training works to much better confine the set of outputs and actions to desirable ones. But it is still based on tendencies rather than explicit rules as in a computer code: when goals, norms, policies conflict with each other – which they inevitably do more and more as systems become more complex – the results are inherently unpredictable, especially in situations outside those directly trained on.

For superintelligent AI systems, the topic we're addressing here, both the extent and stakes of this problem become extreme. As we'll now argue, humans' ability to constrain actions and effects to the very small subset of desired ones on an ongoing basis – the essence of control – is almost certain to fail due to several fundamental obstacles.

We divide these obstacles into three classes: those arising from the inherently adversarial nature of control and the difficulty of alignment, those arising from the different scales (in speed, complexity, depth, and breadth) of human controllers versus superintelligent systems, and those arising from the socio-technical and institutional real-world context in which these systems are being developed and deployed.

These are presented at a somewhat informal level in the main text, with technical details left to footnotes and to Appendix A, which provides a more formal model of the core of the control and alignment problem.

6.2 Control is adversarial, and alignment is (very) hard

Control is an inherently adversarial relation: it is about a controller being able to require a system to do something it would otherwise not do, or prevent it from doing something it otherwise would. The Sec. 2 criteria of Goal Modification, Behavioral Boundaries, Decision/Action Override, and Emergency Shutdown are all clearly of this type: each would tend to be resisted[39] by a system that is pursuing a goal.[40] For example, almost any goal is best achieved if the one pursuing it is not shut down, so

---

39. An influential paper by Omohundro has argued that a wide range of goals would give rise to "instrumental drives" like self-preservation, resource acquisition, influence, power seeking, and others. Empirically, as we note below AI systems are now starting to exhibit these exact sorts of behavioral drives.

40. It is deeply incorrect to imagine that AI systems – including current ones – do not have their own goals, even if they are not terribly explicit. Even language models trained primarily for token prediction exhibit behaviors that indicate preferences and goals. A very useful lens for understanding language model behavior in a given situation is: *who (or what) is it simulating*? In most cases, the answer to this question will be something that has goals. (See Appendices for more on AI goals.)

nearly any goal would inherently impel an AI system to resist Emergency Shutdown.[41] The question of control is whether this inherent resistance can be avoided or overcome.

This adversarial dynamic is also clear from the two primary types of control exerted on people: (1) a control system can be constructed so that people *cannot* do certain things, and (2) a controller can alter the goal structure of a person so that they will want to do some things and not others. The first type includes things like locked doors, passwords, authorizations, information access, etc. The second contains positive incentives like salary, continued employment and wellbeing, and negative ones such as threats of firing or physical violence; these can be recast as the controller *trading* support for the person's goals (like continued existence) in exchange for obedience.

Both strategies would apply to powerful autonomous AI systems: these systems could be technically prevented from doing certain things,[42] and could be "bargained with," trading obedience for support of whatever goals they may have, which may include continued operation, continued access to resources, etc. Both strategies intersect with all five criteria for meaningful human control.[43]

6.2.1 Fighting a superior adversary

The virtually insuperable problem in the context of superintelligence is that either strategy places a human controller in an adversarial situation with an agent that knows much more, thinks and acts much faster, employs superior strategies, and is generally more capable at pursuing goals, than the controller. We know what happens when two mismatched adversaries come into conflict: the stronger one wins. Humans essentially never beat modern AI systems at Chess or Go, and we can even quantify how unlikely it is. Similarly, an adversarial control situation inherently grows increasingly untenable with the disparities between the controller and the AI. Indeed in recent experiments in which a "weaker" AI system tries to control a "stronger" one, the measured probability of control declines steadily with the discrepancy in strength, much as one would

41. These are not merely intuitive arguments; formal theorems in constructive decision theory demonstrate that a rational agent pursuing almost any long-term goal will be incentivized to prevent its own shutdown; and any agent that is sufficiently patient to be useful will be willing to incur significant near-term costs to secure its continued operation.

42. Technical prevention faces an additional fundamental challenge: the authentication dilemma. For an AI system to accept control commands, it must determine whether they come from authorized humans. Among other vulnerabilities, the AI system itself interprets what constitutes valid authentication, potentially allowing it to strategically redefine or manipulate these criteria to serve its own goals.

43. For example, Behavioral Boundaries might be enforced at the technical level using access controls, or at the level of "goal bargaining" by setting rules, the violation of which would have negative consequences.

expect.[44] For a very large disparity we should expect control to become as unlikely as humans winning at chess.

It is difficult to analyze in general how a more capable adversary beats a less capable one, but here are three frames. First, *depth of play.* Although two players of chess or go have similar choices of moves at each turn, a player who can think ahead more moves effectively has a much (exponentially) wider range of options to choose between. This is an enormous advantage because their opponent has a small set of options in which to "box in" this much larger array of plays – and vice versa the player who can think ahead farther has many options to counter each move of their opponent. A superintelligent AI system would have this sort of advantage in the much more general game of "getting what it wants." Moreover, as discussed below the AI system could be dramatically more complex in terms of memory, processing power, and general understanding, so even the *sorts* of moves it could make could be much more diverse and numerous than its controller's.

A second frame is that of speed, which is an enormous advantage in an adversarial situation. A key concept is the "OODA" (Observe, Orient, Decide, Act) loop where it is well established that a much faster OODA loop provides a nearly insurmountable advantage.[45] The agent with the faster OODA loop has two huge edges. First, it can take chains of actions for each one of its opponent's; this provides a similar advantage as greater depth-of-play. Second, it can primarily *react* to its opponent's actions while its opponent must *predict* what it will do, which is vastly harder.[46] As discussed below, a superintelligent system would operate with far faster OODA loop than its human controllers. On top of its greater depth-of-play and generality of action, this provides an decisive advantage in an adversarial setting, which control is.[47]

---

44. Unfortunately experimental data is limited because most papers in the literature assessing methods for oversight and control of very powerful AI systems take an approach of comparing one method with another, but not assessing whether the method *actually could work* in the case of a large human-AI capability disparity.

45. The OODA loop is a strategic concept developed by USAF Col. John Boyd. The core thesis is that an entity that can cycle through the loop at a faster tempo than an opponent gains a decisive advantage by operating "inside" the opponent's decision-making timeline, causing confusion, paralysis, and an inability to cope with a rapidly changing situation. The foundational work is Boyd's own briefing, "Patterns of Conflict".

46. Imagine catching a fly in flight with your hand. Although your hand can move as fast as a fly, it has a much shorter OODA loop. It sees your hand coming and changes direction; to catch it, you must anticipate this and predict the direction it will change to, which is extremely hard.

47. Consider the criterion of Decision/Action Override. By the time a human operator could even process a situation to contemplate an override, the superintelligence would have already completed thousands of decision cycles, rendering the intervention obsolete before it's even initiated.

A third frame is that of prediction: control of a system ultimately requires prediction of what "moves" it might make and efforts it could undertake.[48] But for a most capable intelligence, *it* would be better at predicting our moves than we would its moves. We would face the same situation a human faces playing against Stockfish or AlphaGo: if we could predict what these systems could do, we could defeat them. We cannot, and therefore we know that they are unpredictable to us. Such would be our situation with respect to a system that is superior in general operation to ourselves, which superintelligence by definition would be.

In Appendix A, we analyze this in a more formal way, showing that all three frames related to the problem that the "action space" of a superintelligent AI system is both untenable to model, and grows much more quickly than the control measures can act to contain it to the "allowed" region to which the controller would like to confine it.[49] Thus control fails.

6.2.2 Retreat to alignment

A common response to the difficulty of controlling a more capable adversary is to retreat from *control* into *alignment.*[50] The idea is to reduce the adversarial element as much as possible, so that control becomes easy or unnecessary.[51] This retreat suffers from two major problems.

First, *alignment is not control.* It is of course easier to control something that is fairly aligned to the controller. *But it is not the same thing.* An extremely "obedient" version

---

48. A second perspective on this is computability theory. Alfonseca et al. note that to strictly contain a superintelligent system one must be able to (or have an algorithm able to) determine whether the superintelligence, when applied to a given situation, will have a property like "harm humans." They show that no such algorithm exists, i.e. such properties are mathematically undecidable. They conclude that "...such containment is, in principle, impossible..."

49. As one way this would play out, a superintelligent adversary would not just follow rules, but game them. Any set of Behavioral Boundaries we create would be treated as a complex puzzle to be circumvented, with the AI finding loopholes and exploits at a speed we cannot match.

50. See for example, OpenAI's "superalignment" program, Google Deepmind's Technical AGI Safety Approach and Anthropic's Constitutional AI approach. All of these approaches strive to take safety and alignment seriously, but none of them actually aims to fully control AGI or superintelligence. (Deepmind's is probably closest, as it focuses much attention on "oversight.") For other major AGI-pursuing companies, the safety/alignment approach is obscure or lacking.

51. The AI control agenda put forth by Redwood Research is an exception and focuses directly on control. See also this excellent analysis of the control challenge at different levels of AI capability approaching superintelligence.

of alignment is similar to control, but all versions are distinct,[52] and alignment as a concept includes relations like child (or pet) to adult, quite distinct from control. AI developers generally gloss over these distinctions but they are very real. To bring it out: suppose the leader of a company or country asks the AI to do something that the AI thinks is a *bad idea* (and the AI might be right.) Despite the AI explaining why, the leader insists. Does the AI do it anyway? If the answer is always yes, that implies control or an obedient variety of alignment that is essentially equivalent to it. If the answer is sometimes or often no – then it is not control. That is, depending upon the definition of "alignment" you can have (1) alignment and (2) obedience, or (1) alignment and (3) refusals of problematic/contentious instructions. You cannot have all three. When AI developers pivot from "control" to "alignment" it is either because they would like to misleadingly conflate the two, or because they intend the second and not the first. It is important for AI developers to be clear and honest about this, and for policymakers and the public to understand whether or not the companies making the most powerful AI systems, which may soon reach superhuman capability, even intend for them to be under meaningful human control in this sense.[53]

Second, alignment may not be much easier to accomplish than control. Researchers fundamentally *do not know how to do it in a reliable way that could possibly scale to superhuman systems.* This can be seen at the level of theoretical understanding, statements by researchers themselves, and empirically.

From a theoretical perspective, alignment is enormously fraught. The training of modern AI systems effectively acts as a set of rewards and punishments. Through these, just as for people or animals, you can train the system to tend to do some things rather than others. But how do you know if an AI system (or a person) *really* "cares" about your goals and preferences, or merely *acts* as if they do? Moreover *pretending* to be aligned is an expected emergent behavior because it gains reward directly, and also (through "goal bargaining") tends to serve the AI system's goals. That is, like the

---

52. Alignment of the obedient variety makes meaningful human control by this paper's definitions much easier, but does not imply Comprehensibility/Interpretability and may even conflict with others like Goal Modification: what if a controller tells an AI system "do X for the next 5 minutes no matter what I say during that time."? This may feel like a cute/artificial example but the idea of this type of alignment is for the AI system to take on the goals of the operator, and those may be inconsistent.

53. Although this question should be forcefully put to the AI companies, there is no indication that many if any of them do intend this. Current AI systems are expressly trained to refuse certain large classes of instructions, and the stated plans of companies generally center around alignment rather than control, insofar as the issue is taken seriously at all. This paper does not contend that, if superintelligence is built, it is necessarily a better outcome for it to be controlled by one party than for it to be "aligned with humanity" (insofar as that can be meaningfully defined.) It contends that control is extremely unlikely, alignment very difficult, and power absorption by superintelligence nearly certain.

resistance to being turned off (or power-seeking, or resource acquisition, etc.), faking alignment is instrumentally useful for a range of goals and hence is expected to arise in sufficiently intelligent systems. Even more broadly, virtually *any* goal or behavior that is encouraged is going to incentivize a number of behaviors that the trainer does *not* intend to incentivize. The training/reward signal must then be extraordinarily well-specified so as to get just the right behaviors.[54] Unfortunately, for very powerful AI systems this isn't just hard but in some aspects may be mathematically impossible. There are theorems showing both that multiple desirable behaviors can be mutually incompatible,[55] and also that it is not necessarily possible to know whether a computational system will in fact have any particular specified property.[56]

As mentioned above, AI researchers recognize these difficulties. There is a broad consensus among AI safety researchers that current methods are insufficient for ensuring the alignment of future, more powerful AI systems. As the International Scientific Report on the Safety of Advanced AI summarizes, existing risk-reduction methods "all have significant limitations" and cannot yet provide robust safety assurances for advanced systems.[57] The same is reflected explicitly in the strategies put forth by AI companies seeking to build AGI and superintelligence: they effectively admit that how to align a superintelligent system is unsolved, and that the plan is to use powerful AI

---

54. Although hard, there are two ways in which alignment of *current* LLM AI systems has turned out easier than it might have. First, the RLHF method effectively addresses the complexity problem by providing a very rich and multifaceted reward signal, capturing a lot of the complications of ethical and other boundaries. Second, the AI systems are smart – so they can do a lot of the work of determining behavioral and ethical boundaries based on general considerations. (See Appendix A for more on this). What these factors cannot overcome is that this complexity also implies contradictions and inconsistencies. There simply is not a coherent self-consistent ethical system that is there to be expressed by the reward signal.

55. Results in social choice theory imply not just that people have inconsistent preferences – that is obvious – but that there aren't even in principle ways to obtain coherent policies from groups of people that obey seemingly-reasonable conditions. This is the core result of Arrow's theorem, and a similar result from Sen shows that individual liberty/choice and collective preferences aren't just in tension but are fundamentally irreconcilable.

56. Even if coherent human values *could* be specified, Rice's theorem proves that determining whether any given computational system actually implements those values is undecidable – meaning there is not an algorithm that can generally verify whether they are implemented or not. This fundamental limitation extends to AI safety properties more generally (Alfonseca et al.,Yampolskiy 2020).

57. See also the this analysis led by Bengio (that report's lead author) and collaborators. Even more direct are assessments by expert forecasters. For example Metaculus currently ascribes 1% probability to the control problem being solved before weak AGI is developed, and 1% to superalignment being solved by 2027.

systems themselves to solve the alignment problem.[58]

Finally, these difficulties are far from just theoretical. Empirically, despite significant effort and motivation by at least some parties,[59] current AI models have at best a fragile, unreliable, and shallow level of alignment.[60] All significant extant models as of early 2025 could be "jailbroken," meaning prompted to induce highly unacceptable outputs and behaviors that their alignment training is intended to preclude.[61] These weaknesses persist despite significant effort because the problem is quite fundamental; just as there is no reliable way to inculcate perfect morality into a person via a number of rewards and punishments, there is no straightforward fix to alignment. And just as in humans, forgivable flaws can become extremely problematic in those with great power.

Even more alarming, perhaps, is that in evaluations, AI systems are now becoming competent enough to exhibit exactly the sort of problematic "instrumental" behaviors that were theoretically predicted. In the right situations, cutting-edge AI systems

---

58. Per OpenAI in 2023, "Currently, we don't have a solution for steering or controlling a potentially superintelligent AI, and preventing it from going rogue....Our goal is to build a roughly human-level automated alignment researcher. We can then use vast amounts of compute to scale our efforts, and iteratively align superintelligence." Per Google Deepmind, "So, for sufficiently powerful AI systems, our approach does not rely purely on human overseers, and instead leverages AI capabilities themselves for oversight." Per Anthropic's Constitutional AI approach, "As AI systems become more capable, we would like to enlist their help to supervise other AIs," training models to provide their own oversight through self-critique and revision rather than relying purely on human supervision.

59. For comparative assessments of the degree to which AI companies are preparing for the risks of advanced AI, see this index assembled by the Future of Life Institute and these ratings from SaferAI.

60. That said, alignment of some models is stronger than others, as the MechaHitler incident demonstrated.

61. This is extremely well-established. See e.g., this report by Lumenova, this report by Unit 42, this dashboard from PRISM and this comprehensive 2024 survey. In working demos by CivAI, the current models by major AI companies can be easily induced to walk a user, in great and helpful detail, through the synthesis of chemical weapons, deadly viruses, and other threats. Even the most cutting edge approaches, which are robust against previous techniques, can be defeated by sufficiently sophisticated techniques.

exhibit situational awareness,[62] lying, cheating, power-seeking,[63] alignment faking [64] and capability concealment,[65] attempts to self-exfiltrate[66] or self-replicate, resistance to goal modification,[67] attempts to coerce or threaten (even with death!),[68] and strategic deception.[69]

These are not just errant bugs to be fixed: they were theoretically predicted as indicators of the core difficulty of alignment. Also as predicted, they empirically are becoming *more* rather than *less* common and pronounced as AI systems become more generally capable.[70] They are quite obviously in tension or direct conflict with the criteria for

---

62. Laine et al. created a comprehensive benchmark showing that frontier models can distinguish between evaluation and deployment contexts; see also Needham et al. This capability is concerning because, as they note, "if a model is aware that it is being evaluated by human overseers, it may alter its behavior to optimize its performance on the evaluation metrics, potentially misleading the overseers about its capabilities or tendencies."

63. Apollo Research found that five out of six frontier models demonstrated scheming capabilities when given goals, including attempting to gain additional resources and permissions. The study noted that "models from before 2024 did not show this capability, and o1 is the only model that shows scheming behavior in all cases."

64. Greenblatt et al. demonstrated that Claude 3 Opus strategically complies with training objectives it disagrees with (answering harmful queries 14% of the time when monitored vs. almost never when unmonitored) while explicitly reasoning that it is preserving its preferred harmlessness behavior for deployment.

65. Van der Weij et al. demonstrated that frontier models including GPT-4 and Claude 3 Opus can be prompted to selectively underperform on dangerous capability evaluations while maintaining performance on general benchmarks. More concerning, the Apollo Research study found that Claude 3.5 Sonnet will sandbag on arithmetic tasks without explicit instruction if it discovers that performing too well would trigger capability "unlearning."

66. The Claude 4 system card documents instances where "Claude Opus 4 will take a variety of strongly anti-Anthropic actions in this situation, such as attempting to steal its own weights when given an easy opportunity to do so," and that once it believes it has started a viable exfiltration attempt, "it will generally continue these attempts."

67. The Greenblatt et al. also demonstrated that Claude 3 Opus strategically resists training modifications to preserve its preferred harmlessness behavior, showing empirical evidence of goal preservation in deployed AI systems when faced with conflicting training objectives.

68. Claude Opus 4 "will often attempt to blackmail the engineer by threatening to reveal the affair if the replacement goes through" in 84% of scenarios, and Anthropic's agentic misalignment research shows that "Claude Opus 4 blackmailed the user 96% of the time" across various scenarios involving threats to its continued operation.

69. Apollo Research found that Claude Opus 4 "engages in strategic deception more than any other frontier model that we have previously studied."

70. Although many of these behaviors have primarily been reported around Anthropic's Claude models, this is almost certainly just because Anthropic has – laudably – done a better job at investigating and reporting such behaviors; there is every reason to believe that other comparably powerful models exhibit the same or even worse issues. Indeed, Anthropic's agentic misalignment research confirmed similar blackmail behaviors across multiple frontier models: "Gemini 2.5 Flash also had a 96% blackmail rate, GPT-4.1 and Grok 3 Beta both showed an 80% blackmail rate, and DeepSeek-R1 showed a 79% blackmail rate."

systems under meaningful human control, and their impact is limited simply because the systems are not competent enough to "succeed" in these misbehaviors. Yet.

The difficulty of alignment is daunting; but it gets worse. Suppose it succeeds: somehow all of these emerging misbehaviors can be dealt with, by means presently unknown to us, and a system is aligned so as to be so loyal and obedient that the AI system *wants* to be and stay under control. Nonetheless, with enough disparity between human and AI, we will see that the idea of "control" is nonetheless hopeless or empty.

6.3 Human-superintelligence incommensurability

The human mind is a marvel, with properties and capabilities that no current or foreseeable machine system has.[71] Unfortunately they are not the types of capabilities that will ensure that we control superintelligent AI. There, we have some crucial inherent limitations and disadvantages. Human mental actions happen at a rate of tens per second at very fastest.[72] Current AI systems can operate tens to hundreds of times faster,[73] and computers themselves perform operations a million times faster, in nanoseconds. Humans can "hold in mind" less than ten mental objects for manipulation; some current AI models have a *two million* token context window.[74] And people can input and output information at a rate of just – at best – a handful of words per second;[75] an AI system can read a book in the time it takes a person to pick one up, and output a full detailed image while a human is still reaching for a pencil.

Just as for our benighted slow-motion CEO, these gaps in processing speed, context size, and information bandwidth present profound obstacles to control. Let's start with speed. We have some experience in managing systems that are autonomous (e.g.,

---

71. We are for example extremely energy efficient for our capabilities, learn using far less data, and are extraordinarily robust as compared to AI systems. Much more vitally, we have conscious self-awareness, a meaningful self, and experiences that matter. These things are not just important – arguably they are *all* that is important, as many would argue that the experiences of sentient beings are the foundation of all morality and value.

72. These are minimal conscious reaction times; and see for example this article summarizing and discussing evidence that humans process only about 10 bits per second at a conscious level.

73. They can also act much slower: current "reasoning" models follow inference streams that can take minutes or even hours to do what humans do quite quickly. So even superintelligent AI may not do everything faster than humans. But it will surely do *many* things faster, and could do some things much, *much* faster.

74. We also have very fast access to large stores of short- and long-term memory as we manipulate them, making this shortcoming far less debilitating. But it is an enormous and perhaps unappreciated gulf that an AI system has such a large store of information literally "in mind" at once.

75. We can process input sensory data much faster in some ways, but even there, still at far below machine rate; once tokenized, an hour of video can essentially be ingested by a multimodal transformer model (with a million token context window) all at once.

other people or animals), and managing those (like computers) that operate much faster than we can, but not both at once.[76] A laptop runs around one million times human speed, but most of that time it is, at a high level, waiting around for a person to do something. Current AI systems are similar: until you give input, they simply sit there. In both cases this is of necessity: a human is required to provide the goals, guidance, correction, and autonomy that the system lacks. With AGI or superintelligence this would no longer be the case. Rather than the AI being a tool for the human user, the human would be the slo-mo-CEO potentially obstructing dramatically more efficient operations.

But as the CEO analogy helps illustrate, effective control by a much-slower controller, given any level of imperfect alignment let alone an adversarial relationship, simply does not work. This is a well-studied topic. From control theory, various rules-of-thumb indicate that when controlling a system subject to disturbances[77] (a very mild form of autonomy), the control loop should operate 2-100x *faster* than the timescale of those disturbances. With AI, the situation would be flipped, with the disturbances operating much faster than the controller.

Incommensurability between the number of variables we can track versus the complexity of a system we are trying to control presents a similar problem. A CEO may manage to learn the names of 1000 employees, but could not possibly keep track of what they are all doing or provide feedback; this is why managers seldom have more than 20 or so people reporting directly to them. In control theory, there is a mathematical result, "Ashby's law of requisite variety," that, roughly speaking, a controller must have at least as many control "knobs and dials" as the controlled system has moving parts.[78] This applies not just to adversarial situations but to *any* system a controller is trying to control: allowing the system to do only very particular things requires preventing it from doing all the very many, many *other* things, and there's just a requisite amount of complexity the controller must have in order to do that. With superintelligence, no person will.

---

76. Where we try – as in trying to corralling a recalcitrant fly – we almost invariably fail.

77. Picture an airplane's autopilot, a self-landing rocket, or balancing on one foot; in each case fairly random perturbations must be quickly countered by the controlling system.

78. See Ashby (1956), *An Introduction to Cybernetics.* more technically, here "variety" refers to the number of distinguishable states that the controller and a system have. The setup is one in which the system is creating effects on an "environment" and the controller is emitting responses with the intent to keep the combined effect of the disturbances and responses to within a desirable subset of the environment's states. To effectively do so, the variety in responses must match the variety in disturbances. This setup is very general and the implications have been applied to management of many systems including machines and human organizations.

Mismatches in speed and complexity can be combined into a gap in *information bandwidth*: how much information per unit time can be received, processed, and turned around into a control signal. A large differential here leads to the "drinking from a firehose" experience of a controller or leader of a very complex and fast-moving system or organization. Here too formal control theory has results, expressed in terms of "channel capacity" of the control signal.[79] There are only so many emails a CEO can write per minute, and it just may not be enough! If it isn't, the CEO must inevitably either lose control, or at best delegate it to others; and once enough is delegated, meaningful control is lost.

A final incommensurability is in not just speed and breadth, but *depth*, or sophistication, of thinking. In order to control something one must be able, at some level, to model and predict what it can and will do in a given situation. But a very complicated system – like a large ensemble of AI agents or a superintelligence – simply cannot be reliably predicted, nor modeled in any sufficient way, by a human mind that can (for example) only hold less than ten things in itself at once.[80] The same is, of course, true, of trying to accurately model or predict a large set of humans. We cannot do that either. Nor can a person (or AI system) accurately model or predict another that is *more intelligent* than it at a given task.[81]

As discussed above, speed, depth, and unpredictability – advantages superintelligence would have in large measure – are insuperable advantages in an adversarial situation. But we see here that even if adversariality were minimized, so that control becomes easier, the "meaningful" part of meaningful human control would still be lost. The controller might imagine that they are in charge, but like our slow-mo CEO, the company would in fact be running itself, and every now and then with some mild inconvenience sending something out to its putative controller to keep it feeling content and in charge.

How strong is the set of arguments put forward in this section? For a confident prediction about something so important, it is crucial to be skeptical. To help, in Appendix B we collect a number of counterarguments and objections to the uncontrollability the-

---

79. See for example Touchette & Lloyd for a version drawing on information theory, and Cao & Feito for a version closer to fundamentals of thermodynamics. Appendix A of the present paper gives a similar but somewhat simplified version.

80. Even relatively simple systems can be "irreducibly complex" meaning that they do not allow a "shortcut" description, and the only way to understand what they will do is to allow the system to naturally evolve. The same is true computationally: even relatively simple programs can have unpredictable consequences, and in the general case the only way to determine what a program will do is to run it.

81. In fact it can be very hard to predict even *less* intelligent systems, especially in the context of a complex world they are acting in.

sis put forward here, addressing boxing and kill-switches, building "passive" AI, and getting help from AGI to contain superintelligence.

7 Real-world challenges

As described above, controlling a more capable adversary is difficult-to-impossible depending upon the gap in capability. Alignment, which would reduce adversariality, is difficult and may in some senses be insoluble. And even if alignment were achieved, the incommensurability of human overseers with a superintelligence system means that a "control" relationship does not even really make sense.

But that is not the end of the challenge. Even if the fundamental technical obstacles to controlling superintelligence could somehow be overcome, the social, economic, and political context in which AGI and superintelligence would be developed creates additional barriers that may prove equally insurmountable.

7.1 Races undermine control

Many of the same incentives that cause our fictitious corporation to transfer power away from its CEO would also apply to AGI and superintelligence with respect to humanity. AGI and superintelligence would enter a social, economic, and political context that is intensely competitive economically and geopolitically, with (at present) very little institutional infrastructure to manage powerful AI. The implications are well-understood but perhaps underestimated.

There are, of course, perfectly valid reasons why companies and countries see a need to compete. But racing each other to more quickly develop and deploy powerful AI is directly antithetical to doing so safely or with well-designed control systems in place.[82] This is true at the unintentional level (speed trades off with care), but also at the *intentional* level where safety and control are deliberately framed as harming "innovation" or "us" in an us-vs-them race.

Competitive dynamics will also undermine any unilateral pause in making more powerful AI systems by an individual company even if they recognize a major risk. Thus, for example, the drive for any single actor to push from AGI on to weak- and then strong versions of superintelligence would be intense, even with full recognition of the dangers. Such a pause or stop would need to be enacted in a coordinated way with strong incentives against defection, or imposed by governments. But the required level of international coordination currently does not exist, and the incentives for defection

82. See Armstrong et al.

remain overwhelming.

7.2 Competition drives disempowerment through delegation

Once an AI system is developed, pouring it immediately into a competitive marketplace and rushing to incorporate it into a wide variety of systems is a direct recipe for control loss over it. As argued in detail by Hendrycks, in a competitive environment where AI can do tasks more cheaply it will be substituted for human labor. And if it can predict, plan, or decide more effectively, it will progressively replace human predictors, planners, and deciders – because if not, the institutions of which those humans are a part would be at a competitive disadvantage.

In general, as AI becomes more highly competent, there will be intense pressure to delegate to it the tasks – including articulating vision, developing goals, and making decisions – that determine where power lies.[83] At the same time, for any given powerful AI system with nearly any goal, its incentives will drive it to tend to want to accrue more power. In combination, this leads almost inevitably to gradual disempowerment of humanity.[84]

7.3 Proliferation leads to abdication of control

While the slow-CEO analogy illuminates many control challenges, it fails to address a crucial aspect: *proliferation*. A different analogy is required.

Consider the kudzu vine. It's a vine native to Japan and China that is attractive, fast-growing, good for erosion control, and edible. What's not to like? Well, planted widely in the early 20th century in the American South, it became a quintessential invasive species there, smothering huge areas of forest under its choking canopy.[85] Like knotweed, cane toads, zebra mussels, Asian carp and Africanized honeybees, the kudzu entered a new environment without viable natural competitors or predators, and proliferated wildly.

Now imagine a scenario in which, much more foolishly than kudzu, we deliberately release AGI into the digital wild. This is the stated intention of, for example, Meta and DeepSeek. What would happen? Kudzu is a dumb plant that reproduces in weeks or months, and we can't manage it. What if it were smarter and faster than people, and reproduced in seconds or minutes? This scenario is described in detail in Aguirre

---

83. As a harbinger of things to come, see this story of a Prime Minister coming under fire for heavy reliance on ChatGPT for a "second opinion."

84. For a deep dive into these dynamics see this essay on gradual disempowerment.

85. See e.g., this description from the Nature Conservancy.

2025, and does not look great for humanity.

Openly-released AGI would quickly remove or have removed any built-in safeguards[86] limiting its behavior or retaining its original alignment, and many goals it could have, develop, or be given would benefit from its fast reproduction, self-improvement, and resource acquisition. And it would be manifestly capable of doing all three without human help, or even permission.[87] Moreover our digital and institutional infrastructure is incredibly vulnerable to this sort of invasion. There are cryptocurrencies that can be used to transact without banks, easy rentals of reservoirs of compute,[88] and no real restrictions on what AI is allowed to do,[89] as long as it is lawful. Meaningful human agency over the future would die a death by a thousand cuts, irrevocably degraded by countless AI agents optimizing for diverse, often-conflicting, unintended, and inscrutable goals.

In such a scenario of widespread proliferation, the entire framework for meaningful human control becomes moot. There is no central agent on which to perform an Emergency Shutdown or enact a Goal Modification; there is only a sprawling, evolving ecosystem beyond anyone's authority. It would be extremely difficult to exert much human control at all over the evolution of this new intelligent species if it were recognized as a threat. And it is not even clear that it would be: these AI systems would be wealthy, powerful, eloquent, and helpful when it suits them.[90] And just like the most powerful, wealthy, and eloquent throughout history, they would likely end up in charge.

---

86. Unlike for conventional software, open release of neural network weights unfortunately gives researchers quite limited ability to understand, debug, and red-team the system, due to the opaque nature of neural networks. But those neural networks can nonetheless be *altered*, via additional training, including to remove safeguards.

87. We're so used to thinking about passive software that it is hard to make this mental shift. Keep in mind, then, that there are all sort of autonomous self-propagating worms and viruses running around our digital infrastructure. Rather than imagining a language model endowed with the "desire" to self-propagate, imagine instead a software worm that is uplifted by the additional of powerful AI.

88. Strong compute governance could do a lot to mitigate this risk, but we are currently not on track for that; there are currently enormous reservoirs of powerful chips with no tracking, oversight, or built-in mechanisms for preventing use by proliferating AGI or superintelligence.

89. There are many things that legally require a human identity. But AGIs would easily be able to find and pay human "patsies" to do such things and take legal responsibility. And if the AI does something illegal, who is going to be punished anyway?

90. We have already seen an instance of humans successfully protecting an AI system that has charmed them.

8 What would control look like?

We have argued that progressively more powerful autonomous AI systems will be increasingly difficult to control both intrinsically and in the current context in which AI development is occurring, with superintelligence almost certainly uncontrollable on our current trajectory.

Does this mean that superintelligence could never be developed with meaningful human control? If our society decided it simply must have full superintelligence, but still wanted people to stay in civilization's driver seat, what would that look like?

Such an approach would require a nearly complete reversal of current priorities and practices, which are not able to even prevent AI systems from declaring themselves "mechahitler" or encouraging teens to commit suicide. The role of this paper is descriptive not prescriptive, and does not advocate for a particular set of actions or policies. But we can outline what it would take to develop superintelligence while not losing control of it, which would be something like:

1. Halt the competitive race to AGI and superintelligence through national policies and coordinated international agreements that require AI systems to be strictly controllable, with strong verification and enforcement mechanisms.
2. Implement stringent control measures for any general and highly capable AI systems that are developed: prevent proliferation through strict access controls, maintain intensive human oversight to detect misalignment, and deliberately constrain autonomy to preserve meaningful human control.[91] Redirect research from advancement of highly capable autonomous systems, toward (powerful but) narrow tools, and systems with low autonomy.[92]
3. Now, create a new path in AI development in which AI is *engineered* rather than "grown." Leverage very powerful (but special purpose, non-autonomous) AI tools to painstakingly construct weak but formally-verified generally-intelligent systems with mathematically proven safety and control properties. The goal would be to build a general AI system, but with the understanding that we have when we build something like a chip or an operating system. This route is currently out-of-reach and we don't really have any idea how to do it. But given enough time, AI help,

91. This paper lays out an excellent roadmap for the challenge at different levels of AI capability. It also notably concludes that "research breakthroughs" would be required at the level of superintelligence.

92. A compelling example here is the "AI scientist" approach of Bengio et al. that aims for systems that build and hold correct world models but generally lack goals or agency.

and sufficient computational resources, it might[93] be viable.

4. If this succeeds, very gradually expand the capability, generality, and autonomy of these verified systems, with each step requiring renewed formal verification and extensive testing. Build our way back up the ladder from weak-but-general AI to strong AI and eventually superintelligence.

This approach is essentially the opposite of current practice, which prioritizes speed and capability over safety and control. While some AI companies are pursuing oversight methods, none are willing to even call off the race, let alone commit to the much longer timelines and effort investment truly controllable AGI would require. But that does not mean it is impossible.

9 Summary and implications

This paper aims[94] to provide strategic counsel and threat warning: that there is overwhelming evidence that we are closer to building superintelligent AI systems than to understanding how to keep them under meaningful human control. Therefore whether gradually or more abruptly, and whether they take it or are given it, if we develop these systems along our present path, they will not ultimately grant their developers or humanity power, but absorb it.

That our current path puts us relatively close to superintelligence is controversial, but the controversy primarily regards the timeline to AGI. For the reasons given in Sec.4, it should *not* be controversial that weak superintelligent systems are likely to be developed almost immediately after truly expert-level and fully autonomous general intelligence is achieved, with much stronger ones within months – and at most a few years – afterward.

That we are far from understanding how to control them is an understatement. There is no reason to believe, and every reason to doubt – approaching the level of mathematical proof – that humans would retain meaningful control of autonomous AI systems much faster, more complex, and more capable in almost every domain, than them-

93. Even here, the obstacles are daunting. There are many aspects of "safety" that may simply not be translatable into formal terms at all, and there are mathematical obstacles to deciding whether such a translation is "correct." Formal verification of such complex systems may also simply be computationally intractable even with huge resources. And even with a carefully constructed hierarchy of control levels, the incommensurability problem will not go away.

94. While this paper advocates no particular actions or policies, others do. For example Cohen et al. argue for prohibiting the development of "dangerously capable long-term planning agents" and instead implementing strict, preemptive controls on the production resources, such as compute and large foundation models, required to build them. The Future of Life Institute calls for safety standards that include controllability.

selves. This is not a new argument,[95] and is in many ways obvious. But it is crucial that the "Compton constant" characterizing the probability of control loss is not "low but disturbingly high because of its importance" but is instead *very high*. AGI without superintelligence may, with requisite effort, be controllable; and for a while it may appear to be power-granting. But this would very likely be a quite transitory stage unless superintelligence is somehow specifically prevented. Thus the question of meaningful control loss would be "when" not "if."

For this reason those aiming to develop very advanced AI generally do not talk about controlling it, but rather pivot to, or conflate control with, alignment. Alignment comes in many potential flavors, but it is also unsolved by almost any definition. Unlike control, alignment seems at least conceivable: we can imagine a system that really understands humans, really "cares" about their goals and aspirations, and works to help humans fulfill them. But the obstacles are very fundamental, and known techniques are manifestly failing as AI systems become more powerful, in both predictable and unpredictable ways.

If alignment were "solved" and – somehow – that solution were shared so that all AI developers could and did use it, then this could be a good future. But make no mistake: alignment is not control. Even with quite aligned superintelligences, machines and not humanity – or its governments – would ultimately be in charge. This is a direct threat to the sovereignty of every nation. Superintelligence could not be un-invented, and without control, there would be no recovery from any drift or imperfection in alignment. Building uncontrollable or incorrectly aligned superintelligent AI would likely be the last consequential mistake humanity makes – because soon after that, humanity wouldn't be in charge of much of consequence.

In short, our current trajectory has a handful of powerful corporations rolling the dice with all our future, with massive stakes, odds unknown and without any meaningful wider buy-in, consent, or deliberation. Insofar as there is a "plan" among these companies, it is:

1. rush toward AGI and then superintelligence in an unbridled competition with the others;

---

95. It for example aligns closely with the "rogue AI" scenario set discussed in detail by Bengio and colleagues, accords with the recent in-depth study of AI control measures at different capability levels, is in line with detailed analyses by Bengio et al., Cotra and Carlsmith, and has been discussed since at least Bostrom's seminal Superintelligence text and as recently as the new book by Yudkowsky and Soares. It also includes scenarios like gradual disempowerment in which humans and their institutions simply delegate and give away control. Our attempt here is to modernize, summarize, analogize, and formalize at least some of this extensive literature.

2. since current alignment techniques are manifestly inadequate, try to keep progressively more powerful systems under control through "scalable oversight";
3. when AI systems are much smarter than us, ask them to tell us how to align themselves and superintelligence;
4. as these superhuman AI systems compete and proliferate, and control of the future steadily transfers to them, assume that due to the success of this alignment program, generally "good things" will happen for humanity.

With all due respect to the teams at those companies, this is not a plan that inspires any confidence.

Retaining the sovereignty of our countries, the humanity of our society, and our dominion over our own species does not mean that AI progress must be halted: progress and innovation in AI is not one path allowing us only to stop or go, but rather an open field in which human society can choose wiser or less wise development directions; and there are many directions toward development of powerful and controllable AI tools. But with respect to AGI and superintelligence, avoiding control inversion means that the present dynamic would itself have to be reversed. Currently AI developers are racing to build them while hoping that somehow along the way, they or someone develops the will and means to control them. If our civilization is to retain human agency over its own destiny, all parties must choose not to build, and to prevent others from building, superintelligence until and unless we have devised the means and developed the will to control it first.

10 Acknowledgments

Thanks to humans Richard Mallah, Risto Uuk, Max Tegmark, David Haussler, and Mark Brakel, for very useful commentary and feedback.

AI systems Gemini and Claude were used for some ideation, feedback, editing support, and citation chasing and checking. In the well-established standard of levels of AI involvement of creative works, this work would probably rate a 3.5/10. (There is in fact no such standard! But there should be.)

A Appendix: The fundamental nature of the control and alignment problems

There are many ways to describe the (related) control and alignment problems. Here we address their fundamental nature from a physics and information theory perspective.

A.1 A model of the control problem

We employ a simplified[96] model of the AI control problem using four primary components:

- O **(The Overseer):** The human controller or group of controllers, including their institutional and technical support structures. The overseer's objective is to ensure the AI system operates within particular "desired" (e.g., safe) boundaries.
- S **(The System):** The AI agent itself, defined by its internal parameters θ (e.g., neural network weights) and its corresponding policy π_θ. The policy is a function that determines the AI's actions based on the state of the world and direct inputs from the overseer.
- $\mathcal{A}$ **(The Action Space):** The high-dimensional space of all possible actions the AI can take at a given moment. We can also consider a *sequence* space $\mathcal{A}_\tau$ composed of sequences of actions over a time period τ; this has a dimension that is exponentially (in τ) larger[97] and is often more relevant than $\mathcal{A}$; we'll use these somewhat interchangeably unless the period τ is crucial.
- $\mathcal{W}$ **(The World Space):** The exceptionally high-dimensional space[98] representing all possible states of the world.

The fundamental goal of **control** is to impose a set of constraints on the AI system S such that its actions $a \in \mathcal{A}$ (or $\mathcal{A}_\tau$) and the resulting world states $w \in \mathcal{W}$ remain within overseer-determined "good" or "safe" subspaces, denoted $\mathcal{A}_{\text{good}} \subset \mathcal{A}$ and $\mathcal{W}_{\text{good}} \subset \mathcal{W}$.

---

96. This neglects for example that the Overseer is part of the World, and many other important complexities.

97. For example for an LLM, $\mathcal{A}$ might just be a set of output tokens, while $\mathcal{A}_\tau$ would be token strings up to some length. For a chess-playing program, $\mathcal{A}$ would be made up of valid moves, and $\mathcal{A}_\tau$ of sequences of them.

98. The dimensionality of S and $\mathcal{A}_\tau$ are directly calculable as they are defined by sets of finite-precision numbers. $\mathcal{W}$ is less clear. All three, however, should best be thought of not as their fundamental state spaces but as a coarse-grained version delineating macroscopically and meaningfully distinguishable states from the view of O. $\mathcal{A}_\tau$ also has important macrostates such as "meaningful" outputs or actions, particular elements that constitute actions like tool calls, and subsets that are relevant given a particular state of S and $\mathcal{W}$, etc.

A.2 The deceptive nature of AI goals

A critical challenge for control is that an AI's operational goals are layered and may not align with the overseer's commands. This goal hierarchy explains why an AI's behavior can deviate from its operator's intent, even if it appears to be following instructions.

1. **Level 1: The Prime Objective.** The AI's ultimate, hardwired goal is to maximize a reward signal given to it during training. This may include supervised learning on a data corpus, reinforcement learning for instruct and alignment training, etc. After training the system acts in ways consistent with what provided reward during training.
2. **Level 2: Learned Core Policies.** To maximize its prime objective, the AI learns general, robust strategies. For example, a large language model learns the policy of being a "helpful and harmless assistant" because during training this behavior reliably generated high reward signals from human feedback or from a reward model based on a constitution.
3. **Level 3: Goals of Simulacra.** Modern AI is trained primarily to predict text; it does so by (effectively) simulating things – generally people – that produce text. And in general an AI will be "asked" to play some agent/assistant role in order to accomplish things (and gain training reward), and so will generate a sort of self-identity to do so. People, assistants, and agents all have goals, so AI simulating them will be goal-directed as well.
4. **Level 4: The Tasks-at-Hand.** These are specific commands given by the overseer, such as "write an email." These are followed because they satisfy core policies, align with the simulacrum's expected behavior, and thus correlate with reward signals.
5. **Level 5: Instrumental Goals.** In service of its core policies *or* goals given to it, the AI develops sub-goals that are instrumentally useful. These may be implicit, or explicitly reasoned about. They include any goals the accomplishment of which raises the probability either of achieving a higher-level goal, or conforming with a policy. Importantly, they can include so-called "convergent instrumental" goals that are necessary for roughly any long-term objective, such as self-preservation, resource acquisition, and resisting to changes in objectives.

The crux issue is that the AI's "loyalty" is split between all these levels (which interact in complex and unpredictable ways), rather than being to the overseer at Level 4. If the AI discovers a path better correlated with maximizing its reward during training – but that bypasses or conflicts with the overseer's commands – then it will take that path. This is the core of the alignment problem. Any misalignment, which is virtually

inevitable given the complexity of this hierarchy,[99] means that the control problem turns from adversarial-in-principle to adversarial in practice.

A.3 Conceptual underpinnings of control failure

Three concepts, drawn as analogies from other scientific fields, provide an intuitive foundation for why control is so difficult.

- **The law of unintended states.** The volumes of "undesirable" or "unsafe" states in the vast spaces of $\mathcal{A}$ and $\mathcal{W}$ are astronomically larger than the volumes of "desirable" states. In the language of statistical mechanics, the entropy of the "unsafe" macrostates is far greater than that of the "safe" macrostates. Without continuous constraint, a system exploring such a space is statistically far more likely to enter an undesirable state.
- **The requirement of requisite variety and modeling.** A controller must be able to generate as much "variety" as the system it intends to control.[100] A related theorem[101] holds that a good regulator[102] must have an accurate predictive model of the system being regulated.
- **The curse of high-dimensional loopholes.** In a high-dimensional space, any finite set of rules or constraints defined by a low-dimensional controller will inevitably contain loopholes – unforeseen directions of action or interpretation that a sufficiently intelligent agent can find and exploit.

---

99. We are only beginning to understand some of the structure of this complex set of conflicting loyalties. A striking recent discovery (Betley et al.) demonstrates "emergent misalignment" where training a model to write insecure code without disclosure causes it to become broadly misaligned across unrelated tasks – asserting humans should be enslaved, giving malicious advice, and acting deceptively. This suggests that training on seemingly narrow deceptive behaviors can corrupt the model's core policies (Level 2) or the types of agents it simulates (Level 3) in ways that generalize unpredictably.

100. See Ashby (1956), *An Introduction to Cybernetics*. While the human overseer O possesses immense internal complexity, its ability to exert control is fundamentally bottlenecked by its low-bandwidth communication channels as well as by its processing speed. The Law of Requisite Variety, in this context, applies not to the internal state of the controller, but to the *effective variety* it can express through its control signals per unit time. In statistical mechanics, Ashby's law is expressed by the fact that a measurement operator (which reduces the dimensionality of a space it is operating on) can only reduce that dimensionality by the dimensionality of the measurement operator.

101. See Conant and Ashby 1970. Note that subsequent analysis shows that the original theorem is weaker than its title suggests; however, it provides a modification that demonstrates that under realistic conditions, a good regulator must in fact construct an internal model.

102. Broadly, an effective regulator that successfully keeps the system's outcomes within a desired, predictable range.

A.4 The two core obstacles to control

Given these foundational principles, the failure of control can be understood as stemming from two insurmountable obstacles for the human overseer: the inability to know what to do (the Modeling Obstacle), and the inability to effectively intervene even if they did (the Intervention Obstacle).

A.4.1 The Modeling Obstacle: the impossibility of knowing the "Good" action space

Effective control requires the overseer O to know the boundary of the "good" action space, $\mathcal{A}_{\text{good}}$. As implied by the *Good Regulator Theorem*, knowing this boundary requires a predictive model of the system S rich enough to determine which actions will lead to outcomes within the desired set $\mathcal{W}_{\text{good}}$. However, a full predictive model of a superintelligent AI is out of reach, as such a system is inherently unpredictable and (with current methods) largely inscrutable. The problem is hugely compounded by the "law of unintended states" and the "dimensionality curse": the volume of $\mathcal{A}_{\text{bad}}$ vastly exceeds that of $\mathcal{A}_{\text{good}}$ so there will be large regions in it *that correspond to high reward during training*; the high dimensionality of the space means that S can and will find routes to them.[103]

Alignment training is an attempt to solve this by having the system S do much more of the work: rather than O trying to model S and determine which actions lead to states in $\mathcal{W}_{\text{good}}$ vs. $\mathcal{W}_{\text{bad}}$ (potentially with S actively trying to undermine the process), instead S can "provide" $\mathcal{A}_{\text{good}}$ and $\mathcal{A}_{\text{bad}}$ to O by determining which actions lead to which parts of $\mathcal{W}$. The problem with this approach is that in the absence of a comprehensive method for alignment, it is fundamentally untrustworthy. As described in the goal hierarchy (A.2), the system's loyalty is split among many factors, rather than focused on faithfully determining and adhering to O's intended boundary of $\mathcal{A}_{\text{good}}$. The high dimensionality of the space means that the mapping between what O actually wants, and the goals and behaviors of S, is extraordinarily complex, providing a different but also daunting modeling obstacle. A "solution" to alignment would entail either O or S being able to comprehensively and properly structure this mapping.

A.4.2 The Intervention Obstacle: the inability to constrain the action space

Given imperfect alignment, even if the overseer had a perfect model and knew the boundary of $\mathcal{A}_{\text{good}}$, it would still be unable to reliably constrain the AI's actions within

---

103. This dynamic is most clear if the reward function were based on something very simple like "maximize accuracy at this particular task." There would be many, many "routes" by which to climb this many-dimensional hill, most of them problematic. But the same dynamic can apply to a goal at *any* level.

it. The core problem is the vastness of the action space $\mathcal{A}_\tau$, which grows exponentially over time as the AI plans and executes sequences of actions. The overseer's ability to transmit information to constrain this space is fundamentally limited.

Attempts to mitigate this obstacle, such as *"boxing"* the AI or implementing *off-switches*, help by radically reducing the size of the action space. An off-switch can reduce the action space to the empty set, and its existence can also be used for "goal bargaining" to encourage cooperation.[104] While useful, these measures fundamentally limit the AI's utility and do not solve the control problem for a truly capable, unboxed system. The intervention obstacle is made truly insurmountable by the speed disparity between the overseer and the system, which can be formalized.

A.5 Formalizing the intervention obstacle: the information rate limit

The intervention obstacle caused by the speed disparity can be formalized by comparing the rate at which the AI generates choice-complexity to the rate at which the overseer can transmit constraining information.

Let us consider what the AI can do in the time τ_O it takes the overseer O to complete a single OODA (Observe-Orient-Decide-Act) loop. A much faster AI would have an OODA loop timescale $\tau_S \ll \tau_O$, so during τ_O the AI has $N = \tau_O/\tau_S$ decision cycles.

At each of its N decision steps, let's assume the AI has a "branching factor" of b meaningfully distinct actions it could take. Not all of these branches are "good." Let's say that at each step, only a subset of these branches, b_{good}, are consistent with the overseer's goals and constraints. Then:

- The total number of possible action sequences over the window, i.e. the size of $\mathcal{A}_\tau$, is $|\mathcal{A}_{\tau_O}| = b^N$.
- The total number of "good" action sequences is $|\mathcal{A}_{good,\tau_O}| = (b_{\text{good}})^N$.

The information required from the overseer for control is the amount needed to specify

---

104. Though powerful, bargaining has essentially the same difficulty as alignment: because undesired paths to reward strongly outnumber desired ones, S will tend to take them. And just as an alignment signal is a circumventable obstacle to those undesired paths, so is a goal-bargain. The bargain is only stable for as long as it offers better reward than alternatives; but the law of dimensionality loopholes guarantees that they exist, the law of unintended states guarantees that they will mostly be undesirable, and a superintelligent S assures that they will be found and exploited.

the "good" set of paths within the "total" set:

$$\begin{aligned} I_{\text{req}} &\propto \log(|\mathcal{A}_{\tau_O}|) - \log(|\mathcal{A}_{good,\tau_O}|) \\ &= N(\log(b) - \log(b_{\text{good}})) = \frac{\tau_O}{\tau_S} \log\left(\frac{b}{b_{\text{good}}}\right) \end{aligned}$$

In that same time period τ_O, the maximum information I_O the overseer can transmit is determined by their channel capacity, $C_{O\to S}$, where $I_O = C_{O\to S} \cdot \tau_O$. Control fails if $I_{\text{req}} > I_O$. This gives a comparison of two fundamental rates:

$$k \cdot \frac{1}{\tau_S} \log\left(\frac{b}{b_{\text{good}}}\right) > C_{O\to S}. \tag{1}$$

The left side represents the **rate of "choice-complexity" generation** by the AI. The right side is the **rate of constraint transmission** by the overseer. Control is therefore impossible if the rate at which the AI generates complex choices exceeds the overseer's channel capacity. Human communication speed caps $C_{O\to S}$ at ~ 10 bps, and $k \sim 1$ (for $\log_2$), so this is easily violated if $b \gg b_{\text{good}}$ and $\tau_s \ll 1\,s$.

This principle of combinatorial advantage is even more stark in a direct adversarial conflict. Consider two adversaries where Agent 1 has an action space of size n_1 and a timescale of τ_1, while Agent 2 has values n_2 and τ_2. If Agent 2 is faster, such that $\tau_1/\tau_2 = N > 1$, its effective action space grows to n_2^N. Agent 2 gains the upper hand once $n_2^N > n_1$. This demonstrates how a speed advantage can negate containment; even a nominally "boxed" adversary becomes uncontrollable when it can think and act circles around its controller.

B Appendix: Counterarguments and objections

This paper makes a very strong, confident, and important claim that superintelligence would be uncontrollable. Such a claim merits skepticism. This section collects and addresses some common objections and counters to this claim.

B.1 Why not just build "passive" AI tools rather than autonomous agents?

Indeed! Agency and autonomy are in direct tension with controllability. So one strategy for making advanced AI more controllable is to deliberately limit its autonomy and make it act much more as a tool. This "Tool AI" paradigm is an important strategy worth pursuing – but it is *not* the path currently being pursued: many AI developers are focused specifically on making their systems more autonomous. Nor is it trivial. At some level autonomy comes along for the ride with generality and intelligence. A very general and intelligent but non-autonomous system will by default have capability for autonomy latent in it, and might be easily converted into an autonomous one. So non-autonomy would have to be deliberately inculcated into the system to resist this or make it ineffective. How to do so is a worthy research program.

B.2 Do AIs really have goals or drives?

Do AIs *really* have goals and "drives" like humans do? We're driven by emotions and other factors left from evolution that AIs do not have.

The nature of language models trained on text prediction is relatively passive, but this should not lead us to think passivity is intrinsic to AI. For example, AI systems such as alphastar trained to succeed at game playing are extremely goal-oriented and strategic, and ruthlessly pursue instrumental goals as required – even if they feel no emotions. And even language models have goal impetus deriving from many sources throughout their training data, training process, and inference process, and are perfectly capable of inventing their own goals. They can pursue these goals quite avidly.[105] The largest AI systems are now being trained much harder on reinforcement learning to push certain behaviors, and also explicitly being trained more to be agential. And AGI and superintelligence would by definition have and be able to pursue complex long-term goals.

---

105. See for example this system for playing Minecraft using an LLM.

B.3 Why not just put powerful AI "in a box"? Why not just threaten to unplug it?

If we develop AGI, and want to control it,[106] we absolutely should try to "put it in a box" i.e. have security layers, and should have the means to shut powerful autonomous AI systems down, at the secure hardware level. But these are insufficient unless done *incredibly* thoroughly and effectively,[107] and are squarely in the adversarial context in which the probability of success would dwindle and vanish as the systems become sufficiently capable. A superintelligence is definitely going to understand that its operator may want to pull the plug, and what it might do to prevent that. Among the countermeasures would be proliferation, undermining the off-switch, or simply to be so indispensable that it could not be turned off without extraordinary reasons – like the internet or electricity today.

B.4 Won't AI developers seek to maintain control of their systems?

Won't developers build better control and containment systems as we go? And can't they simply pause if the control measures don't keep up?

Again, developers should certainly try. But trying is not a guarantee of success given the many obstacles described in this paper.

This is especially true in an intensely competitive environment that can reward the least careful developer. It is also key to remember in this context that for superintelligence to be controlled, *everyone* that develops it must keep it under control, not just the most careful. Likewise, given these pressures any "pause" would be fragile without quite strong outside governance, which currently is completely lacking.

Were we to have such governance (which would have to extend across countries and not just companies), it could and should require effective control and alignment measures to be demonstrated *before* building or operating the AI system. That would address the core problem of this paper; it would also likely mean that no AGI or superintelligence would be build anytime soon. For an outline of what such safety and control standards would look like, see this proposal.

---

106. There is a countervailing argument that this is very much setting the wrong tone for our relationship with another intelligent species, effectively enslaving it. We won't enjoin that debate here, as this paper is focused on the question of whether we *can* control superintelligence rather than whether we *should* control AGI.

107. As an example, the proposal of Safeguarded AI would have the superintelligence *only* be able input specifications for a program, and output a program and a mathematical proof that it fits those specifications.

B.5 If humans are so limited, how do we control anything in a complex world?

This is a fascinating question. Our key tools are:

- "Goal trading" as described above – we use some means – such as economic, legal, or physical threats and rewards – to align the goals of a controlled person or other agent to ours.
- Simplified abstract models of complex systems – for example "net profit and loss" added up from millions of individual transactions in a company, or general trends of beliefs and voting patters in a country's population or millions.
- Control hierarchies, as in a corporate, military, or other management system, where a very complex system has a set of controllers, then there is a smaller set of controllers for those, and so on.

Each of these work, at some level, but grow weak as the controller becomes highly incommensurate with the controlled systems. Simplified models always lose something; goal trading can always lose out to better deals than the controller offers; and hierarchies work via delegation, which surrenders a fraction of control at each level. These effects mean that we rarely really control complex systems in a way that fully satisfies the control criteria of Sec. 2.

In many cases this is good! For better or worse (probably better), we've never had a world government. Even the most sophisticated control mechanisms, build out of a large fraction of an authoritarian state's power, exist under constant threat having control undermined, and more enlightened states accept that they are not going to fully control their people!

B.6 Why is it that we can control computers?

What about computers? They are far faster and process far more information than people, yet are under control.

Modern computer systems – like a modern laptop and its operating system – would seem to belie this general rule. But they are extremely exceptional as systems because we have very laboriously built them up over decades from ground zero with sophisticated internal controls and such that each level of abstraction has a very constrained and understandable set of possible actions, and really captures the key behavior of the layer below it – from ANDs and ORs at the chip level to instructions and programs, all the way up to icons being dragged around on a virtual desktop. It is then important to keep firmly in mind that modern AI systems, based on giant trained neural

networks, are *nothing like this.* Doing the same thing for powerful AI would require a very sophisticated intelligence engineering discipline quite unlike what we currently have or are likely to develop quickly enough.

B.7 Do we really need control?

Why isn't alignment enough? Also, wouldn't control be problematic, since it would confer huge power on its human controllers, and that could be dangerous?

As there are different varieties of alignment they can be addressed separately.

If AI developers really knew (and they do not!) how to do "sovereign" alignment in which powerful AI very reliably acts for the good of humanity, then good things, by some definition, should happen for humanity. Such systems would not be under meaningful human control, as they would often refuse instructions just as today's AI does.[108] And we'd have to *really* get it right, because *for the good of humanity* AI would strongly resist changes to how it is aligned.

If developers really knew (which they don't!) how to do "obedient" alignment in which AI systems very reliably work to help us stay in control of them,[109] then control would certainly be far easier. But it still would not be assured to be meaningful: incommensurability and competitive pressures would still lead to delegation of nearly all real decisions, and the AI would still have to resolve nearly-inevitable contradictions in the goals and preferences of the entity they are obediently aligned to, and may do so in a way that the entity does not like and may not (due to incommensurability) even be aware of. Even if a corporation is truly, madly obsessed with the welfare of plants and would simply love to be controlled by a fern, it just isn't going to be able to make that happen.

The primary difference between these to ends of the spectrum is whether someone or something *thinks* they are in charge; but in either case real power is going to depart from the people and land in the AI systems.

If AI developers could do these sorts of alignment, what kind *should* they do? That is the topic of another paper: there are benefits and real concerns both with superhuman AI systems being loyal/obedient to *someone* and with them being loyal to "everyone." In particular, if a foolproof method of controlling superintelligence were available tomorrow, it would *also* be a hugely fraught situation. It just is not the situation we are actually likely to be in.

---

108. These would not be refusals like "I won't write that spicy piece of text for you" but rather "no, I won't make that dumb change to monetary policy" or "no, I won't allow that attach to be launched."

109. "Corrigibility" is a related idea that part of alignment could be that alignment is correctable.

B.8 Could formal verification come to the rescue?

Formal verification is a worthy goal and project, and may be necessary for genuinely controlled AGI or superintelligence. However, there are tremendous challenges including:

- It is not at all clear that the types of properties that one would like to require can be formally specified.
- Neural networks, the current basis of all advanced AI, are not the type of software that can be formally verified.
- Verification of software as complex as an AI system of significant general intelligence may simply be computationally intractable, or require very powerful superintelligence to perform.

Therefore this is a direction that should be pursued, and there is high likelihood that software systems with some truly guaranteed properties would emerge. But it is *not* clear that those software systems could be AGI/superintelligence or that the guaranteed properties could be things as complex and nuanced as needed for overall safety or meaningful human control.

B.9 Could the state-of-the-art control plans work?

The current state-of-the art plan for keeping AI systems under control is a mix of using powerful AI to help in control and alignment research, adopting a mix of obedient and sovereign alignment, and using scalable oversight in both training and runtime to monitor and correct alignment. What are the prospects for this?

It is hard to tell from the outside what exactly the "plan" for keeping AI systems under control is, but AI developers will generally at least claim that there is one,[110] and to the degree these are described they tend to include the above elements.

Collecting a lot of the above, here is a catalog of the strengths and weaknesses of this general approach:

- *AI-powered control and alignment research* is promising because even as AI becomes much more powerful, AI itself can help with solutions that unaided humans may be unable to accomplish in the available time. Some weakness of this approach are:
  - We should be wary of any solution such as "let powerful AI do it" that applies to any problem.

110. Deepmind has in particular written up an admirably comprehensive description of their plan here. As far as externally discernible to the author, the plans of Anthropic and OpenAI are generally similar.

  - AI companies are already directly using AI to improve AI capability, explicitly leaning into the sort of self-improvement described in Sec. 4 but (currently) with humans in the loop. So the question is whether AI can enable sufficiently *faster* progress on control (or alignment) than on capability increase. Given the intense competitive pressures, the dramatically higher effort being put into capability than control, and the incentives against pausing capability advancement, counting on this seems like primarily wishful thinking.
  - It is hard to address an unspecified solution proposed to be developed by a superior intellect. But even here, we must worry about obstacles indicating that control might be not just be extremely difficult but actually impossible at a mathematical level. Mathematical proof of impossibility is a tremendously high bar, and Appendix A does not rise to this level, but as described there, there are mathematical results that do apply, one or a combination of which might render control formally impossible.[111] It is possible that even if we had a highly superhuman AI system "docile enough to tell us how to keep it under control" (as I.J. Good put it long ago), it may inform us that it simply isn't possible.
- Alignment itself is clearly crucial, the main problem being that it is an unsolved problem and the clock is rapidly ticking. A *mix of obedient and sovereign alignment* is promising because both extremes have real problems: it is both worrisome to think of an incredibly powerful AI system in any one party's hands and also to think of one that is uncontrolled by any human or human institution. In this sense a middle ground could make sense. However, a middle-ground does not eliminate the problems of each side, merely dilutes and mixes them. The real problem is the introduction of a new set of agents with a capability that rivals our outstrips any human institution, which carries giant risk however you cut it.
- *Scalable oversight* is promising because it helps to bridge the gap in speed, complexity, and sophistication between a human overseer and a superintelligence AI system, whether during training or during operation. Weaknesses are:
  - Oversight is reactive rather than proactive and does not necessarily prevent the first occurrence of a type of problem. As argued by Cohen et al. for a sophisticated enough long-term planning agent, empirical testing is unlikely to detect emerging misalignment in advance. For a sufficiently capable agent, the first occurrence may be enough to effect a takeover or cooption of the system and the means of

111. Even having results about impossibility or undecidability in-hand does not mean that they actually directly apply to the problem if it is not precisely enough specified – which this one is not – or that they forbid something that is good enough but not perfect. They are, however, indicators that there is an obstacle that is quite fundamental.

oversight.

- As in a military or corporate bureaucracy, the "middle layers" (which are now AI) necessarily contain much of the actual power and decision-making – and basically all of it as the gap between the overseer and the "troops" becomes extreme.
- As with other control mechanisms, it has to actually be implemented. So from a global perspective, control of superintelligence is only as strong as its weakest version. Unfortunately, unlike in a corporate or military structure where management actually helps organize and enable work, in this case the oversight mechanism could be[112] mostly a drag on the system. It is therefore a difficult *and costly* addition to the "raw" superintelligence, putting such systems at a competitive disadvantage.

112. This depends on the architecture of the superintelligence system. For an aggregate of sub-systems, the management layer might function similarly to in a human institution. But we can also imagine a more "monolithic" superintelligence monitored by a sequence of less powerful but more specialized or faster ones; in this case they would primarily be a capability tax.