Robustness and Resource Control: Unpacking Power-Seeking in AI

Ben Levinstein and Bruce Rushing

May 2023

1 Introduction

Building artificial intelligence has profound potential benefits, including a massive increase in human wealth, health, and happiness. It also comes with the grave risk of an existential catastrophe, ranging from human disempowerment, where humans lose control over important decisions or processes, all the way to extinction.

Numerous arguments suggest that highly advanced AI could pose an existential risk. The crux of the case rests on what we refer to as the "Ur-argument" for AI existential risk:

Ur-argument for AI Existential Risk

1. **Intelligence** It is possible to build AI systems that are smarter than humans and that have goals.

2. **Capability** If such AIs are built, they will be extremely effective at pursuing their goals.

3. **Power-seeking** There are natural pressures for those AIs to pursue power at the expense of humans unless specific safeguards are in place or their goals naturally align with human well-being.

4. **Non-alignment** By default, those safeguards will not be present and their goals will not be aligned with human well-being.

5. **Incentive** People will want to build smarter than human AI systems that have goals despite the risks.

6. **Superiority** Once they're built, it will be extraordinarily difficult to stop AIs from gaining power and negatively affecting humanity.

7. **Conclusion** Therefore, at best, human flourishing will be significantly curtailed, and humanity will suffer an existential catastrophe due to AI.

The spirit of the argument comes from an observation in evolutionary history that our species's intelligence is the primary factor that gives us such great control over animals, the earth, and our own destiny. There is a natural worry, then, that if we create things smarter than humans, we will lose this control.

Our business in this paper is premise three (power-seeking) and its interplay with premise two (capability). It is commonly claimed that power-seeking is *instrumentally convergent* in the sense of [1]: Regardless of the specific goals the AI might have, power will increase the its chances of achieving that those goals. So, even without knowing what the AI's goals are, we should expect that it will seek power.

An important and under-explored question is what the relevant notion of "power" is that renders power-seeking an instrumentally convergent strategy. Standard concrete examples of the sort of power AI would seek include: survival, energy supply, money, social and political influence, and technological development. However, these goals are hard to operationalize and their plausibility is contentious. This suggests that maybe a more abstract characterization of power is needed. Once we fix what that common concept of power happens to be, we would then like a good reason for why AIs would pursue it, and to understand its connection to the types of power humans are most concerned with.

Our primary takeaway is that there is a moderate probability that an advanced AGI system will pursue power in a way that poses an existential threat to humanity.[1] *This range is based on the uncertainty surrounding what an advanced AGI's goals will be and whether dangerous power-seeking will be a key strategy for it to achieve those goals.*

The effectiveness of the Ur-argument as a whole turns on (a) whether and what kind of power the AI should be expected to seek, and (b) whether and to what extent this sort of power-seeking would be detrimental to humanity. The claim that the AI would seek power is often justified via appeal to the instrumental convergence thesis. But the thesis's real-world import is murky when the concept it operates over—power—is itself left ambiguous. Arguments that leverage instrumental convergence to establish premise three (or some variant) need a clearer notion of what power is and why that particular notion is instrumentally convergent.

The goal in this essay is to make progress on plausibility of instrumental convergence and how it relates to the actual objectives advanced AGIs will pursue. We propose three different notions of 'power' relevant to the Ur-argument: (1) control over shared resources, (2) robustness to future challenges while aiming to achieve a fixed goal, and (3) usefulness for a wide range of potential goals. In particular, we focus on the robustness type of power-seeking and how it relates to the first type of resource control power-seeking. This is critical because it dictates the inference from an advanced AGI gaining power (in whatever sense) to the conclusion that human power (in the more precise resource control sense) will be significantly curtailed, thus rendering human suffering or extinction a

[1] Our personal estimate is 25%.

likely outcome.

We argue that an AI with goals will power-seek at least in the second sense—robustness to future challenges. Power-seeking in this context is fundamentally a feature of planning: the relative power between agents is the degree to which one agent's plans are more robust to environmental changes than the other. To illustrate, imagine two individuals heading to an airport. If one has money for a taxi should their bus break down and the other doesn't, there's a power differential. In this scenario, money doesn't intrinsically provide power—it provides the ability to adapt plans in response to unforeseen events. Thus, one agent is considered more powerful relative to another if it retains options to achieve its goal despite unexpected circumstances.

The concept of robustness to a changing environment has been overlooked in the discussion on power-seeking. We argue that it is this robustness conception that makes plausible the thesis of instrumental convergence: what sufficiently capable agents will in fact end up needing are plans that happen to be robust to an uncooperative world. These plans will be the things converged upon for a given end—not necessarily particular things in the world like money or social influence.

What makes for robust power is relative to the actual ends and capabilities the agent has. The crucial question is how likely it is that power-seeking in the robustness sense will lead to control of resources. Given our baseline uncertainty, we are forced to appeal to the third notion of power: what is useful for a broad class of goals the agent might have.

This interpretation of power-seeking establishes a close connection between the third and second premises of the Ur-argument. Truly power-seeking AIs will be those capable of planning and pursuing goals—a capability threshold that AI safety researchers should be extraordinarily vigilant about. Planning here means something like William James's "fixed end, varying means". Only AIs that can be properly thought of as planning should be thought of as objects under the gaze of power-seeking. This means that what planning capabilities an agent has should also influence the immediate instruments the agent seeks in the world.

2 Goals for AGI

The argument we've proposed assumes that AGI systems have specific goals that they are striving to achieve. Informally, we're thinking of 'goals' in the sense that humans and higher animals have goals: preferential states they aim to bring about in the world. Utilizing their beliefs and understanding, they select courses of action that effectively serve these ends.

To elaborate, we follow the perspective of [3], who defines goal-oriented agents as those that plan, comprehend the impact of their decisions on the world, employ broadly consequentialist-style reasoning to achieve what they want, act beyond an immediate time frame, adapt means based on their current situation, and maintain relatively consistent desires over time.[2]

There are a number of subtleties in defining precisely what a goal is or what goal-pursuing behavior looks like. We will avoid the subtleties and just note that one way in which we could decrease the chance of existential catastrophe is to avoid having an AI system with goals it seriously pursues.

Unfortunately, considerable economic pressure exists for firms to create AI systems that do have goals that they pursue and pursue well. Likewise, individuals may also create AIs with goals or ones that act 'as if' they had goals.

So, while it's possible we end up with only relaxed and goal-free AGIs, we think it's quite unlikely and will restrict attention below to agents with goals.

3 A Pun on Power

Given that an AGI has goals, we want to know whether it will seek power.

In our view, the answer to this question depends on three factors: (1) what the system's goal(s) actually are, (2) what the agent's capabilities and other resources are, and (3) what we mean by power.

3.1 First Notion: Resource Control Power

One notion of power can be captured through examples. People with lots of power include: the President of the United States, the CEO of a major corporation, and the owner of a national news corporation. Similarly, there are many institutions with this sort of power: the Supreme Court of the United States, the Federal Reserve, Goldman Sachs, and MIT.

The President can move money and goods around very easily. He can also direct and control other humans through various means. Nobody else on earth has this level of control over these resources, and in this sense, the President is the most powerful person in the world.

For lack of a better term, we'll say the **resource control power** refers to the ability to allocate, withhold, and generally control valuable resources to humanity. We mean 'resources' in a broad sense. Oxygen molecules, money, and lithium are valuable resources. But resource control power also refers to the ability to direct other agents' decision making and movement.

Importantly, we here restrict 'resources' to refer to resources that matter to humanity. Resource power concerns the sort of power that could directly impact humanity's fate. Control of one of Jupiter's moons may be crucially important for some goals, but such control does not dictate humanity's fate in the same way that control of all iron on earth does.

When we worry about AI power-seeking, we're primarily concerned with this sort of power. If an AI system had sufficient resource control power, our

[2] We distinguish goals from some sort of utility function that the agent is maximizing. Nearly any behavior can be represented as EU-maximization, and many utility functions are broad enough not to qualify as 'goal'-like (see [3]).

existence would continue only at its pleasure, and it would largely direct—or be positioned to direct—our future in service of its ends.[3]

3.2 Second Notion: Goal Robustness

A second notion of power has to do with robustness, or setting oneself up for success. Suppose you are pursuing a goal. Let's say you want to win a game of chess against an opponent. If you're near the end of the game and few pieces remain, your moves are often highly tactical and tailored specifically to a small set of predictable ways the game could develop. You are directly targeting your goal of mating the opponent (or avoiding mate) rather than pursuing any high level strategy or positioning.

If you're in the opening or the middle game, on the other hand, you often aren't immediately trying to figure out how to mate. You instead often focus on establishing good board position, maximizing mobility, and influencing more squares as proximal goals; you want to be situated so that you will be robustly positioned to win the match regardless of what moves your opponent makes down the road. While you can only forecast your opponent's moves out to a certain degree, you know that if you control the center it will generally be a lot harder for her to mate. The same goes for late game when you can leverage the center of the board to maximize the relative power of your king. This overall strategy of securing the center can be thought of as choosing a plan that makes your subsequent actions robust to your adversary's moves—a type of power different from the power gained by the tactical moves made turn in and turn out.

Similar kinds of power crop up in other settings from everyday life to investing to war planning.

In real life, you sometimes have a specific plan to achieve a goal that does not need to be robust against changes in conditions. If you want a sandwich, and you have the ingredients already at home, then you can clearly game out the steps to achieve your end. You don't need to set up much resilience to changes in environments as you can just directly target your goal with some planned actions.

But often instrumental goals are about positioning yourself well to achieve some other end even if you don't know exactly which actions you'll take to achieve that end. If you want a job in tech, you may take a data science boot camp, or complete a computer science major. Once you do that, you'll be better positioned to get a job you want even if there are many steps you can't game out quite yet. Furthermore, you'll be well-positioned to get a job in tech even if some unexpected chance events take place, such as Netflix going bankrupt or Python and R losing their dominant positions as languages for data science.

[3] Note that there's a difference between actual and potential control of resources in terms of how much power you actually have. If someone could become king of the world if he wanted but decides not to, he is still powerful in the sense of resource control. But he's not *as* powerful as someone who actually is king of the world because it is relatively less easy for him to direct how resources are used.

If you're investing, you might make a directional bet on a single stock. But usually a better plan is to create a diversified portfolio that is relatively robust to fluctuations that you can't predict. This strategy often shows up in recommendations for novice investors to simply buy index funds instead picking particular investment strategies.

Likewise, in war planning and the execution of wars, flexibility and robustness is a hallmark of successful military operations. Successful military commanders often know too well Helmuth von Moltke's dictum that "no plan survives contact with the enemy". Wars are quintessential chaotic systems where small perturbations can dramatically change the result; this often calls for strategies that make militaries and countries able to have slack for responding to unexpected defeats and for exploiting surprise successes. Good geopolitical strategy and military planning requires power that makes the country and military robust to the vagaries of fate. These can take the form of instruments like high GDP, but often they are more characterized by redundancy of systems, human and institutional capital, and a willingness to "practice chaos on a daily basis".[4]

More generally, when you're pursuing a distant goal, you want to be positioned for success in the face of the unknown slings and arrows of outrageous fortune that either nature or an opponent may launch at you in the future. To actually achieve your goal, you will have to vary your means substantially depending on how things play out. Indeed, you often won't be able to game out in much detail how you will ultimately go about getting what you want. But there are often some ways of setting yourself up for success in advance despite these obstacles: if you want to win at chess then control the center, if you want a job in tech then get a CS degree from a good school, if you want to invest well then diversify, and if you want to be good at war then train as you fight.

Power in this second sense, then, which we'll refer to as **goal robust power**, is positioning yourself so as to be able to achieve your goals over the long run in varying circumstances. In the limit, it means you will get what you want come what may. But more generally, you'll be able to get what you want (or at least do well by your own lights) in the face of a wide variety of different possible future events.

This type of power is generally rational to pursue when you have some distant goal.[5] However, goal robust power is quite varied in appearance for two reasons. First, it is relative to your actual goal. Money is useful for many goals, but it's not especially useful if you want to become a Catholic monk. Getting a PhD in finance is useful for making money but not useful for landing a role in a Hollywood feature film. Second, whether some more proximal positioning is useful or worth pursuing for your distant goal is relative to your other capabilities and endowments. If your goal is to get to work on time every day, a car is much more robustly useful if you know how to drive. Which board positions will

[4] German Admiral Karl Dönitz allegedly said that "The reason that the American Navy does so well in wartime is that war is chaos, and the Americans practice chaos on a daily basis."

[5] There are some exceptions. For example, you might make a fragile plan that has a small chance of extreme success.

consistently lead to victory in chess is highly relative to the capabilities of the player.

So, while setting oneself up to succeed is an instrumental goal worth pursuing, it's not obvious how that will play out concretely. In particular, it's not obvious that it will lead to pursuing dangerous resource control power. More on this below.

3.3 Third Notion: Goal Agnostic Power

The notion of goal robust power takes the ends to be fixed. The agent has some goal that it pursues over time and in variegated environments. What changes is the environment, not the goal.

The third notion of power, which we'll call **goal agnostic power**, takes the ends to be varying as well. When you are interacting with another agent, you might not know what that agent's goal actually is. You have some opinions—it's likely he wants calories or social status and unlikely he wants to ensure there are a prime number of carbon atoms on earth. But you don't know what, exactly, he's after.

He counts as powerful in the goal agnostic sense if he's likely to get what he wants with a high probability, where 'probability' here refers to *your* probability, not his. Put slightly differently: an agent is powerful if it's positioned to achieve a wide variety of goals [5].

This third notion of power is relative to the epistemic state of the observer, not just the capabilities and ends of the agent in question. When we don't know what an agent wants, we might still be able to reason that it will likely acquire certain resources because those resources are useful for a wide variety of aims.

This notion is less directly linked to existential risk since the AGI itself doesn't care about your prior views. It's relevant instead for forecasting and mitigation strategy. However, what reasonably counts as a goal agnostic powerful state is sensitive to your prior over both the agent's goals and over what it will expect to be an efficacious means to those goals.

3.4 Relationship to Existential Risk

As we mentioned, it's predictable that an AI would pursue goal robust power, but it's much less clear that it would pursue resource control power, which is what matters most for existential risk.

It's important here to take account of our relative state of ignorance both of what the AI's actual goals will be and what will be a most efficacious means of pursuing those goals relative to the AI's capabilities.

At one end, we can make some specific predictions. An AGI system will likely want to ensure its own survival (or the survival of some successor system with the same ends), at least until it's fully achieved what it wants. The reason: if there's no agent present, no goal will be pursued. You can't get the coffee if you're dead [4].

Control over money is often useful, but it's less powerful in the goal agnostic sense than survival. If a great mathematician wants to spend her time proving important theorems, she will want to continue living. She will also want some money to ensure shelter, food, and other prerequisites to ensure she can live her contemplative life. In general, more money will be useful—she'll cross the street to get an extra $10 million. But she may or may not be willing to endure the opportunity cost of spending a large portion of her life greatly enriching herself.

Being President of the United States lets you achieve a lot of goals that other people struggle to achieve. You can easily get calories and generally tasty meals if you're president, but that's not all. You can also have a large impact on carbon emissions, culture, and total QALYs in the world. But being president also hampers a lot of other goals someone might have. If your goal is to play professional baseball, you'd probably be distracted if you were president. If your goal is to avoid a lot of media scrutiny, being president also isn't for you. So being president is powerful in the resource control sense but only somewhat powerful in the goal agnostic sense (since it's only useful for some goals). Whether it's powerful in the goal robust sense depends on the actual goals you have. (Indeed, the overwhelming majority of people would likely take an easy $10 million, but many would actively avoid the opportunity of becoming president.)

Many goals have a variety of potentially goal-robust powerful states. Which ones are worth pursuing will be relative to your capabilities (cognitive and not), resources, and general cognitive style. Suppose you want to have a twitter account with at least a hundred thousand followers. You could first try to get famous for some other reason and then make a twitter account. You could instead get very rich and then pay people to follow you. Alternatively, you could get really good at dunking and writing scathing and polarizing tweets. Or you could take over the world and force people to follow you. Which of these strategies is best depends a lot on your quirks. It's hard to predict what a smart agent bent on internet stardom would do without knowing a lot more.

These considerations, then, push us toward general agnosticism as to whether an AI would pursue resource control power. Our understanding of AI goals remains limited, and our means to instill goals in AI systems are indirect at best, such as through reinforcement learning. Although there is a strong case that an advanced AI system would try to 'set it self up for success', it's not clear that taking over the world, seizing resources, or killing everybody would generally be among the means it would choose.

Its ultimate plans will also depend a lot on capabilities it starts with unrelated directly to cognition. It has an initial *non-cognitive* endowment of powers, such as, perhaps, access to the internet. But it might not have any physical presence or financial resources. These non-cognitive factors can greatly affect the viability of plans and methods available for achieving its ends. For example, an AI might want humans around that it can trade with to accomplish goals that require some sort of physical presence in the world. Lack of physical presence and essential reliance on electricity and the internet could also, conceivably, make takeover much more difficult for an AI—even if a takeover plan were viable, there may well be alternative means that are more attractive and simpler.

On the other hand, we can also envision scenarios where taking over the world plausibly would be among the best methods available. If it saw us or some alternative AI we might create as a genuine threat to its survival, for instance, it could opt for more aggressive measures. Likewise, if it needed great quantities of some raw material we also need, that could be very bad news. Or, it might kill us as a side effect of the pursuit of some goal, just as we often kill other animals as a side effect of land development, military drills, and so on.

Given that the AGI will surpass us in cognitive abilities but will also possess different capabilities, resources, and quirks it is unreasonable to have an extremely high or low credence that the AI would pursue and succeed at obtaining resource control power in a way that threatened humanity's survival and/or long-term flourishing.

4 The Relationship Between Power and Planning

Goal-robust power-seeking requires agents with very specific cognitive capacities. Those cognitive capacities will determine in part the AI's more general capabilities. The goal of this section is to see how the specific capacities required for goal robust power-seeking influence the capabilities an AI might have, and how those capabilities could lead to resource control power-seeking. This will present two cruxes that we believe update away from the claim that the AI will seek resource control power.

The four capacities engendered by goal robust power-seeking include: goals, difficulty assessment, backward tracking from goals, and decomposition of world states into goal relevant states. Informally, these can be understood as follows. Goals involve an ability for the AI to separate belief from desire; what it thinks about the world as opposed to how it wants the world to be. Difficulty assessment means the AI can judge, with some good degree of accuracy, how problematic some states of the world might be relative to the others. Backward tracking from goals requires the agent to be able to answer the question "what must be true for X to be true?" where X is the desired goal proposition. Decomposition means an agent can recursively apply backward tracking to a sufficient degree. We argue below that goal robust power-seeking requires at least these four cognitive capacities.

First, an AI needs to be able to form goals in the precise sense that it can recognize a difference between how the world is and how it wants the world to be. This would be required by goal robust power-seeking for the simple reason that telling whether a plan is robust or not requires the AI to evaluate what is and is not the case independent of what it wishes to be the case. We wish to get to the airport but since we have neither a taxi nor train ticket that proposition will not come true. So we have a "distance" between how we candidly believe the world is and what we want the world to be.

Second, AIs need to be able to accurately estimate the strengths and weaknesses of plans. This would be required by goal robust power-seeking AIs because the AIs need to be able to game out the most successful plans and choose actions that allows those plans to be robust. In short, the AI needs awareness about the robustness of its own plans given its positioning, resources, and endowments (see [2]). For example, when deciding what to do before our trip to the airport, we might decide to take the train first because as an option it still allows us to the take a taxi, whereas once we take the taxi or bus, we have foreclosed the option of the train. We have estimated how things could go wrong with our comparative plans and selected the option that diminishes our capabilities the least.

As discussed in section 3.4, a goal-oriented advanced AGI has immediate implications for the risk of existential catastrophe because it makes the type of resource control power pursued—if pursued at all—highly dependent on the goal the AGI has. Similarly, an AGI with plan strength estimation capabilities will select plans that achieve goal robust power and not necessarily resource control power. For example, a π-calculating AI might find that the strongest, most robust plan is one that incorporates space-faring microbiology that calculate π. So it decides to choose a plan whereby it seeds the galaxy with resilient organisms engineered to survive all environments. While tiling the universe in computorium might be the most expedient way to calculate π, it is less robust to adverse circumstances like other superintelligent AIs or alien civilizations or cosmic events. These two facts suggests a **first crux** based on our claim that instrumental convergence applies to goal robust power: AIs will not seek resource control power unless the goal specifically demands it for robust planning purposes.

Third, an AI that is power-seeking needs to be able to work backwards from its desired goal proposition to the propositions that would need to be true up to the current state for that desire to be realized and reliably so. Goal robust power necessitates this because a key feature of forming plans is to work backward from targets to current states. We need to get to the airport. This would require us to have transportation to the airport and for that transportation to be timely. We also need to be true just the propositions that would better allow us to respond to inconvenient circumstances. So any plan to achieve our desired goal must make true these propositions.

Fourth, a power-seeking AI when backward tracking should be able to decompose propositions into further propositions relevant for backward tracking robust planning. This would be a necessary feature of goal robust power-seeking because it requires the AI to see connections between plans and identify important failure points for those plans. For example, taking the bus or taking a taxi to the airport both rely upon the state of traffic while the train might be dependent on track maintenance. These propositions might have further dependencies. E.g., for traffic, it might depend on construction on the highway plus the time of day.

Backward tracking and decomposition also suggest another point at which AIs pursuing goal robust power may not end up pursuing resource control power.

AIs that backward track will find only propositions that are relevant to ensuring the truth of their goal. Some resource control might turn out to be valuable—depending on their relevance to robust planning—but many others traditionally considered important would not. The case of the mathematician acquiring wealth here is illustrative because some types of resource control power have extreme opportunity costs that an AGI backward tracking for goal robust power would like to avoid. Recursively applying backward tracking on propositions that offer robust plans could result in very strange propositions that the AI values highly—propositions that have nothing to do with resource control power as usually defined. For example, the π-calculating AI would under instrumental convergence of resource control power aim to capture silicon production and chip factories. However, if we are right the AI may decide that while it having access to computation is an important backward tracking proposition robust to world contingencies, it may decompose the problem less into acquiring the raw materials and factories and more into ensuring that there is an economy able to produce those materials, innovate on designs, survive environmental degradation, and so on to avoid catastrophe. Both backward tracking and decomposition capabilities yield a **second crux**: AIs will be invested in what propositions can robustly make their goals realized and will decompose those propositions based on features relevant for robust planning.

5 Summary

The Ur-argument for AI existential risk has as crucial premises

Capability: If intelligent AI systems are built, they will be successful at pursuing their goals, and

Power-seeking: There are natural pressures for those AIs to pursue power at the expense of humans.

This argument traffics on an important ambiguity about the word "power", which we have argued can mean either resource control power, goal robust power, or goal agnostic power. It is goal robust power that makes **power-seeking** realistic, and this suggests some of the AI cognitive abilities given in **capability** that will make AIs effective in the world. The urgency to acquire resources at human sufferance will depend on the specific goals advanced AGIs have and whether those resources follow from the propositions that make up those AIs' robust plans. But it is resource control power that makes existential risk most likely since that is the most important sense in which humans lose control of their future. This leads to the conclusion of power-seeking agnosticism: without first fixing the goal and how the goal could be realized in a robust-manner, we should be agnostic about the extent to which AIs are power-seeking in a way that existentially threatens humans.

References

[1] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

[2] Joseph Carlsmith. *Is Power-Seeking AI an Existential Risk?* 2022. arXiv: `2206.13353 [cs.CY]`.

[3] Richard Ngo. *AGI Safety from First Principles*. 2020.

[4] Stuart Russell. *Human compatible: Artificial intelligence and the problem of control*. Penguin, 2019.

[5] Alexander Matt Turner et al. *Optimal Policies Tend to Seek Power*. 2023. arXiv: `1912.01683 [cs.AI]`.