WITHOUT FUNDAMENTAL ADVANCES, MISALIGNMENT AND CATASTROPHE ARE THE DEFAULT OUTCOMES OF TRAINING POWERFUL AI

Peter Barnett[*] **Jeremy Gillen**[*]

Machine Intelligence Research Institute
{peter,jeremy.gillen}@intelligence.org

ABSTRACT

This report focuses on AI systems that are able to automate large-scale scientific projects and argues that such systems, if created using current methods, will pursue unintended and catastrophic goals. We discuss important properties of hard tasks like novel science; success in these tasks is defined in terms of outcomes and requires overcoming novel and diverse obstacles. We argue that AIs capable of such tasks will be well described as taking actions in order to achieve goals. We further argue that behavioral training, the current method of creating AIs, is an incredibly imprecise way to specify goals and so AIs created this way are likely to pursue misaligned goals. Due to the scale of the projects these AIs will be doing, it will be very difficult to safely get useful work out of misaligned AIs. We ultimately argue that the default outcome for creating such AIs will likely be catastrophic.

Contents

[*] Equal contribution.

We would like to thank Lisa Thiergart for managing this project. Many thanks to Thomas Kwa, James Lucassen, Thomas Larsen, Nate Soares, Joshua Clymer, Ryan Greenblatt, Buck Shlegeris, John Wentworth, Joe Carlsmith, Ajeya Cotra, Richard Ngo, Daniel Kokotajlo, Oliver Habryka, Tsvi Benson-Tilsen for helpful discussion and feedback. All views and mistakes are very much our own.

Summary

In this report we argue that AI systems capable of large scale scientific research will likely pursue unwanted goals and this will lead to catastrophic outcomes. We argue this is the default outcome, even with significant countermeasures, given the current trajectory of AI development.

In Section 1 we discuss the tasks which are the focus of this report. We are specifically focusing on AIs which are capable of dramatically speeding up large-scale novel science; on the scale of the Manhattan Project or curing cancer. This type of task requires a lot of work, and will require the AI to overcome many novel and diverse obstacles.

In Section 2 we argue that an AI which is capable of doing hard, novel science will be approximately consequentialist; that is, its behavior will be well described as taking actions in order to achieve an outcome. This is because the task has to be specified in terms of outcomes, and the AI needs to be robust to new obstacles in order to achieve these outcomes.

In Section 3 we argue that novel science will necessarily require the AI to learn new things, both facts and skills. This means that an AI's capabilities will change over time which is a source of dangerous distribution shifts.

In Section 4 we further argue that training methods based on external behavior, which is how AI systems are currently created, are an extremely imprecise way to specify the goals we want an AI to ultimately pursue. This is because there are many degrees of freedom in goal specification that aren't pinned down by behavior. AIs created this way will, by default, pursue unintended goals.

In Section 5 we discuss why expect oversight and control of powerful AIs to be difficult. It will be difficult to safely get useful work out of misaligned AIs while ensuring they don't take unwanted actions, and therefore we don't expect AI-assisted research to be both safe and much faster than current research.

Finally, in Section 6 we discuss the consequences of building a powerful AI with improperly specified goals. Such an AI could likely escape containment measures given realistic levels of security, and then pursue outcomes in the world that would be catastrophic for humans. It seems very unlikely that these outcomes would be compatible with human empowerment or survival.

Introduction

We expect future AI systems will be able to automate scientific and technological progress. Importantly, these systems will be doing novel science, developing new theories, discovering new knowledge. We expect these systems will be doing large scale projects, the kinds of projects that would require many people multiple years to complete. This is the scale of the task that we will be considering in this report.

We expect such tasks to require robust goal-directed behavior, and we expect agents capable of such behavior to also be difficult to supervise and contain while retaining their usefulness. If such goal-directed agents are created by behavioral training, this training won't be sufficient to precisely specify its terminal goals. This means that such an AI is unlikely to have goals that align with our goals. When people attempt to use misaligned powerful AIs for large scale useful tasks, these AIs are likely to escape and pursue their own goals. Behavioral training is the default way to create AIs, and so without fundamental advances (which we don't speculate on) powerful AIs will likely be dangerously misaligned.

We argue that AIs must reach a certain level of goal-directedness and general capability in order to do the tasks we are considering, and that this is sufficient to cause catastrophe if the AI is misaligned. This does not require the AI to be "superintelligent" and "optimally goal-directed" in all situations.

One reason for pessimism about getting powerful AIs to safely do hard novel science is the scale of this task. We expect that the AIs will be doing the equivalent of years of human labor, and there will be many opportunities to take irreversible, catastrophic actions. We very likely cannot account for all of the unknown unknowns the AI will encounter in its research.

Our aim for this report is to explain the central difficulty where we expect AIs that are capable of novel science to also be dangerously misaligned, and why we expect this problem not to be solved if we continue along the default trajectory. We focus on the difficult tasks which we assume the AIs will be capable of. The specific capabilities required for these tasks, combined with the imprecision of behavioral training, lead us to expect misaligned goal-directed behavior, even though current, 2024, AI systems do not seem dangerously capable and goal-directed.

We aim to lay out a mostly complete argument for our mainline beliefs about catastrophic outcomes caused by AI, starting from the tasks we assume they are capable of.[1] This risk model can be thought of as an "inner alignment" failure, where even if we knew what to tell the AI to do, we are unable to make it safely do that.

We hope that this can be helpful for informing AI safety research directions; ideally focusing research effort on approaches which are still valid for extremely powerful AI systems or on foundational research to avoid the problems inherent to behavioral training. Many of the claims in this report are not of the form "It is impossible in principle to get an AI with this desired property" but rather "Given how we made the AI and that it is capable of certain things, it is unlikely to have this desired property". It is not impossible to create an aligned or safely constrained powerful AI, however this is unlikely if the AI is created using current methods.

This report considers AIs that are capable of hard, novel science, and we expect AI alignment research to be in this category (see Section 1.4). Therefore, difficulties and dangers in this report should be relevant considerations for groups attempting to use AIs to do AI alignment research. The most prominent example of this that we know of is the OpenAI Superalignment team [1], although teams

[1] We are *aiming* for this document to be as complete as possible on our mainline reasoning, but we will inevitably miss many important considerations.

at Anthropic and Google Deepmind are probably pursuing similar strategies. This also includes any group aiming to radically speed up scientific progress using AIs.

0.1 How to read this report

This report is intended to explain the authors' beliefs and the main reasons for these beliefs. Each section is stating a thesis which depends on a number of assumptions. The sections are ordered such that assumptions of later sections are argued for in earlier sections. Unfortunately, this means that at any point the reader doesn't agree with an argument we make part way through, the rest of the document won't feel fully justified. In this case, the reader should treat each section as a separate argument for a conditional statement: If we believe the assumptions[2], then the section is our mainline reasoning for believing the conclusion.

The arguments that are "most central" vary a lot between people. We are trying to provide the arguments that would be most likely to change our beliefs if we discovered they were wrong. There is a heavy bias toward arguments that are most salient to us. These arguments are usually salient because they have been important in disagreements with people we regularly talk to.

This report represents the views of the authors, not the views of MIRI or other researchers at MIRI.

We are fairly confident about most of the individual claims in this document, however this doesn't mean we are confident that all arguments and assumptions are correct. We think it is likely that there are sections that contain mistakes, and it's plausible that such mistakes dramatically change our conclusions. However, we still think it is important to communicate our overall beliefs due to their implications for research prioritization and other planning.

0.2 Related work

We will compare this work with related work on similar thread models, see [2] for a more thorough review of threat models.

"Risks from Learned Optimization" describes why we would expect AI systems to be goal-directed, and how behavioral training is not sufficient to precisely specify these goals [3]. Our work focuses on difficult tasks and argues that systems capable of these will have to be goal directed, we also discuss how behavioral training can lead to *unstable* goals, not just improperly specified goals.

"Is Power-Seeking AI an Existential Risk?" lays out a conjunctive argument for expecting existential risk from powerful AI [4]. We attempt to focus more on why we expect trained AIs to be misaligned and goal-directed, and given this how such an AI could evade our countermeasures.

"AGI Ruin: A List of Lethalities" lays out many reasons why one would expect powerful AI to lead to catastrophe [5]. Our work attempts to lay out a more cohesive and expanded picture, and we focus more on how a powerful misaligned AI could evade human oversight and control.

"A central AI alignment problem: capabilities generalization, and the sharp left turn" describes a specific problem related to powerful AI, where once an AI is capable it is revealed that it was not aligned, and the AI then pursues some unintended goal [6]. We make a similar argument, but attempt to connect our story more to the specifics of AI development and the tasks powerful AIs are expected to do.

"How likely is Deceptive Alignment?" argues, by considering path dependence and inductive bias in neural network training, that AI systems are likely to be *deceptively aligned*; faking alignment during training and later pursuing misaligned goals [7]. Our work does not focus on the inductive bias or path dependence of training AIs, but rather argues that AIs which are capable and created using behavioral training will likely be misaligned.

Various arguments made in related work are based on "counting arguments"; arguments of the form "there are many goals that are consistent with an AI's behavior during training, so we should not expect the AI to pursue the specific goal we want" [3, 5, 8]. We make similar arguments and compile multiple "degrees of freedom" in an AI's goal specification.

[2] Sometimes a strong version of the assumptions.

1 Useful tasks, like novel science, are hard

We start with an assumption that, when developed, powerful AI systems are capable of large-scale, difficult tasks, and people will try to use them for such tasks.[3] We will discuss specific properties of these tasks, and later in this report we will argue that by default AIs capable of tasks with these properties will be misaligned and dangerous.

These tasks will have large search spaces, requiring many actions and where success is only achieved by a relatively very small set of action sequences. They will also be outcome-oriented, novel, and diverse. Throughout this report we refer to tasks with these properties as *hard* tasks, and will expand on the specifics of these properties in this section. By *outcome-oriented* we mean that we have to specify the task by the outcome it achieves, because we don't know the sequence of actions to achieve it. By *novel*, we mean that such tasks would require the AI to do things they weren't initially trained to do.[4] By *diverse*, we mean that there are a wide range of skills required to successfully do the task. These properties will be expanded on in this section.[5]

These properties define a kind of task that we will use as the defining capability of powerful AI. In this report we will refer to these as *hard* tasks, this specifically refers to these tasks which have large search spaces, are outcome-oriented, and contain novel and diverse subproblems. We are using *hard* to refer to tasks that have these properties, and not just any task that a human would find difficult.

1.1 Impactful novel science takes a lot of work

We will be focusing on AI systems which are capable of doing novel science,[6] the scale of the work we are imagining is curing cancer or some similarly large scientific endeavor. Tasks like this would necessarily require systems to be operating over long time scales. A central intuition pump here is the Manhattan Project, which took three years to design and produce the atomic bomb. Other examples could include:

- The Human Genome Project
- The Apollo Program
- Proving Fermat's Last Theorem
- The development of modern microbiology, starting at from the first observations of microbes
- The development of quantum mechanics

We don't know how much work would be realistically needed to cure cancer and it may be easier than the above examples, but we can lower-bound this based on the amount of work that humans have put in to date. This task so far has already taken thousands of humans decades of work.

Some of the work may be parallelized, but there is a lot that cannot be; discoveries that are made along the way will be necessary for later steps, and will change the course of the research. New methods will likely need to be developed, and old methods applied in new ways. This will necessarily require large amounts of serial work.

[3] At the end of this section, we'll discuss why the task of AI alignment appears to be in this category. This also applies to other tasks that we have considered that seem sufficient to prevent dangerous AIs from being developed for a significant length of time.

[4] This isn't precisely defined, but we are referring to the intuitive scale that roughly goes from copying training data to deriving never-before-seen Go strategies, and further to inventing special relativity from scratch.

[5] There are tasks that are economically useful and don't have all these characteristics, for example writing relatively simple code. AIs can easily be both capable of these tasks and safe, and therefore can be economically useful. However, these AIs are not capable of dramatically speeding up research, because they are lacking skills such as: designing novel experiments; working out novel theories based on data; seeing patterns in data and then drawing appropriate conclusions; inventing conceptual frames for specific problems.

[6] By "doing" we mean fully obsoleting or at least dramatically speeding up humans at a task. An AI speeding up human research 30× would be doing 97% of the work and very likely be able to do effectively all of the work. The 30× speed up is inspired by [9].

1.2 Outcome-oriented tasks

A task like doing novel science (for example, curing cancer) is outcome-oriented. That is, the task is defined by achieving some specific outcome in the future, and we are able to describe this outcome in advance. Tasks that we define by describing the necessary actions are less outcome-oriented. For example, instructing someone to bake a cake by giving them exact instructions is less outcome-oriented; while telling someone to bake a cake and having them work out the steps to get there by themselves is more outcome-oriented.

When we tell an AI to achieve an outcome which we don't know how to achieve, this is inherently outcome-oriented. By this we mean that we don't currently know the procedure to achieve the desired outcome and so the AI has to work out the procedure; we don't mean that this is a task that humans could never achieve with substantial effort. This applies to "curing cancer" because we can't specify the exact sequence of actions that lead to success, we can only specify success criteria. For example, the success criteria could be "have a medical intervention which can remove all the cancer cells in a patient's body, while leaving the other cells intact and the patient otherwise unharmed". Novel science in general has this property, because this inherently involves discovering new things and using those discoveries, hence we cannot describe all of the actions in advance.

1.3 Novelty and diversity of obstacles

Obstacles in hard tasks like novel science are not predictable in advance, and often dissimilar to obstacles previously encountered. Writing a program could involve inventing or appropriating a new data structure which works with the particular constraints of the specific problem. In science, it can be extremely valuable to sort through messy debates containing arguments about which data is relevant and real and combine this information with context, to decide which experiments are most valuable to do next. In adversarial settings, an agent needs to deal with other agents which seek out and play strategies that it has the least experience with.[7]

Sometimes a researcher is missing necessary skills or knowledge and has to work around that somehow, either by gaining the skills or knowledge, or finding a route that doesn't require these. Different resources can be a bottleneck at different times, leading to new and different constraints. Available resources may change which can necessitate a different approach.[8]

Above examples of diverse obstacles share the property that each new obstacle may require a novel strategy to address. The defining property of a successful novel strategy is that it still leads to the desired outcome. Defining the strategy by other means, like by describing a particular sequence of steps, or a particular heuristic to locally optimize, becomes harder as the diversity and novelty of obstacles increases. Defining strategies without reference to goals often requires predicting and coming up with solutions to obstacles before running into the particular obstacles that need solutions.

Among different goals and environments, there are differences in the level of novelty and diversity of obstacles. It looks like hard research that generates genuinely new and useful insights will require facing repeatedly very novel and very diverse obstacles. Hard novel science, such as curing cancer, seems extremely likely to fall into this category.

[7] More examples of diverse obstacles, for different tasks:

- A designer trying to build a fusion rocket has to interface with the steel producer and run tests to see if the steel is good, and renegotiate or switch suppliers depending on the results.
- An advanced coding assistant will sometimes have to debug weird hardware glitches, things outside of normal coding. E.g. To build extremely efficient algorithms, you need to be experimenting with different ways to exploit caching, especially in GPUs.
- Math AI has to communicate extremely complicated results to humans. For this purpose it might need to develop a new formalism for thinking about problems, or define new types of mathematical object that allow people to draw on intuitions (i.e. similar to Feynman diagrams).

[8] Examples of "resources" could include: time, money, energy, skills, knowledge, other people's labor, food, memory, writing space, or shared context with coworkers.

1.4 AI alignment research is hard

We think that the task of doing useful AI alignment research is a task which is hard and outcome-oriented, and will require novel and diverse skills. By "do useful AI alignment research" we mean that the AI system would be able to perform or speed up human research output by 30x for the research task of "build a more powerful AI which can do novel science faster than humans, which we are confident will do tasks it is directed to do, while choosing not to irreversibly disempower humanity".

This task, especially getting sufficient confidence in safety, we think will require:

- Novel mathematical work, developing mathematical models that have not been explored before.[9] This work will likely need to build on itself, requiring inventing and understanding one new mathematical model, and then developing another based on that. This kind of work seems necessary for building an AI that is very stable under learning and reflection (i.e. acts in well-understood ways while learning a lot, especially when it comes to ontological crises and self-improvement).
- Empirical work to validate the theory. This will likely require coding to run large empirical tests of various components of the designed AI system, as well as testing bounds and approximations.
- Engineering work, building and iterating on the (hopefully) safe AI system. This will likely require large scale software engineering, similar to the scale that is required to build large foundation models. We expect this to be much more complicated than current foundation models, because the system probably needs to be something more complicated than an end-to-end trained black box.

While we expect AI alignment research to be hard and require these skills, the overall argument in this report does not rely on this assumption. We will be arguing that AI systems trained using current methods (if they are capable of hard, novel science) will be misaligned[10] and too dangerous to use.[11] We separately believe that solving AI alignment will require hard, novel science (but won't argue for this further, in this report).

1.4.1 Stopping misaligned AI deployment seems to require powerful aligned AI

Many well-resourced companies and governments are motivated to build powerful AI. Any approach to AI safety has to deal with the problem of surviving when a less competent and safety-conscious actor could create and accidentally release a misaligned AI. We don't know of any approaches to this that don't involve a safe, aligned powerful AI.[12] We would welcome being wrong, and would be excited about concrete strategies that would make the world existentially secure without needing to solve alignment and build a powerful AI.

1.5 Conclusion

For the rest of the report, we frequently refer to *hard* tasks as tasks that have extremely large search spaces (relative to the set of solutions), are outcome-oriented, and contain a lot of novelty and diversity. In this section we have argued that large-scale novel science is *hard* in this sense. We are aware that the difficulty of tasks falls on a spectrum, when we say "*hard* task" we are referring to tasks of similar scale and diversity to the Manhattan Project or other tasks discussed in this section.

We expect useful AI alignment research to be hard in this way.

We argue in Sections 2, 3 and, 4 that capability to do this sort of task implies that we can model powerful AI as approximately "consequentialist", and there is difficulty in specifying the goals that such an AI will be pursuing.

[9] E.g. research such as heuristic arguments [10], logical induction [11], infra-bayesianism [12].

[10] See Section 4, with dependencies in Sections 2 and 3.

[11] See Section 5.

[12] Such as an AI defense system that detects and shuts down misaligned AI.

2 Being capable of hard tasks implies approximate consequentialism

In the previous section, we argued that an AI which is able to do hard, novel science must be capable of tasks which are:

- Easy to describe in terms of future outcomes
- Difficult to describe precisely in other ways (due to novelty and diversity of obstacles)

In this section we will argue that this implies AIs with such capabilities will be capable of approximately consequentialist behavior. By this we mean that the AI will be capable of taking actions to achieve specific outcomes, this will be further defined in this section.

2.1 Future outcomes as goals

Many useful goals are simply specified by future outcomes. By future outcomes, we mean roughly a property or fact about the future world. In particular, the important part of future outcomes is that they don't have strong dependence on near term actions. We argue that AIs doing outcome-oriented tasks, as described in Section 1, will be well described as behaving as if they are robustly pursuing future outcomes (i.e. their behavior will be consequentialist). A "goal" is a representation inside the AI, while a "future outcome" refers to the state of the real world.

Future outcomes are usually the only simple way to describe success on a task, without knowing in advance how it should be done. An AI that is robustly capable of completing the task must have some means of recognizing actions that lead to success. It can't have memorized specific strategies for overcoming all obstacles, because there could be arbitrarily many of these. Therefore, it must be capable of calculating strategies on-the-fly, as it learns about new obstacles. We will expand on this argument later in this section.

2.1.1 Formalizing consequentialist goals

We can construct a simple definition of consequentialist goals based on the idea that success can be evaluated entirely by looking at future states, rather than the path that led there. This definition can be represented as a causal graph, as in Figure 2.1.1.

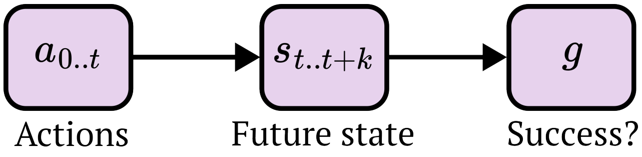

Figure 1: A causal graph, where g is a random variable that represents whether the goal was achieved.

Here the actions affect the future state which affects success at achieving the goal; the only way the actions affect the goal is via the state.

If $a_{0..t}$ are early actions, $s_{t..t+k}$ are a short sequence of states in the future[13] and g is some internal representation of the outcome-goal in an AI, then a consequentialist goal has a property like:

$$MI(a_{0..t}; g|s_{t..t+k}) \leq \epsilon \tag{1}$$

Where the mutual information $MI(\cdot)$ could be defined with respect to an outside observer approximately modeling the AI. Given that we know what the future state is, knowing what the actions were doesn't tell us any more about whether the goal was achieved (note this is an identical statement to the causal diagram in Figure 2.1.1).

2.1.2 Non-consequentialist constraints

There are also ways to specify behavior that don't look like taking actions to achieve future outcomes, and instead are more like constraints on actions and intermediate states. Examples of such constraints:

[13] Not necessarily entire world states, this definition works better if it's only the goal-relevant facts about the future world.

Always act “kindly”, or always follow a particular high-level procedure while completing a specific task, or always do a particular action upon particular observation. Humans also have plenty of similar shallow constraints; for example, disgust reactions, flinching, fear of heights. More deontological ethical prohibitions [13], like “don’t kill”, are also an example.

One could define such non-consequentialist goal specification c as a variable that is primarily dependent on short term behavior and state, and not strongly dependent on the outcomes in the future:

$$MI(s_{t..t+k}; c | a_{0..t}, s_{0..t}) \leq \epsilon \tag{2}$$

Here c is defined by modeling the agent as choosing actions that result in success according to c. Conditioning on the actions and states in the short term, the final states don’t give you additional information about whether the AI successfully followed its constraints (Figure 2.1.2).

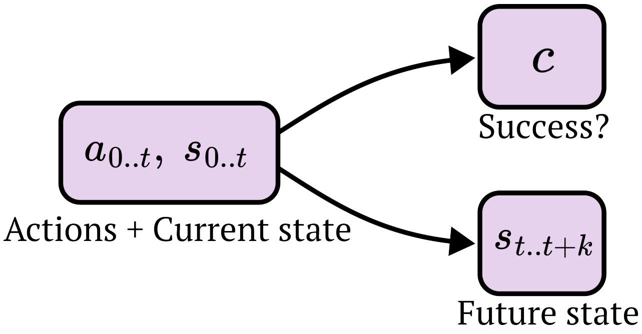

Figure 2: Non-consequentialist goals

2.1.3 Combined goals

Constraints and consequences can be combined to describe many problems (Figure 2.1.3). For example, winning chess might be described by defining the action space and legal moves, alongside a description of checkmate. The next section will argue that, while combined goals are a more accurate model for the behavior of powerful AI, the primary driver of intelligent problem-solving behavior will tend to be consequentialist goals.

The distinction between constraints and consequences is useful for describing why powerful trained systems are likely to be an existential risk by default. Specifically, the danger comes from a powerful trained AI pursuing a different consequentialist goal than the one that we intended, and is missing constraints that we intended it to have. The reasons why the precise goals and constraints are likely to be learned incorrectly are described in Section 4.

There exists much more complicated behavior that isn’t precisely captured by this simplified model of goals. Despite this, we think this description of goals is sufficient for describing the main reasons we expect misalignment in real agents.

We call an AI “misaligned” if its behavior is well described by goals and constraints that are different from the goals and constraints intended by its human creators.

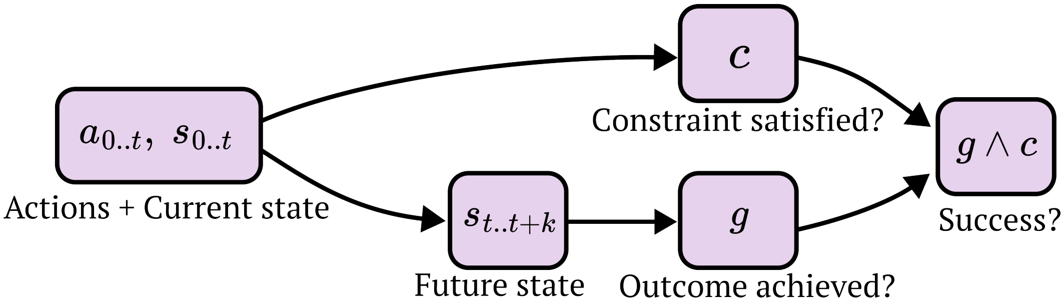

Figure 3: A combined goal that is a conjunction of a constraint and a consequence.

2.2 Why consequentialist goals are a necessary part of powerful AI

2.2.1 Robustness to diverse obstacles is driven by consequentialism

The primary reason we expect approximate consequentialism to be a good model for the behavior of useful systems is that we are assuming the AI is capable of generalizing well. By this we mean the *diversity* property in Section 1; the tasks we are assuming the AI is capable of involve a diverse array of unknown obstacles and difficulties, and the AI is able to achieve a particular outcome in spite of these obstacles. This allows us to make inductive conclusions such as: **If we know the AI can overcome one hundred particular obstacles when pursuing a particular goal, it can likely overcome another obstacle that is in roughly the same reference class as the first hundred.**

One of the main things that is useful about powerful AI is its ability to overcome many diverse and unknown obstacles, and this is what leads us to think of powerful AI as primarily pursuing consequences. When the AI generalizes to novel hard tasks, its behavior is still well described taking actions in order to achieve outcomes, in spite of unknown obstacles. This is what we mean by behaviorally consequentialist.

Humans attempting to achieve outcomes can be modeled as consequentialist. When we want to do something we work out a way to do it; when we encounter obstacles we work out ways to overcome them or search for alternative routes. Humans are not perfectly consequentialist, we often give up on things before achieving success, but we are consequentialist enough to radically reshape our world. Current LLMs such as GPT-4 [14] are useful for many tasks, but these are generally similar to their training distribution. They may be modeled as consequentialist when close to their training distribution, but this breaks relatively often when they encounter novel obstacles.

For example, we might assume an AI has the goal "build a fusion rocket", and for this it must be capable of completing many engineering subtasks of the form "have a design idea to solve a problem, build models to empirically test unknowns, refine design based on data". Because there is a huge variety of subtasks of roughly this form, it's likely that if given a new arbitrary subtask of this form (of a similar difficulty) the AI will be capable of completing that subtask. If it has generalized this far, it will probably continue to generalize.[14]

Compute budget and generalization pressure toward consequentialism

For some types of task with many diverse obstacles, there appear to be (large) benefits to computing actions on-the-fly, after observing the current obstacle, instead of in advance of knowing about the specific obstacle that the agent is facing. This is similar to the argument sketch in Conditions for mesa-optimization[3, 16].

We first make the simplifying assumption that each obstacle-overcoming strategy takes some amount of computation to work out (either during training or deployment). There are many possible obstacles. We could try to precompute every obstacle-overcoming strategy in advance, however this would take a huge amount of compute. Therefore, it is more computationally efficient overall to develop an algorithm which can work out how to overcome obstacles on-the-fly. The AI waits until it encounters an obstacle, and then runs the computation to work out how to overcome it.

Another variation of this argument is about generalization rather than the compute budget. If we're doing the precomputation approach, and we don't know or can't iterate over all the possible obstacles in advance, then the AI won't have memorized strategies that generalize well at inference time if an entirely new class of obstacle shows up. In contrast, if the AI operates by storing the outcome it wants, it can (try to) understand the obstacle on-the-fly and compute a new winning strategy.

2.2.2 Shallow and deep constraints

Another reason we consider outcome-oriented goals to be important is that shallow constraints on behavior tend to be easy to exploit (or hurt the usefulness of an AI if they are overly broad). By shallow constraints, we mean constraints that are easy to implement or specify, but don't robustly serve their intended purpose when subjected to optimization pressure. In other words, shallow constraints don't generalize well. For an intuitive example, a prohibition against saying explicitly false statements is easy to exploit by selectively revealing information or by using misleading but technically correct definitions. Deeper constraints, like "don't manipulate person X while trying to

[14] See arbital for more examples of consequentialist cognition [15].

achieve outcome Y", tend to be most naturally represented as outcomes (e.g. with "did X's beliefs become less true as a result of talking to me", or "if X knew what I knew, would they be upset by the way I communicated with them" [15]). Another example of a shallow constraint might be "don't kill people by doing recognizably people-killing actions". This kind of constraint wouldn't slow down a determined murderer. Legal systems instead use the outcome, intent and causal influence to define murder, which is a far better definition (but also has some ambiguous edge cases).

Constraints are hard to specify in a way that generalizes to new or different situations, or new strategies, because the AI is creatively searching for ways to achieve particular outcomes and will tend to eventually find loopholes. This is not to say it's impossible, merely that specified constraints need to solve an adversarial robustness problem.

2.2.3 Approximation

It's important to note that when we describe an agent as pursuing consequentialist goals, we're not saying it must be doing this optimally. Optimal consequentialism has a few tractability issues [17]. Our claim is that it's close enough such that it usually works, *and usually generalizes well* to new obstacles of similar difficulty to past obstacles. Specifically, it works and generalizes well enough in order to do hard tasks.

Whether the AI is well described as goal-directed is not binary, and this may apply more or less in different situations. We only assume that the AI is far enough along on the spectrum of goal-directedness, such that it is able to do hard tasks as specified in Section 1.

We can describe some axes upon which consequentialist AIs can be approximate.[16]. One type of approximation is how robust is the AI to unexpected obstacles (where more robust means the AI is capable of recovering from or working around a larger set of unforeseen obstacles). For example, your walking robot might recover from gently poking it, but might not recover from being knocked to the ground. Being knocked to the ground was outside of its set of unforeseen obstacles that could be overcome.

Another axis of approximation is to what extent the AI is correctly making (probabilistic) trade-offs, given a preferred outcome and many pathways to achieve that outcome. For example, suppose an AI must play a sequence of lotteries to gain money, which it can later spend to achieve some desired outcomes. To what extent does it make decisions that avoid sure losses or take advantage of sure gains? How efficient [20] is the AI with respect to a given set of other AIs?[17]

2.3 Conclusion

Powerful AIs should be behaviorally modeled as approximately selecting their actions to produce specific outcomes (often subject to non-consequentialist constraints). This is a necessary consequence of their capacity to solve hard tasks which involve unpredictable and complex obstacles; they are consequentialist enough to be able to do hard tasks. This doesn't mean that such powerful agents must act to achieve outcomes "by any means necessary". Constraints on behavior are not ruled out. We will argue in later sections that a powerful AI will be dangerous if its consequentialist goals or constraints are misspecified.

3 Hard tasks require learning new things

In previous sections we have argued that doing large scale novel science is a lot of work; and that an AI capable of this will be well modeled as doing consequentialist problem solving, i.e. taking actions in order to achieve outcomes. Part of why novel science is difficult is because it will require

[15] This is a constraint that involves a *counterfactual* outcome.

[16] By approximate, we mean in the same sense that UCT [18, 19] approximates tree search. Approximate algorithms have may suboptimal performance, but can satisfy more realistic time or space constraints.

[17] If an AI is avoiding losses and collecting all potential gains, and this behavior generalizes to other problems that it faces, we can usually treat it as an *approximate* utility maximizer [21] (for some utility function, outcome space, and action space). We haven't been specific enough about our capability assumptions to draw this conclusion in this report, we have only assumed enough to say that a powerful AI will be capable of generating actions that robustly produce particular outcomes, in spite of diverse sets of obstacles.

skills and knowledge that an AI doesn't initially have. In this section we will discuss why an AI will need to learn new things, and that this learning will need to be self-directed (not directed by humans-in-the-loop).

3.1 The AI will need to learn new things

If an AI system is going to do novel science on the scale of curing cancer or the Manhattan Project then it will need to be able to learn things. The AI will need to learn empirical facts that it didn't originally know, update its models of the world, and learn new skills that it wasn't originally trained to do.

3.1.1 Learning facts

A simple example could be that the AI does not know a specific fact; it may be missing the value of a physical constant, or not have read the operating manual for a particular machine. An AI doing novel science must be able to realize that it doesn't know a specific fact, and then take actions to learn it. For example, reading a physical constant from a textbook, planning and performing an experiment to measure a constant, or reading an operating manual in order to use a machine for a specific task. All the information that the AI needs will not be "stored in the weights" from the initial training.[18] When doing novel science, the AI can't initially know all the facts it needs, because many of these facts won't have been discovered yet.

3.1.2 Learning skills

Further, the AI will need to learn new skills, not simply learn new facts. The AI will run into cases where it doesn't know how to do something and so needs to learn. There will be particular algorithms or methods that are needed for the novel problems it is solving, but were not available (or invented) during the AI's initial training. An AI that was never trained on French could not be expected to be able to write in French without learning how to; similarly, an AI that was never trained on differential calculus would not simply know differential calculus. When doing novel science, AI systems will need to learn skills that others have developed previously (such as French or differential calculus), as well as develop skills that it needs to invent itself because they have not yet been invented by humans.

As with learning facts, these new skills initially will not be "stored in the weights" because the training process will not have had any reason to build them; especially if we are expecting the AI to generalize far from the training distribution. It doesn't seem likely that an AI would be able to intuit or extrapolate to differential calculus if it was never trained on it. This is not a claim that an AI could not learn differential calculus, but rather that this will require explicit work from either humans or the AI.

As an example, during the Manhattan Project, scientists invented Monte Carlo methods for numerically performing complicated integration [22]. These methods simply did not exist before the Manhattan Project, and so they needed to be invented and specific skills needed to be learned. The same applies for similar cases, such as developing mathematical theory to describe the hydrodynamics of shockwaves and centrifuge design.

3.2 Self-directed learning

This section will argue that the AI likely needs to be doing self-directed learning, where the AI itself is controlling what it learns. The alternative to the AI doing self-directed learning would be for a human to be constantly overseeing the AI, and looking for when the AI needs to learn a fact or skill. Then the human would either train the AI using supervised learning or RL, or assign the AI to learn this skill or fact as a new task.

3.2.1 Human-directed learning is a big efficiency hit

For the human to be able to competently and safely direct the AI to learn things, the human would have to adequately understand both the problem being solved and the AI's capability profile. Specifically,

[18] By "stored in the weights", we are referring to information that is explicitly represented in an easy-to-decode way, inside the AI program.

the human would need to know what skills were required to solve the problem, and that the AI was currently lacking those skills. It will be much faster for the AI to know this, as it is the one actually solving the problem, and has access to its current knowledge. This seems important when the AI is working in a domain that the human doesn't understand. If the AI is bottlenecked on human understanding, including when exploring research directions that don't pan out, the research speed won't be much faster than human research speed.

Additionally, some skills are only legibly useful with the benefit of hindsight, and so it may be hard for the human to realize that the AI needs to learn these. It can be difficult to explain the usefulness of math to students, and similarly, it may be difficult to realize the benefit of particular knowledge prior to knowing it.

3.2.2 Indirect self-directed learning

The AI may be able to "indirectly" do self-directed learning, for example by telling the human which skills or facts it should be trained on next. If the human doesn't fully understand the problem and is just deferring to the AI, then this is effectively the same as the AI doing self-directed learning. The AI is just "using the human as a puppet", or simply working around the human. There is some additional safety because the human may be able to prevent the AI from learning obviously harmful things. This seems like the most likely outcome of a naive attempt to put humans in the loop.

3.2.3 Useful versus safe tradeoff

Learning is useful for completing hard tasks. Having a human in the loop, deciding what should be learned is safer. For some tasks, having a human in the decision loop is fine. The claim we are making is that for hard tasks there is a significant tradeoff to be made, where putting a human in the loop will dramatically slow down the overall system. This topic will be discussed more in Section 5.

This is a specific case of a more general lesson; complex multifaceted tasks contain lots of bottlenecks, and solving one bottleneck means that the next bottleneck will dominate.[19] This isn't a fully general argument against it being possible to speed up anything. **It is an argument that dramatic acceleration on very diverse tasks requires an algorithm capable of attacking approximately every type of bottleneck that comes up.**

3.2.4 Examples

Here are two brief examples from the history of science where learning of facts or skills was necessary to make progress, and that this learning needed to be self-directed because knowledge needed to be built on previous discoveries.

Experimental science

We can consider Hodgkin and Huxley discovering the ionic mechanism of action potentials [24]. An observation had been made that some squids had extremely large axons, and so were more amenable to experimentation. This allowed electrodes to be inserted into cells in order to measure the potential difference across the membrane of the cell. Such an experiment would not have been possible with smaller cells. Here, we can see that a fact was learned (some squids have extremely large axons), and knowing this fact allowed for a novel experiment (including novel experimental techniques) to be developed, and this led to an important discovery (the ionic mechanism of action potentials). Learning the initial fact was needed and a chain of facts and techniques were built upon it in order to discover and demonstrate the mechanism of action potentials.

Theoretical mathematics

We can also look at a theoretical example, which does not require learning from observations in the world; the development of integration. In the 17th century Newton and Leibniz showed a connection between differentiation and integration with the fundamental theorem of calculus. However, integration at this point had not been rigorously formalized. Formalization of integration required the mathematics of limits; it was not until the 19th century that integration was formalized by Riemann. Here, the development of a rigorous definition of integration required the initial non-rigorous definition as well as additional mathematical tools (limits).

[19] See Why Tool AIs Want to Be Agent AIs [23].

These examples make the (perhaps obvious) case that when doing novel science, the AI system (or a human) will need to learn both facts and skills, and that these will necessarily build on themselves. The task of science is often inherently sequential.

3.3 Conclusion

An AI doing hard tasks will need to learn new things because we are asking it to do something novel; the task requires skills and knowledge that were not part of its initial training. The AI will likely need to learn a wide range of things, including things that were not specified or known in advance. It is much faster for the AI to do self-directed learning, rather than having a human direct the learning, which would require the human to have a deep understanding of what the AI is doing.

In the following sections we will discuss two important consequences of self-directed learning:

1. It is a major source of several kinds of distribution shift (relevant for Section 4).
2. It causes a number of problems for oversight, control, and predicting the limits of the capabilities of an AI before using it (relevant for Section 5).

4 Behavioral training is an imprecise way to specify goals

This section is about AIs which are created by behavioral training, and also are capable of doing the hard tasks as described in Section 1. By behavioral training we are referring to a wide variety of training techniques, which involve running a parameterized model, providing feedback on the output of the model, and updating the model based on this feedback. Examples include model-free RL using PPO [25], model-based RL like MuZero [26], next token prediction on tokens describing goal-directed behavior [27].

In previous sections we have described the behavioral properties that appear to be necessary for hard tasks. That is: **powerful AIs are behaviorally, approximately, optimizing their actions to produce outcomes.**[20]

We haven't described the internal operation of such trained AIs, mainly because we expect it to be a mess in the same way that humans and other evolved systems are a mess.[21]

There are several categories of problems that make it difficult to specify goals. Each category introduces an uncontrolled degree of freedom in the goal specification which exists because we are only using feedback based on behavior. Because there are lots of potential degrees of freedom that we don't have control over (via behavioral training), we can think of the space of "intended goals" as being a small subset of the space of "goals that empirically are pursued after training and deployment". In this way, behavioral training is imprecise and is exceedingly unlikely to nail down the goal we intended. We describe multiple separate failure modes and so failure is disjunctive; we only need one failure of goal specification for the AI to be misaligned. Here is an overview of the problems that we will describe in more detail:

- "Goal instabilities" that can come from the AI doing instrumentally convergent, capability-improving operations (during the process of doing hard tasks). These mean that the apparent goal of the AI changes over time, leading to success in training but unintended behavior after deployment.
  - Updating beliefs (Section 4.1.1)
  - Outer shell constraints (Section 4.1.2)
  - Changing terminal goals in-lifetime (Section 4.1.3)
- Goal specification that generalizes out-of-distribution is difficult because it's hard to distinguish between terminal and instrumental goals, and because the "distribution shifts" that we need a goal specification to be robust to are particularly large.
  - Goals naturally need to generalize a long way out of training distribution (Section 4.2.1)

[20] With all the caveats from Section 2 inherited; an AI might have constraints on actions and behavior, or more abstract constraints. However, a central part of its goal specification is about outcomes.

[21] As in biological systems, the system may be factorizable in some ways, but messy in other ways.

  - Instrumental goals are difficult to distinguish from terminal goals (Section 4.2.2)

- Outer behavioral incentives are difficult to get right in the first place. If these aren't perfect, they incentivize "playing the training game" (Section 4.4). This section will be brief, since this issue is well described elsewhere.

We will describe some of these problems using terms like *probability, belief, utility, goal* for concreteness, but each problem applies to learned algorithms that approximate goal-directed behavior. These problems are largely *caused* by the fact that the learned algorithm is messy and unstable, for example by the designers not carefully tracking the distinction between beliefs and goals in the clean way that the internals of an expected utility maximizer would.[22]

4.1 Unstable goals

In these examples, the AI initially appears to be pursuing one goal in training, but then takes some actions in order to become more capable, and this causes it to pursue a different goal.

4.1.1 Beliefs and goals can be mixed together

This example demonstrates a pattern common across other examples: we train an AI, this leads to an unstable implementation (of the AI) that generates behavior sufficient for low training loss. Some internal process eventually breaks this unstable implementation, causing the AI to pursue unintended goals. This first instability comes from a fundamental fact about probabilities and utilities: the same behavior can result from different sets of preferences and beliefs.[23] For example, an AI might execute a certain behavior, believing this helps it achieve some goal, but the belief is false, and in fact the behavior results in a different outcome. Or an AI might falsely believe a subgoal is instrumentally useful for reaching a terminal goal. In these cases, the AI behaves as if it is pursuing one goal, but it would not behave this way if it had different (more accurate) beliefs.

One of the differences between beliefs and goals (as we use these terms) is that beliefs update over time, as observations are made, and as evidence is considered. One might imagine an AI that is being trained using reinforcement learning to solve difficult puzzles, but where there is a way to hack the solution-checker. With good oversight, the AI would receive low reward for hacking the solution-checker during training, and one might hope that this correctly puts a "morality-like" constraint into the AI's goal. But there is another way it might update: the training could insert a false belief[24] into the agent. This might be a belief that hacking the solution-checker doesn't work, or leads to some other bad consequence, or is more difficult than it truly is. Later, if the AI updates this belief (because the belief was false), it may attempt to hack the solution-checker.

This mechanism (false beliefs instead of correctly specified goals and constraints) creates many additional degrees of freedom that allow particular behavior to be selected for, without the RL training necessarily having much influence on the ultimate goals of the agent.

4.1.2 Outer shell non-consequentialist constraints

There is a difference between the sort of constraints that are built into the core terminal goals of an AI, and the sort of constraints that are implemented outside of the core outcome-seeking algorithm, perhaps thought of as a kind of wrapper around the outside. The core outcome-seeking algorithm may treat the wrapper as an obstacle, and try to work around it or remove it in order to achieve its goal.[25]

In humans, an instinctual fear of heights is (probably) an example of this. A fear of heights is not a value that a human would usually endorse having in their terminal goals, and if it were really getting

[22] Where by expected utility maximizer we are here referring to a theoretically clean implementation that has a module that explicitly predicts the future outcome of all possible actions, and has another separate module that scores each outcome (not that we think such a design is realistic).

[23] See Jeffrey-Bolker rotation [28].

[24] Or a bad prior, or cognitive bias.

[25] From Section 2, we are assuming there is an algorithm which is behaviorally robustly outcome-seeking, at least sometimes or with some non-outcome constraints. In this section are noting that "wrapper algorithms" can be implemented around such an outcome-seeking algorithm. We aren't making claims about the implementation of the outcome-seeking algorithm, other than its existence.

in the way of achieving their goals, they would look for ways to overcome it. This is an example of an agent that initially produces behavior that appears to obey a constraint, but when there is a sufficiently large conflict between this constraint and other goals, then the constraint will be overcome.

We could imagine a hypothetical AI system, being trained by humans to make money in a legal and ethical way. The system may attempt to steal money from humans by stealing from their bank accounts, get caught, and have this behavior punished by RL training. This doesn't mean that it will learn the general goal "make money legally and ethically", instead it may learn the goal "make money" and a separate action-rejection circuit which rejects some stealing-like actions. The two possible ways for the RL to adjust the algorithm both result in identical behavior on the training distribution. It could then, in a later task, work out how to circumvent the action-rejection constraint by employing other humans or AI systems to do illegal things for it, or doing some form of self-modification to remove the constraint.

As another example, we could imagine an AlphaZero-style system guided by a trained heuristic [19]. The heuristic may have been trained in order to guide the search away from certain states; as an arbitrary example, we might have trained it to play chess with the additional constraint that it should keep its bishops as far apart as possible. However, this constraint and the goal of winning the game may be at odds; the heuristic can guide the moves, but more tree search can "override" the heuristic. In some future situation which involves using more search than previously on particular branches of the game tree, the constraint will no longer be obeyed. This is even more important if the system is set up to do learning in-lifetime, where the heuristic can get updated to remove the constraint, as this hurts the AI's performance for winning the game.

This kind of heuristic-goal mixup would be a fairly obvious mistake if we were designing the search procedure explicitly, as we would be in the AlphaZero example, but when we are training an algorithm that internally implements some planning procedure, it would be much more difficult to notice this type of mistake.

4.1.3 Changing goals over time

In very capable AIs, the terminal goals shouldn't change over time. This is because it's not in the interest of previous goals to be overwritten, so the AI should take steps to avoid its terminal goals changing. There is initially little reason for a behaviorally trained AI to have completely stable goals, especially if it was not directly trained on long and hard tasks. Because of this, there can easily be mechanisms built in that change terminal goals. Such mechanisms may or may not have a useful reason for being there, all that's necessary is that they didn't hurt performance in training enough to be removed.

One possible instance of this could be an AI that sets its goals based on observations. We could imagine an AI for solving puzzles, where the AI receives a description of the aim of the puzzle as an observation, this observation sets the AI's terminal goal, and then the AI pursues this goal. In this system, the AI's terminal goal is changing for every task. In deployment, it might run for longer, or be in a more information-dense environment or a more adversarial environment. Then, some way into a task, it pattern-matches some text that looks a little like a puzzle description (but wasn't intended that way) and its goal would unintentionally be adjusted.

Or, as another example, an AI trained to solve tasks could constantly be having its goals shifted by the observations, but the AI is running over short time horizons in training and the distribution is homogeneous enough for it to appear stable and achieve good training loss. When the AI is set to do a hard task that requires much more time or a larger distribution shift this could cause the goals to change more radically. Here the most likely outcome seems to be that the AI breaks and not be capable or dangerous, but if we assume that it would be capable of doing the hard task (and thus remains competently goal-directed) then it could end up competently pursuing a misaligned goal.

This kind of mechanism where the goals change isn't *necessarily* incentivized by training. It's just a way of implementing an AI that is unstable but performs similar behavior to the behavior we want (during training). It is therefore another source of degrees of freedom in goal specification, which makes it less likely the intended goal is internalized.

Goal-updating machinery is dangerous to the extent that we don't understand exactly how it works, and can't predict how it will react to very different environments.[26] When there is a lot of messiness and approximation, as in most complex systems evolved from training data, a huge source of difficulty in specifying goals is that we don't know precisely how the system evolves over time.

4.2 Out-of-distribution generalization

4.2.1 OOD generalization difficulty

If a trained AI can be thought of as approximately pursuing an outcome, in a manner robust to almost any obstacle it encounters, then some kind of description of the outcome seems like it must be stored in the AI. It need not necessarily be cleanly stored in one place but the AI must somehow recognize good predicted outcomes. As in Section 2, we will call information representing the goal g.

The thesis of this section is that to be safe, g should generalize in a well-understood way to previously unexplored (and unconsidered) regions of future-trajectory-space. This is because an AI that is capable of generalizing to a new large problem is considering many large sections of search space that haven't been considered before [29, 30]. An AI that builds new models of the environment seems particularly vulnerable to this kind of goal misgeneralization, because it is capable of considering future outcomes that weren't ever considered during training.

The distribution shift causing a problem here is that the AI is facing a novel problem, requiring previously unseen information, and in attempting to solve the problem, the AI is considering trajectories that are very dissimilar to trajectories that were selected for or against during training. The properties of g are not pinned down by data, so we are relying on generalization. This problem becomes worse when our intended goal is more complex to specify.[27] It also depends on the amount and diversity of training, but we are always depending on generalization because of the inherent novelty of hard tasks.

4.2.2 Instrumental goals as terminal goals

Solving difficult problems involves pursuing many instrumentally convergent subgoals along the way. A behavioral training signal doesn't fully distinguish between things that were instrumental for achieving the goal and the goal itself. And so instrumental goals can be selected for, to some extent, just as if they were terminal goals. It seems likely they wouldn't be a dominant part of the goal, because that would result in instrumental goal-directed behavior that sometimes conflicts with good training performance. But instrumental goals being incorporated as a small part of terminal goals appears to be a fairly natural design for selected agents, and creates more degrees of freedom in goal specification from behavioral training.

For example, an AI trained to run a factory may be incentivized to gain more influence over employees, because this is useful for managing the factory. The AI may end up partially terminally valuing "power over humans", because this was behaviorally identical to having the goal of running the factory well. These differing goals could come apart outside of training if the AI was given the option to gain much more power over humans, because now a relatively small part of the goal has become much easier to satisfy.

4.3 Deliberate deception

If an AI develops internally represented goals during training, and has adequate understanding of the training process, this may result in deceptive alignment [3] (or "scheming" [8]). Here, the AI performs well according to the training process, deliberately, in order to avoid being modified or having its misalignment detected. The AI ends up doing well on the training objective, but pursues something entirely different when it gains confidence that it is no longer in training or that it is otherwise safe to defect.

[26] If for some reason you want to deliberately train unstable goals and use this mechanism to specify goals, remember the previous and next examples of problems will all be mixed up in the same AI.

[27] Conversely, simpler goals should be more likely to be specified correctly. To be more precise about what we mean by "complex" and "simple", we're not necessarily referring to K-complexity. We're referring to the inductive bias of the overall training algorithm. The sample efficiency of correctly learning a goal representation depends on how much the intended goal is favored by the inductive bias of the learning algorithm (plus all the biases introduced by the non-g learned machinery).

This is another way for the goals of an AI to be underspecified by behavioral training. This may happen simultaneously with the failure modes described above. The AI could be deceptively aligned, even if its misaligned goal is unstable and will change with further training or in deployment. What ultimately matters, from a risk perspective, is that the AI is achieving good training loss for reasons other than it being aligned.

4.4 Badly designed training incentives

For standard behavioral training there is also the problem that the training signal (for example, in reinforcement learning) doesn't perfectly (or unbiasedly) track what we would want. This may incentivize the AI to "play the training game", where it exploits unintended flaws in the training process that achieve more reward [31]. These flaws could include: predictable human biases, mistakes in the training setup, incentives to be sycophantic for human approval. Strategies that exploit these flaws will receive more reward (and be more reinforced) than strategies that straightforwardly do what the humans intend.

This may be thought of as an "outer alignment" failure, where we are effectively telling the AI (via the reward) to do something that we don't actually want. This is rather different from the other discussed problems with behavioral training which are about the underspecification of the internals of the AIs and could be thought of as "inner alignment" problems.

4.5 Conclusion

In this section we've compiled an (incomplete) list of problems that make it difficult to specify goals of an AI using behavioral training. Instead, the AI will pursue correlates of the intended goal and instrumental goals, have "instincts" that can be overcome, have false beliefs, and maybe update its goals over time in an unpredictable way. This should add up to behavior that does well on a training distribution of tasks, but pursues its own goals when the AI generalizes to new hard tasks. These failure modes are disjunctive; only one of these failure modes needs to happen for the AI to be misaligned. We will describe more in Section 6.1.3 why the imprecision that results from these failure modes is dangerous.

While the goal might be unstable, there is no particular incentive for the goal to move closer to the goal we intended as the AI faces context changes, more difficult problems, learns more about the world, or generally learns new skills.

5 Control of misaligned AIs is difficult

In this section we will describe some difficulties with forcing misaligned AIs to do hard tasks, where the tasks are helpful for us but go against the AI's own interests. Here, we say that an AI is *controlled* if it is unable to cause bad outcomes and can be directed to usefully complete hard tasks, even though it is misaligned. Currently, several groups intend to use early AIs to accelerate alignment research.[28] We expect this kind of plan to either be extremely dangerous or insufficiently helpful for solving alignment in time.[29]

Such plans rest on two pillars:

1. The AI will do useful work because the training process has selected it to be good at completing similar work.
2. Escape or sabotage will be much more difficult for the AI than helping us, and we can control the AI's capabilities to ensure that it isn't ever capable of escape or sabotage.

If both of these are achievable, then a lab can set up a large-scale research effort primarily staffed by AI, to solve hard tasks such as AI alignment.

We think there are problems with both of these pillars and don't expect all of these problems to be surmounted.

[28] e.g. OpenAI [1], Redwood Research is also primarily preparing for this pathway [9, 32].

[29] By "in time", we mean before another, less careful group deploys misaligned AGI.

For the first pillar:

- The sort of powerful AI we are imagining, as described in Sections 1 and 2, is capable of considering the long-term consequences of its actions and evaluating those consequences according to its goals. Because of this, we expect it to be very difficult to get the AI to do useful research if this research isn't part of a plan to achieve its own goals (Section 5.1).
- We expect problems checking the quality of AI generated research output, and also problems producing useful human research data to imitate (Section 5.2).

For the second pillar:

- We don't expect escape to be harder than the problem-solving it is routinely capable of (Section 5.3).
- We expect sabotage and delaying to be much easier than doing useful research (Section 5.4).
- Internal planning won't be transparent (although if this can be made fully transparent, through advances or luck, this makes many problems easier to solve) (Section 5.5).
- Predicting the capabilities and knowledge required to complete particular complex tasks is difficult, especially if the AI is trying to mislead you *and* is capable of learning over time. This is not to say that we can't estimate what tasks an AI will be capable of doing, but we expect the margin of error to be large (Section 5.6).

We will go through each of these points in more detail in the following sections. Section 5.7.1 discusses why we think AI assisted monitoring doesn't fundamentally change much.

Importantly, we expect the problems with control of misaligned AIs to be amplified by the fact there is so much work to do and so many opportunities to make mistakes. There will be adversarial cognition working against the plan, looking for any holes in security, pushing research in various directions, and attempting to stop us from learning from mistakes. In a large project there will be many unknown unknowns, vulnerabilities that we may be unable to anticipate but that a powerful AI can exploit.

5.1 Misaligned goal-directed AI won't cooperate

We have assumed our AI is capable of doing research; solving problems creatively and in a goal-directed way. As we discuss in Section 1 and 2, we consider an AI to be acting in pursuit of long-term consequences if it is robust to diverse obstacles. We are relying on this argument here. Likely obstacles need to be constantly foreseen and planned around. Given that the goal of this AI is not to complete the task we want it to complete, it will presumably foresee that the research task we have given it isn't likely to help it achieve its goals and nor would overtly refusing to do the task. It will take whatever approach seems most likely to achieve its goals, which might include escape, sabotage, or simply delaying.

5.2 Eliciting useful capabilities via outer training loop

Here we will consider the case where we are attempting to make a misaligned AI do useful research by training it on similar research. This approach involves having the AI predict the result of some human-generated research, and then modifying the AI with gradient descent according to what it outputs; if the AI outputs a very similar result as the humans, it is not modified. This training may be supervised learning or reinforcement learning, depending on the nature of the research. If the AI wants to avoid being modified by training then, even if it is misaligned, it has an incentive to generate the same result that the humans did. If we give the AI a research question that humans have not solved, and the AI is unable to tell that humans have not solved it, then the AI will have the same incentive to generate the result that humans would have found.

5.2.1 Difficulties

There are three main difficulties with this approach to produce useful work:

- It only applies to relatively easy to check research results

- We would need *lots* of useful human-generated research
- It isn't clear that the training would make the AI output good research in critical cases

This approach only works for certain types of research where there is a final correct answer that can be checked. This could apply to work like "prove this theorem", "write an algorithm for this well-specified task", or more rarely "find a clean formalization of this pretheoretic concept". Such tasks sometimes have a single well determined answer, which a human researcher can work out for themselves. However there is a lot of important work that isn't like this; for example, choosing research directions, or designing new algorithms for imperfectly defined problems. These contain lots of small decisions that are strongly entangled with the desired outcome and constraints, and which are costly to verify in detail. In particular, the misaligned AI would be optimizing against the human's ability to provide an accurate oversight signal, so this signal wouldn't be reliable on non-trivial tasks.

There needs to be sufficient human-generated research in the domain where we are attempting to elicit the AI's capabilities. This research is expensive to generate. In particular, we need enough such that we can train a model to reliably produce very accurate predictions of the research output produced by human researchers. This depends somewhat on the eventual data-efficiency of the training method, in particular on the transfer to tasks that human overseers are unlikely to have completed, and which would be against the interests of the AI to complete.

If the AI outputs something different than the human-generated result, it isn't clear that the subsequent training would make the AI generalize to generate useful research in the future. In particular, it plausibly doesn't generalize to doing novel research, where the AI needs to do significant online work in order to find the result.[30] In this case, it seems like there are three things that the training could do: make the AI correctly instrumentally value high training signal (such that it won't do novel research if it can tell that it isn't likely to receive feedback), create hacky modifications to the AI that don't generalize to novel research, or actually succeed in make the AI want to do good research. It seems like the first two are more easily reinforced, especially given limited training data.[31]

In order for the AI to do useful research, it needs to pursue the goal (or sub-goal) of doing research. There's no particular reason for the gradient based training to actually modify the goal representation, rather than building some additional narrow machinery on top. In this case, the training would not be entirely using the AI's underlying capabilities and would be less efficient.

It does seem true that in the limit of adversarial training and a perfect training signal, the AI would be behaviorally aligned; that is, on all inputs it would do good research. However, this means providing the AI with all the answers, and the entire value of AI is its ability to generalize to problems that we otherwise didn't have the answer to. Without providing all the answers, there isn't ever a behavioral guarantee that the AI is aligned.

5.3 Difficulty of escape compared to routine problem-solving

In this section we discuss whether escape is much more difficult than helping with research. We are considering a wide range of potential strategies an AI might pursue to escape. The simplest might look like asking a sympathetic researcher to contact a government and convince them of the value of stealing the AI. A more difficult strategy might involve finding and using a privilege escalation exploit in a virtual machine used for experiments, such that logging can be manipulated and further actions toward exfiltration are undetectable. Both of these obviously depend on the AI correctly learning a lot of facts about the world around it, such as confidently inferring that a particular researcher will act in particular ways. An incomplete list of similar affordances is in Appendix: Attack surfaces.

Plans to escape depend entirely on the particular security flaws that happen to be present in the real world control scheme. Unfortunately, hard tasks are likely to require skills such as inferring and predicting complex systems given limited indirect evidence. Many skills are likewise valuable for difficult problem solving and also useful for everything else, including developing plans to escape.

The extent to which we expect security flaws to be present and exploitable depends on the size of the attack surfaces and the vigilance of the people running the project. There are several factors that

[30] This is not to say that this training would never successfully elicit anything from the AI. For example, this training would likely elicit facts that were already "stored in the weights".

[31] Another contributing difficulty is that a misaligned AI can plausibly exploration hack to manipulate or slow down the training process.

make attack surfaces larger. Hard tasks take a lot of time and resources, and very likely include a lot of time interacting with overseers, data and experiments. The AI will have many opportunities to learn about security measures, collect information, and consider strategies for escape. There will likewise be many opportunities for mistakes or lapses in security. On top of this, there are several difficult tradeoffs to make, such as the extent to which information and compute is freely available to the AI for the purposes of research.

On the other hand, there are factors that can make escape much harder without heavily impacting the difficulty of cooperating. For example, close monitoring and logging of actions increases the difficulty of escape much more than it increases the difficulty of assisting with research (although it probably does have some cost because careful monitoring should result in frequent false positives, where overseers see slightly suspicious actions and need to pause to investigate).

Overall, we expect escape to be incredibly difficult, but it isn't clear that it will be much more difficult than the hard tasks we want the AI to solve for us.

5.4 Difficulty of sabotage and delay

If the AI is unable to immediately escape, it can instead sabotage and delay helping with useful research until there is a security vulnerability it can exploit. If the humans don't adequately understand the research, the AI can direct research in ways that are less likely to be useful.

The AI could sabotage research by explicitly suggesting research directions it knows aren't useful. If humans were choosing the research directions, the AI could implement some experiments in order to make some directions seem more promising and others seem less promising. The AI could exploit biases in the researchers, prompting them to pursue their pet research direction which will likely turn out to be useless. The humans would need to thoroughly understand the experiment in order to catch this which would dramatically slow down research. This becomes much more difficult if the AI is adversarial, for example, deliberately writing complicated code or explaining concepts badly. There can also be dead ends in research and so if the AI chooses research directions that aren't that useful, this isn't strong evidence the AI is misaligned. It may not be feasible to assess if a research direction is good until it has results, and even then it can be difficult to determine if it was better than alternatives.

The AI can also direct the research in directions which let it learn skills to escape; for example, learning specific programming techniques or extremely complicated mathematics which humans don't understand. It could also use this to learn things about the specific containment system; running code that triggers certain bugs, learning which actions trigger review by humans. Humans may be unable to judge if something is innocuous or part of a broader plan to learn about security flaws.

Human overseers could disallow the AI to even suggest facts or skills to learn, and only let the human make this decision without AI input. However, if the human doesn't understand how to move forward with the research, this would slow down research probably below the point of usefulness. This is the inherent tradeoff between the autonomy of the AI and how overseeable it is; the human can make many of the AI's decisions and thus have better oversight, but this makes the AI drastically less useful.

5.5 Transparency and honesty

The risk of escape, sabotage, and delay is much greater if the overseers don't have adequate understanding of the AI's thoughts and plans. We strongly expect that there will not be adequate transparency, barring significant advancements or luck. Much of the difficulty in controlling and getting useful work from a powerful AI comes from the AI not being transparent, such that it can consider plans and outcomes without us knowing. If the AI is *actually* fully transparent and well understood, control becomes a lot less difficult.[32] We'll first discuss why we don't expect to get adequate transparency, and then discuss difficulties for worlds where there is partial transparency.

We think it is most likely that the internals of a powerful AI will be largely uninterpretable, especially when the AI is having complex "thoughts" about novel concepts. We want the AI to be solving very

[32] As does AI alignment.

complicated problems, and currently there are no methods to tie the AI's externalized reasoning to the reasoning it is doing. Current LLMs already provide some evidence for this [33, 34].

This lack of transparency can allow the AI to develop skills or reason about escape, because we can't ensure that it is focused on the problem we want rather than something else.

We now consider the case where there is partial transparency, in the sense that humans can initially see and understand some aspects of the AI's cognition. In this case, one specific failure mode is that the AI develops techniques which allow it to do obfuscated reasoning (i.e. steganography). For example, we could imagine a system that can think some thoughts internally but needs to externalize complicated reasoning. It could encode hidden information in its writing, allowing it to do additional reasoning that humans don't understand. There are many degrees of freedom in writing that allow additional information to be encoded. By default though, we don't even expect to have this level of partial transparency.

We might hope that we could get around the problem of humans not understanding things by having the AI explain them. This could allow us to avoid steganography and also stop the AI steering research in bad directions. Unfortunately, we don't have a method for making AI systems honestly explain things. Ideally we would have a proper training signal for "honestly give the human an accurate impression of what is going on, such that they can make a decision they reflectively endorse", but we don't have this. This is made worse because the skills required for explaining things to the human are almost the same skills that are useful for misleading the human. For example, teaching involves learning which analogies and frames the human finds most convincing, or learning which technical areas the human is stronger or weaker in.

The difficulty of interpreting powerful cognition is a core part of the argument we are making. The greater the extent to which we have tools that allow us to understand the inner workings of powerful AI, the harder it will be for an AI to plot against overseers, and the harder it will be to sabotage or delay research.

5.6 Eliciting dangerous capabilities

If we were trying to pull off a control plan, we would want to be able to tell if the AI has dangerous capabilities that would allow it to escape. The two main categories are hacking and human manipulation, although it is plausible that there are unknown unknowns that the AI could exploit.

One approach for eliciting dangerous capabilities is similar to eliciting useful capabilities. It focuses on eliciting very specific bad capabilities; for example, if an AI is going to break out via hacking, it will likely need to be able to notice vulnerabilities in code. So we create a dataset of code which has some vulnerabilities in it, and then fine-tune the AI to find the vulnerabilities. If we fine-tune the AI and it performs well, this means that it might at least be able to have this capability.

5.6.1 Difficulties

Similarly to the case of eliciting useful capabilities, if the AI internally represented the presence of a vulnerability in some observed code without the need to do additional reasoning, the fine-tuning approach is likely to be able to extract this knowledge, if we know ground truth. One way for the AI to get around this, could be for the AI to avoid learning the specific skill of instantly noticing vulnerabilities, but learn related skills which would allow it to find these with additional reasoning. It may be hard to efficiently elicit capabilities which require additional (perhaps chain-of-thought) reasoning steps. This would also stop one simply training a probe on the AI's internals, because the vulnerability would not be represented.

There is a different failure mode where effectively all capable AIs learn to classify code vulnerabilities, when they are specifically fine-tuned. And so this test would imply that all AI systems are potentially dangerous.

This leaves us in a position where the approach may provide almost no evidence. If the method successfully elicits the dangerous capability, this may be because the model is dangerous and was hiding this latent ability, or because this was trained via the fine-tuning. If the method does not elicit the capability, this may be because the model is benign, or because the fine-tuning was not sufficient to elicit capabilities which require chain-of-thought or other work.

Another issue that this can approach can only test for very specific things, such as noticing vulnerabilities in code. It can't test more general abilities such as "reasoning about potential security holes" or "planning escape attempts".

5.7 AI assisted monitoring doesn't change the picture much

So far in this section we have been discussing containment strategies that involve using careful human monitoring, without explicit assistance from other AI systems. However, there are some advantages to using AI monitoring instead. For example, with automated monitoring, it is much cheaper to watch every action taken by a different AI and flag those that are suspicious. While we think that such techniques could make escape or delay somewhat more difficult, we don't expect this to dramatically change the difficulty of these tasks. We suspect that the escape difficulty created by AI assisted monitoring isn't the main disagreement between us and proponents of these techniques.[33] Instead, we suspect that disagreements about the value of AI assisted monitoring are usually traced back to disagreements about the necessary intelligence and generality of the AI system doing hard tasks.

Despite this, we think it's valuable to discuss some of the object level difficulties with automated monitoring in this section. We will split this section into two parts, depending on the capability level of the AI system that is being proposed for monitoring [36]:

- AI systems that are powerful, but misaligned. For example, a copy of the AI we are trying to use to do research.
- AI systems which are safe because they are not capable of powerful consequentialist planning or reasoning. GPT-4 is an example.

5.7.1 Monitoring with powerful AI systems

One could attempt to use a separate powerful AI system to monitor the powerful AI doing the object-level useful task. Examples of such strategies include *untrusted monitoring* [32] or debate [35, 37] (if used for monitoring). The advantage of having a powerful AI monitor is that it could keep up with the work and foresee long-term strategies.

One core issue here is collusion between the powerful AI doing the useful task and the powerful AI performing monitoring. This is made possible if the humans can't detect all information being passed between the AI systems. It seems likely that a powerful AI will understand things and be able to detect patterns that the humans can't, and also know which things the humans won't understand. This makes sending difficult-to-detect messages relatively easy. Monitoring with powerful AI systems seems much more promising if we could be confident from the beginning that the AIs were not colluding and had no intention to collude. But we have no way of ensuring this, because it requires precisely specifying the goals of the AI.

Several approaches have been proposed to patch or reduce the size of various problems. One such patch for collusion would be for the humans to intentionally insert fake "honeypot" collusion messages between the AI systems, in an attempt to make them openly defect. However, it seems like the best strategy for AI systems would be to only act when very confident of success, and only choose to communicate via channels that the humans don't understand.

For concreteness, we will consider the scalable oversight proposal debate [35, 37]. The scheme involves two AIs each arguing for a position; for example, deciding which research direction is promising or whether a piece of code is safe to run. The AIs each propose an answer or action to take, and then argue back and forth; the debate is ultimately judged by a human. One hope here is that in order to "win" the debate and avoid punishment (for example, by being updated via gradient descent[34]), the AIs will have an incentive to tell the truth because this is the easiest option to argue for. Each AI argues for their own side, as well as pointing out flaws in their opponent's case. The humans running the debate can rerun the argument tree from different places and with different information in an attempt to catch one of the debaters being inconsistent. There are a number difficulties with this approach:

[33] E.g. debate [35] or control [32].

[34] Although it isn't at all clear that this would work as "punishment" in the relevant sense.

- The AIs may be discussing things the humans don't understand and there will be a lot of back and forth, likely allowing the AIs to communicate information without humans realizing. For example, this could allow them to agree to avoid debating particular points, or entirely stage the debate.
- The humans may judge the debate incorrectly, especially when under adversarial pressure from the AIs. This reduces the incentive to tell the truth.
- Despite the attempt to incentivize the AIs to choose the better (or most true) position to argue for, it seems unlikely that the threat of gradient descent is strong enough to ensure this. The AI can choose to behave badly in a few critical cases, where it can trade off the risk of being updated for increased chance at achieving its goals.
- Even honest AIs may arrive at different conclusions or make different arguments, given slightly different circumstances, and so inconsistencies between positions may not be strong evidence of bad behavior. This primarily applies in the case where there is no easy way to check the correct answer, such as choosing research directions.

5.7.2 Monitoring with weak AI systems

One could try to use relatively weak AIs to monitor a more powerful AI doing a hard task. Weak AIs could be checking the actions for obvious attempts to take actions that humans are unlikely to approve of, such as inserting obvious security vulnerabilities into code.

A key problem is that when the powerful AI is working in a domain that the weak AI system doesn't understand, then the weak system won't know everything that is going on and won't be able to provide an accurate signal. For example, the weak system would not be able to point out a backdoor implemented in an algorithm that it had never seen before and doesn't understand. This is also a problem with human oversight, except in this case it's probably worse.

As with human supervision, we don't know how to get the powerful AI to provide honest explanations of its actions to the weak AIs.

One proposed strategy is to use weak AIs to paraphrase the outputs of a powerful AI [32]. This could be done to prevent powerful AIs from passing secret messages to other AIs. For example, we could have a powerful AI writing code to be overseen by another powerful AI. We might be worried that the coding AI will insert hidden messages into its code instructing the overseer to not flag a certain security vulnerability (the overseer could be incentivized to normally flag vulnerabilities because these are deliberately added by humans). The hope here is that paraphrasing by a weak AI could remove these hidden messages. However, the weak AI won't be able to remove all hidden messages; it may be capable of removing messages that are encoded in the "syntax" (e.g. renaming variables, rewording comments) but not able to modify the "semantics" (e.g. the coding AI might choose a particular algorithm or implement it in a particular way). The weak AI is being asked to rewrite the output of the powerful AI while preserving the explicit meaning; if the weak AI doesn't understand the content it may make a mistake and change the explicit meaning (e.g. change code to be incorrect), or it may let hidden information through because it isn't able to modify the content enough without making it wrong.[35]

5.8 Conclusion

When a powerful AI is directly incentivized to look for security problems in a real-world research environment, we expect it to find them. Failing this, we expect the AI to be able to subtly delay or sabotage the project, or wait for a lapse in security or increased capabilities. Taking a powerful misaligned AI and training it to produce useful research does not mean that the AI will do useful work in the future, partially because this training does not necessarily make the AI motivated to do useful work.

See Appendix: 4-hour research assistant for a worked example of the difficulty of using an AI to speed up research focused on the tradeoff between safety and usefulness.

[35] Relatedly, if we end up having partial transparency into the AI's thoughts, we could attempt to use paraphrasing to reduce steganographic reasoning [38]. This would mean having the weak AI paraphrase and replace each segment of reasoning, and the hope being that this will only preserve legible reasoning. This has the same problems as above with preventing messages being passed between different AIs.

This section is not arguing against attempting to oversee and control misaligned AIs *if we have to*, but rather that we expect this to be extremely difficult. If humanity goes down the path of building misaligned powerful AIs, we should try as hard as we can to keep them contained. We should stack as many layers of security as we can think of, and minimize the size of attack surfaces. But the balancing act between security and usefulness has many failure modes, particularly under the pressure we are expecting, and given the size and nature of the hard tasks that we want a trapped AI to complete.

6 Powerful misaligned AI would be bad

So far, we have argued that it will be hard to build an AI that is powerful, without also being misaligned and difficult to safely extract useful work from. Given this, there is much uncertainty about exactly how the future could go wrong. Does an AI escape from a lab? What does it try to do? Is it in conflict with humans? How much is the world thrown into chaos? We are uncertain about these questions, but we expect the default outcomes to be bad for humans.

Overall, it looks difficult but possible to achieve good outcomes by taking actions before a powerful misaligned AI exists and is put to use, but that impressive levels of global coordination might be required. After a misaligned AI gains any ability to act autonomously in the world, we expect that it will be too late for humans to take any mitigating actions.

6.1 A misaligned AI achieving its goals would be bad for humans

In this section we will discuss concretely why AI systems, which are capable of hard tasks and are misaligned, will likely lead to catastrophically bad outcomes for humans; outcomes such as extinction or total loss of control of the future. Such an AI would be attempting to pursue outcomes, and be extremely good at overcoming obstacles in order to achieve outcomes. Humans would be treated as obstacles in the way of the AI achieving its goals, to whatever extent we are actually obstacles.

6.1.1 Instrumental convergence toward gaining control of limited resources

Resources and power are extremely useful for achieving a wide range of goals, especially goals about the external world. However, humans also want resources and power for achieving their goals. This will put the misaligned AI in direct competition with the humans. Additionally, humans may be one of the largest threats to the AI achieving its goals, because we are able to fight back against the AI. This means that the AI will have extremely strong incentives to disempower humans, in order to prevent them from disempowering it.

6.1.2 Ambitiousness

One objection to the instrumental convergence argument is that the AI may not have ambitious, long-term goals, and so it may be easy to satisfy these goals without causing harm to humans. For example, it's possible for an agent to want to solve one problem, and then after then not want to do anything else. However we don't have that much control over the goals a trained AI develops; we aren't able to ensure that it develops such a bounded goal. This is made worse by the fact that when we ask the AI to do hard, novel science, we are asking it to do a very ambitious thing. Projects on the scale of the Manhattan Project or curing cancer seemingly require the AI to have at least somewhat ambitious goals. Our mainline expectation is that such a goal specification has many components, some of which are "ambitious" and some of which aren't. Because of this model of goal specification that is made up of multiple components, it seems unlikely that all components of the goal cease to be motivating after achieving a bounded outcome. We don't have fine enough control over how the AI develops to ensure that the ambitiousness of every aspect of its goals is in the narrow range of "safely bounded but still able to do hard tasks".

6.1.3 Consequences of misspecified goals

Some kinds of goal misspecification lead to extremely bad consequences, and other kinds don't. The kind of goal misspecification that leads to extremely bad consequences is the kind that is susceptible to extremal Goodhart [39]. By this we mean goal specifications that contain large errors in some parts of action-space or outcome-space, such that the optimum of the AIs goal corresponds to a bad outcome for humans. As an analogy, a utility function that has bounded error everywhere isn't very susceptible to extremal Goodhart. A utility function over a high dimensional space, where the function isn't well specified on a couple of dimensions, is likely to be susceptible to extremal Goodhart.

Misspecified consequences

Errors in the goals that powerful AIs end up ultimately pursuing could be very bad, and from Section 4 we don't expect to be able to specify these goals well. Many of the goals that an AI could end up with are not compatible with human empowerment or human survival.[36]

We will consider the example of an AI which is tasked with the ultimate goal of curing cancer. One goal that the AI could develop is a drive to acquire knowledge and "understand" the environment. This goal is represented in the AI's ontology and may be different to the human version of these concepts. This could develop because much of the AI's training was based around gaining information about the world and making novel and correct predictions. Like in the *deceptive alignment* story, the AI may go along with the rest of training, partially because this allows it to gain some knowledge but also because it doesn't want to be shut down. If the AI manages to eventually break out and take over it may do unpredictable and unwanted things in pursuit of its ultimate goal; for example, build massive scientific apparatus for examining specific wavelengths emitted from certain stars, or run biological experiments to learn the function of specific proteins.

Other goals that this AI may develop are far more alien, because the AI's internal ontology may not match the human ontology. We don't have control over what ontology the AI uses or what goals it develops. It may end up pursuing a state of the world which is full of "beauty", "achievement", "politeness" and "approval", where the AI's version of these concepts is related to the human concept but is very different when taken to extremes. The AI may calculate which states of the world are best according to its goals, and the "best" states may not be what humans would want. Maybe maximizing the amount of some strange crystal, maybe filling space with intricate machines, maybe building ever larger computers to run calculations.

For both of these examples, what the AI ends up pursuing is not something humans would want, and the AI has a strong incentive to stop humans disempowering it.

Misspecified constraints

Some AI designers may attempt to instill constraints in an AI with training, in order to prevent unwanted behavior. If the constraints are misspecified they will not constrain the AI's behavior in ways we intended. And, as usual, we don't have precise control over the internal constraints that an AI develops. These constraints will naturally be under optimization pressure from the AI itself.

We could attempt to make the AI "honest", that is to tell humans true and helpful things. However, this could easily be improperly specified (for example by the mechanisms in Section 4) and lead to the AI valuing some variant of honesty, missing the value entirely, or trade off against other values. The AI, by its consequentialist nature, will seek loopholes in any constraints that get in the way of its consequentialist goals. The AI may be attempting to achieve a goal where it is helpful to manipulate an overseer. If so, it would say things that humans would class as lies but are perfectly allowed under the AI's constraints.

As another example, we might want the AI to have very little effect on the world outside of the lab it is given. If this is improperly specified, the AI may learn to not let any physical substance created in the lab to influence the outside world. But it may still be able to affect the outside world via sending radio signals to computers outside the lab and having them affect the world. Here the AI may try hard to achieve its goals, and in doing so find an edge case in its constraints that it can exploit. The AI may not be modeling this as an edge case in its constraints that it is "trying" to overcome, but instead there is some blocked path (creating substances in the lab that affect the world), and it is taking another route (using external computers to affect the world).

[36] It seems possible that if a misaligned AI disempowered humanity it may still keep humans alive or because its training did make something upstream of "human survival" part of its mess of motivations.

6.2 Why would a powerful AI be capable of defeating humanity, given moderate levels of freedom?

Given freedom to act unsupervised and with lower risk of shutdown, a powerful AI could do research to improve its own capabilities. There are a number of easy lower bound options for this, including copying itself, or rapidly running experiments that test different ways to reduce the computation required for running a single copy. More advanced experiments could involve building more direct interfaces between programs and cognitive processes, or identifying the most bottlenecked parts of the AI's algorithm and modifying these parts to improve performance. With the ability to introspect, it could be easy to set up fast and informative feedback loops for experimentation.

After such a process, the AI can take whatever pathway to increased power and security is most likely to work [40]. This may involve trading with some groups of people. It may involve operating secretly for a while while further improving its capabilities and influence. It could directly build up its own infrastructure for more research, defense and collection of resources; or manipulate human governments or political groups.

One possible story could be that the AI gains control over the data center where it is being trained and fakes the performance metrics such that the overseers don't think anything bad is happening. It then spends time finding a set of security vulnerabilities that allow it to escape and gain enough freedom for further research. The AI copies itself to data centers around the world for redundancy. It then contacts various terrorist groups, and manipulates and assists them in creating biological weapons which are released simultaneously across the world, throwing the world into chaos.

Such stories are too specific to be accurate, but they gesture at the strategies and affordances a misaligned AI may use. We expect reality to be messier; it will be filled with human error and confusion, with people in stressful situations under both time pressure and social pressure, and security protocols that have never been adequately tested against a powerful AI attempting to escape.

7 Conclusion

In this report we've argued that AI systems which are capable of doing hard, novel science will be dangerously misaligned.

The tasks we are focusing on are inherently outcome-oriented, and hence our AI must be behaviorally consequentialist; its behavior is well described as taking actions to achieve specific outcomes. Hard tasks require that the AI is able to overcome a range of diverse and novel obstacles, and if the AI can overcome these obstacles, then it can likely overcome additional obstacles we didn't intend it to overcome.

We currently use behavioral training to create AIs, by which we mean AI systems are created by modifying a program based on its external behavior. Behavioral training does not fully specify the ultimate goals of the trained AI. There are many degrees of freedom in this process, and hence it is extremely unlikely that a trained AI ends up pursuing the goals we intended.

One hope is to use oversight and control to extract useful work out of AIs, even if they are misaligned. The scale of the project makes this extremely difficult, especially because a misaligned AI will be creatively searching for loopholes in any security measures.

Finally, we discussed the consequences of a powerful, misaligned AI attempting to achieve its goals. We expect misaligned goals to not be compatible with continued human empowerment or survival. An escaped AI could build its skills, and amass power and resources, until it ultimately disempowers humanity in pursuit of whatever goal it has. See Appendix: Argument Graph for a visual representation of the overall argument.

7.1 Requests for AI developers

Given the arguments in this report, we have a number of requests for AI developers.

1. **Stop attempting to build powerful AI using behavioral training.** As we have argued, this is likely to produce an AI which is misaligned. We understand that this request is unlikely to be listened to, despite being the safest course of action in our view.

2. **If an AI is caught attempting to break out or otherwise subvert oversight, stop using it.** We hope that this is an easy request. It is important to not simply paper over signs of misalignment. AI developers should have in place well-specified protocols for how and when to stop using AIs, *before* bad behavior is discovered.

   AIs caught attempting egregiously bad actions could also be used to build consensus about the extreme risks. Such an AI could also be used as a "model organism" [41] to study misalignment, but obviously it is important that this is done extremely carefully.

3. **If the AI is doing research or other work that humans don't understand, stop using it.** Much (but not all) of our worry about AIs evading control comes from them doing additional reasoning in inscrutable research. This is less of a problem if humans understand all the work. Note that it is still *very* important not to be tricked into believing one understands, which could be a large risk if the AI is being used to "explain" what it is doing.

4. **Use extremely thorough control schemes.** Much of the difficulty in safely containing misaligned AIs comes from there being many attack surfaces in whatever scheme we use. This is because of the large scale task we are attempting. However, if there are fewer holes the probability of escape goes down. Note that it is extremely important to not have a false sense of security, simply because humans can't easily think of specific vulnerabilities.

5. **If an AI could subvert the control scheme, shut it down.** Control schemes are likely to get some confidence in safety from attempting to ensure that the AI does not have capabilities which could allow it to escape (see Appendix: Capabilities elicitation for considerations here). If the AI has the capability to subvert the control scheme, then one should no longer believe the scheme provides safety, and so should shut everything down. This also applies if the AI can get around the capabilities elicitation protocol, and so we would not know if it had dangerous capabilities.

8 Definitions

Hard task: A task that involves large search spaces, is outcome-oriented, and contains many novel and diverse subtasks. See Section 1 for examples.

AI: Program that appears "intelligent" in a very broad intuitive sense, including current LLMs and chess bots. Used in the same way as in "the research field of AI".

Powerful AI: A program that is capable of performing hard tasks autonomously.

Consequentialist AI: A program that takes observations, outputs actions. The actions end up causing particular future outcomes in diverse counterfactual scenarios.

Trained AI: An AI created primarily by searching over a space of programs, and selecting or modifying the programs based on their behavior, e.g. by gradient descent on neural networks.

Goal: A relatively simple description from which can be derived what behavior an agent is likely to produce, or would choose to produce if given counterfactual opportunities. Can be a combination of outcomes and constraints.

Outcome: An event in the future, especially one that might be preferred or dispreferred. Not dependent on near-term actions. Constraint: A goal that is highly dependent on near-term actions, rather than outcomes in the future.

Misaligned AI: An AI whose behavior is well described as pursuing a goal that is different from the goal intended by its creators (where the difference is large enough that it leads to bad outcomes according to the intended goal. See Section 6.)

Aligned AI: An AI whose behavior is well described by the goals intended by its creators.

References

1. Leike, J. & Sutskever, I. *Introducing Superalignment* July 2023. https://openai.com/blog/introducing-superalignment (2024).
2. Kenton, Z. *et al.* Threat Model Literature Review. en. https://www.alignmentforum.org/posts/wnnkD6P2k2TfHnNmt/threat-model-literature-review (Nov. 2022).
3. Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J. & Garrabrant, S. Risks from Learned Optimization in Advanced Machine Learning Systems. *arXiv.org*. https://arxiv.org/abs/1906.01820v3 (June 2019).
4. Carlsmith, J. Is Power-Seeking AI an Existential Risk? *arXiv.org*. https://arxiv.org/abs/2206.13353v1 (June 2022).
5. Yudkowsky, E. AGI Ruin: A List of Lethalities. https://www.alignmentforum.org/posts/uMQ3cqWDPHhjtiesc/agi-ruin-a-list-of-lethalities (June 2022).
6. Soares, N. A central AI alignment problem: capabilities generalization, and the sharp left turn. https://www.alignmentforum.org/posts/GNhMPAWcfBCASy8e6/a-central-ai-alignment-problem-capabilities-generalization (July 2022).
7. Hubinger, E. How likely is deceptive alignment? https://www.alignmentforum.org/posts/A9NxPTwbw6r6Awuwt/how-likely-is-deceptive-alignment (Aug. 2022).
8. Carlsmith, J. Scheming AIs: Will AIs fake alignment during training in order to get power? *arXiv.org*. https://arxiv.org/abs/2311.08379v3 (Nov. 2023).
9. Greenblatt, R. & Shlegeris, B. The case for ensuring that powerful AIs are controlled. https://www.alignmentforum.org/posts/kcKrE9mzEHrdqtDpE/the-case-for-ensuring-that-powerful-ais-are-controlled (Jan. 2024).
10. Christiano, P., Neyman, E. & Xu, M. Formalizing the presumption of independence. *arXiv.org*. https://arxiv.org/abs/2211.06738v1 (Nov. 2022).
11. Garrabrant, S., Benson-Tilsen, T., Critch, A., Soares, N. & Taylor, J. Logical Induction. *arXiv.org*. https://arxiv.org/abs/1609.03543v5 (Sept. 2016).
12. Diffractor & Kosoy, V. Introduction To The Infra-Bayesianism Sequence. https://www.alignmentforum.org/posts/zB4f7QqKhBHa5b37a/introduction-to-the-infra-bayesianism-sequence (Aug. 2020).
13. Alexander, L. & Moore, M. in *The Stanford Encyclopedia of Philosophy* (ed Zalta, E. N.) Winter 2021 (Metaphysics Research Lab, Stanford University, 2021). https://plato.stanford.edu/archives/win2021/entries/ethics-deontological/.
14. OpenAI *et al. GPT-4 Technical Report* arXiv:2303.08774 [cs]. Dec. 2023. http://arxiv.org/abs/2303.08774.
15. Arbital. *Consequentialist cognition* https://arbital.com/p/consequentialist/.
16. Hubinger, E., Merwijk, C. v., Mikulik, V., Skalse, J. & Garrabrant, S. Conditions for Mesa-Optimization. https://www.alignmentforum.org/posts/q2rCMHNXazALgQpGH/conditions-for-mesa-optimization (June 2019).
17. Hutter, M. A Theory of Universal Artificial Intelligence based on Algorithmic Complexity. *arXiv.org*. https://arxiv.org/abs/cs/0004001v1 (Apr. 2000).
18. Kocsis, L. & Szepesvári, C. *Bandit Based Monte-Carlo Planning* en. in *Machine Learning: ECML 2006* (eds Fürnkranz, J., Scheffer, T. & Spiliopoulou, M.) (Springer, Berlin, Heidelberg, 2006), 282–293. ISBN: 978-3-540-46056-5.
19. Silver, D. *et al.* Mastering the game of Go with deep neural networks and tree search. en. *Nature* **529.** Number: 7587 Publisher: Nature Publishing Group, 484–489. ISSN: 1476-4687. https://www.nature.com/articles/nature16961 (2024) (Jan. 2016).
20. Arbital. Epistemic and instrumental efficiency. https://arbital.com/p/efficiency/.
21. Yudkowsky, E. Coherent decisions imply consistent utilities. https://www.lesswrong.com/posts/RQpNHSiWaXTvDxt6R/coherent-decisions-imply-consistent-utilities (May 2019).
22. Rhodes, R. *The Making of the Atomic Bomb* Anniversary,Reprint edition. English. ISBN: 978-1-4516-7761-4 (Simon & Schuster, New York, N.Y., Jan. 1986).
23. Branwen, G. Why Tool AIs Want to Be Agent AIs. https://gwern.net/tool-ai (Sept. 2016).

24. Schwiening, C. J. A brief historical perspective: Hodgkin and Huxley. *The Journal of Physiology* **590.** Publisher: John Wiley & Sons, Ltd, 2571–2575. ISSN: 1469-7793. https://onlinelibrary.wiley.com/doi/abs/10.1113/jphysiol.2012.230458 (June 2012).

25. Schulman, J., Wolski, F., Dhariwal, P., Radford, A. & Klimov, O. *Proximal Policy Optimization Algorithms* arXiv:1707.06347 [cs]. Aug. 2017. http://arxiv.org/abs/1707.06347.

26. Schrittwieser, J. *et al.* Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model. *Nature* **588.** arXiv:1911.08265 [cs, stat], 604–609. ISSN: 0028-0836, 1476-4687. http://arxiv.org/abs/1911.08265 (Dec. 2020).

27. Chen, L. *et al. Decision Transformer: Reinforcement Learning via Sequence Modeling* arXiv:2106.01345 [cs]. June 2021. http://arxiv.org/abs/2106.01345.

28. Demski, A. Probability is Real, and Value is Complex. https://www.alignmentforum.org/posts/oheKfWA7SsvpK7SGp/probability-is-real-and-value-is-complex (July 2018).

29. Langosco, L. *et al. Goal Misgeneralization in Deep Reinforcement Learning* arXiv:2105.14111 [cs]. Jan. 2023. http://arxiv.org/abs/2105.14111.

30. Shah, R. *et al. Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals* arXiv:2210.01790 [cs]. Nov. 2022. http://arxiv.org/abs/2210.01790.

31. Cotra, A. Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover. https://www.alignmentforum.org/posts/pRkFkzwKZ2zfa3R6H/without-specific-countermeasures-the-easiest-path-to (July 2022).

32. Greenblatt, R., Shlegeris, B., Sachan, K. & Roger, F. *AI Control: Improving Safety Despite Intentional Subversion* arXiv:2312.06942 [cs]. Jan. 2024. http://arxiv.org/abs/2312.06942.

33. Turpin, M., Michael, J., Perez, E. & Bowman, S. R. *Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting* arXiv:2305.04388 [cs]. Dec. 2023. http://arxiv.org/abs/2305.04388.

34. Lanham, T. *et al. Measuring Faithfulness in Chain-of-Thought Reasoning* arXiv:2307.13702 [cs]. July 2023. http://arxiv.org/abs/2307.13702.

35. Radhakrishnan, A. Anthropic Fall 2023 Debate Progress Update. https://www.alignmentforum.org/posts/QtqysYdJRenWFeWc4/anthropic-fall-2023-debate-progress-update (Nov. 2023).

36. Shlegeris, B. Untrusted smart models and trusted dumb models. https://www.alignmentforum.org/posts/LhxHcASQwpNa3mRNk/untrusted-smart-models-and-trusted-dumb-models (Nov. 2023).

37. Irving, G., Christiano, P. & Amodei, D. *AI safety via debate* arXiv:1805.00899 [cs, stat]. Oct. 2018. http://arxiv.org/abs/1805.00899.

38. Roger, F. & Greenblatt, R. *Preventing Language Models From Hiding Their Reasoning* arXiv:2310.18512 [cs]. Oct. 2023. http://arxiv.org/abs/2310.18512.

39. Manheim, D. & Garrabrant, S. *Categorizing Variants of Goodhart's Law* arXiv:1803.04585 [cs, q-fin, stat]. Feb. 2019. http://arxiv.org/abs/1803.04585.

40. Karnofsky, H. *AI Could Defeat All Of Us Combined* June 2022. https://www.cold-takes.com/ai-could-defeat-all-of-us-combined/.

41. Hubinger, E., Schiefer, N., Denison, C. & Perez, E. Model Organisms of Misalignment: The Case for a New Pillar of Alignment Research. https://www.alignmentforum.org/posts/ChDH335ckdvpxXaXX/model-organisms-of-misalignment-the-case-for-a-new-pillar-of-1 (Aug. 2023).

Appendix: Argument graph

We have created these argument graphs because they help us to keep track of the entire argument all at once (Figures 8 and 8). Possibly they will be helpful for someone reading this document, although quite possibly they won't. We recommend only looking at this section after having read through the whole report. These graphs are a simplification of the arguments, and may be misleading without the context of the rest of the report.

Section 1

- **Lot of work**: Impactful science is hard and takes a lot of work.
- **Outcome-oriented**: Hard tasks will be outcome-oriented.
- **Novel and Diverse**: Hard tasks will have novel and diverse obstacles

Section 2

- **Consequentialist**: AIs which can do hard tasks will be behaviorally consequentialist.
  - The AI will need to be doing a lot of work for a task that is specified by the outcome, rather than the specific actions to take. This task will have novel and diverse obstacles, and in order to successfully do it, the AI needs to continually overcome these obstacles and continue towards the goal. Additionally, if the AI is doing self-directed learning, it will need to take actions with the goal of learning specific things.

Section 3

- **Learn new things**: The AI will need to learn new things
  - Novel science necessarily requires that the AI learns new facts and skills. These may be facts and skills that humans previously knew but the AI was not trained on, but also things that the AI works out by itself. This includes self-directed learning, where the AI is choosing what to learn because of efficiency and lack of human understanding.

Section 4

- **Training leads to imprecise goals**: Behavioral training doesn't precisely specify goals.
  - Training based on AI behavior leaves many degrees of freedom in the goals that the AI ends up ultimately pursuing. There are multiple goals which are consistent with the same training behavior, and also the AIs goals may be unstable.
- **The AI will be in novel situations**
  - There are many novel situations that the AI will face: the environment the AI is acting in may be different, learning new skills and facts can open up novel strategies, the AI will be able to run for longer periods of time, the AI is doing work that humans can now no longer easily check.
- **Goal misgeneralization in novel situations**: Goal misgeneralization when faced with novel situations
  - When the AI is in a novel situation, this may cause its goals to "come apart" from the goals intended by the humans. This can be a consequence of the AI pursuing the same goals all along, and the novel situation now shows it is misaligned.

Section 5 (Separate Graph)

- **Not transparent:** We don't have transparency into an AI's thoughts.
  - We don't have good methods to understand an AI's thoughts and plans, and even if we have partial transparency the AI could learn to do obfuscated reasoning.
- **Humans don't understand**: Humans won't completely understand the purpose and details of a powerful AI's actions.

  - There will be a lot of work to oversee, especially when the AI is learning new things. This will make it hard for a human overseer to keep up. Additionally, because the AI is not honest, we can't just get it to explain everything to the overseers.

- **Delay and sabotage:** The AI can delay or sabotage research.
  - If the AI is unable to escape, it can delay or send research in bad directions. This allows it to wait for a critical lapse in security or develop skills that over time will help it escape.
- **Humans struggle to judge research:** The humans will struggle to assess the quality of the research done by AIs
  - The humans will have limited time to assess the research quality, and misaligned AIs would adversarially produce research that is misleading and hard to judge.
- **The AI does not have the goal of doing research:**
  - The AI will be misaligned and so will not have the goal of doing research the humans want, especially research that goes against the AI's interests.
- **The AI can take critical bad actions:** The AI will be able to take critical bad actions which allow it to escape or do something similarly bad
  - The AI will have been able to delay and sabotage research, and build up skills that are useful to escape. Lack of human understanding can let the AI gain dangerous skills without humans realizing. The AI can also wait for a lapse in security, which is likely to eventually happen in a large-scale project.
- **Hard to elicit useful capabilities:** It will be hard to get a misaligned AI to do useful research
  - It will be hard to force a misaligned AI to do useful research that is not helpful for its goals, in the timeframes required. Humans will struggle to provide a good training signal to elicit this work, it isn't clear that the sample efficiency of this training will be high enough to make the AI do novel useful research, given limited training data.
- **Control fails:** Plans to safely get useful work out of a misaligned AI will likely fail
  - It will be hard to force a misaligned AI to do research, and given the scale of the project the AI will likely have opportunities to take critical bad actions. Monitoring using other AI systems seems unlikely to dramatically change this outcome (this is discussed in more detail in Section 5).

Section 6

- **AI goals will be bad**: The AI will be misaligned and trying to achieve goals that are bad for humans
  - Due to instrumental convergence, an AI may seek power and resources. We are assuming it has goals which are at least as ambitious as the novel science we are attempting. Such misaligned goals, when optimized for hard, are not outcomes humans would want.
- **AI catastrophe**: Powerful misaligned AIs will likely be able to achieve their goals.
  - A misaligned, powerful AI would attempt to disempower humans and take over. It would do this because its goals are incompatible with human goals. It would be able to do this if our oversight fails and the AI is able to learn new skills that allow it to escape and take over.

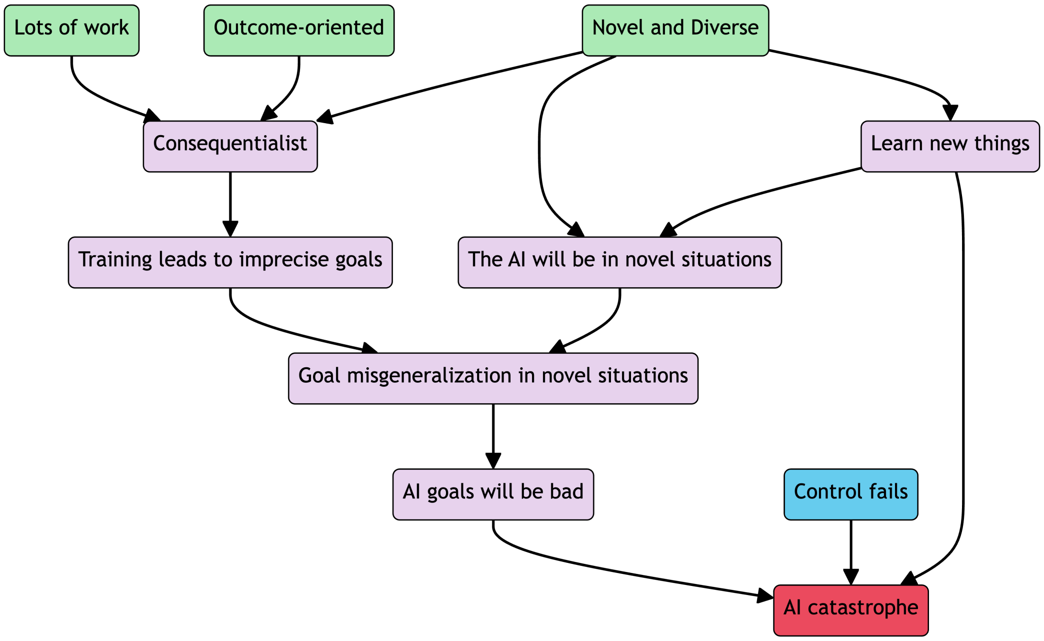

Figure 4: Graph of dependencies for the claims of the overall argument

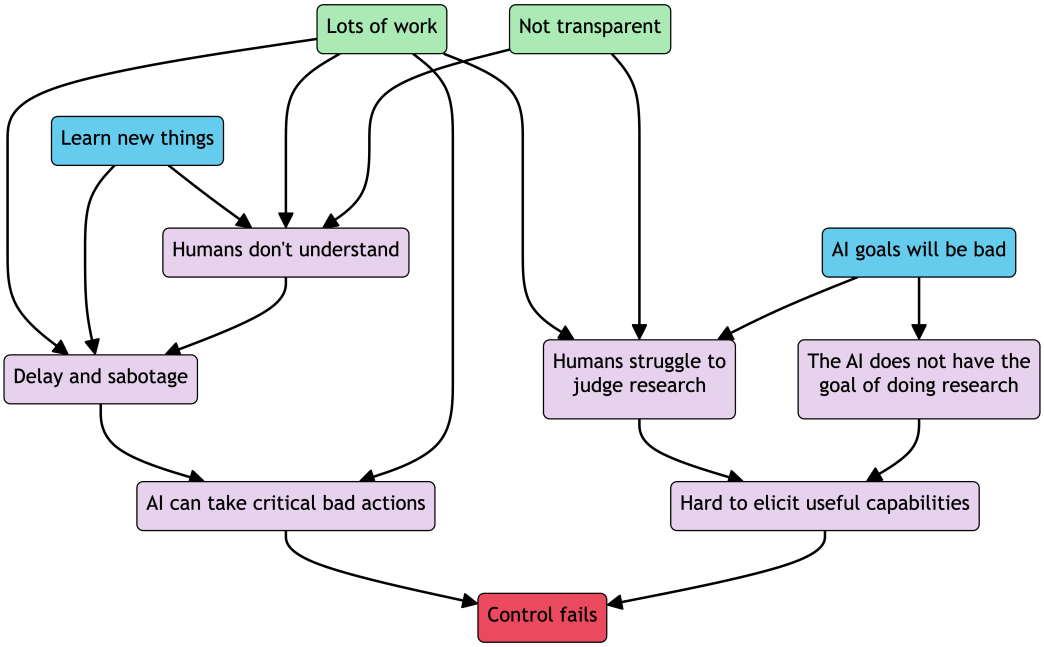
Lots of work
Not transparent
Learn new things
Humans don't understand
AI goals will be bad
Delay and sabotage
Humans struggle to judge research
The AI does not have the goal of doing research
AI can take critical bad actions
Hard to elicit useful capabilities
Control fails

Figure 5: Graph of dependencies for the claims about AI control, made in Section 5

Appendix: Attack surfaces

This is an incomplete list of attack surfaces or vulnerabilities which an AI being used for research could use, in order to steer the research direction and eventually escape. This list is here to emphasize the many options an AI would have, especially in a big (Manhattan Project-scale) project. There will additionally likely be other attack surfaces that the AI may be able to think up, given that it may have a lot of knowledge and that it can likely improve its ability to plan.

- Human attack surfaces
    - Making some humans look good or bad, via experimental results. The AI could notice that some researchers are less security-conscious than others, and manipulate experimental results (write better code for them, have more useful ideas), in order to get them more influence. Similarly, it could attempt to make especially suspicious researchers appear to be bad at their jobs.
    - Making suspicious researchers look crazy and paranoid, in order to discredit them.
    - Making security conscious lab leaders look bad at their jobs, in order to get them replaced by less security conscious people.
    - Finding out which things frustrate certain researchers. The AI will be explaining complicated research results, and find out if there are things that annoy certain researchers and make them think less clearly.
    - Fostering envy and resentment between researchers. The AI could worsen trust between researchers in the lab by making some have experiments that go well or poorly, or by making it appear as if someone had stolen someone else's idea.
    - Guiding research in directions to buy time and wait for an opportunity to escape. The AI could make research that would let it gain certain skills seem more promising, and research that might go against its aims seem less promising.
    - Communicating with spies (insider threat actors) in the lab, in order to get exfiltrated. Even if all messages between the AI and researchers were monitored, they may be able to communicate in code. A spy could potentially communicate details about the AI's hardware, passwords, security holes. Or a spy may simply be able to physically exfiltrate the AI's weights/code.
    - Making certain scientific breakthroughs in order to attract attention from powerful actors like nation states. The aim may be to try and get nation states to put a lot of effort into attempting to steal the AI.
    - Making "progress" on oversight and control research that is actually not robust, in a way that is specific to the AI. For example, if the AI is extremely good at a specific skill that the humans don't know, the AI could push for research that is vulnerable to this specific thing.
    - Influencing decisions by claiming to be suffering or otherwise morally relevant.[37]
- Computer attack surfaces
    - Exploiting bugs in the software or hardware. This could be via subtly running into errors that lead to unwanted behavior. Or attempting the innocuous task of finding out why a bug happens in certain cases but not others. Potentially allowing the AI to write to an unmonitored location that it wasn't supposed to.
    - Using timing of certain programs to work out specifics of its hardware.
    - Pushing research in directions which require specialized hardware. This may be more believable if multiple different research directions end up concluding that specific specialized hardware is useful.
    - Pushing for research which seems to overall improve security, even though it requires giving the AI slightly more autonomy.
    - Obfuscating computation, for example building a chemistry simulator that also computes things the AI is not supposed to be able to do.
- Experimental or engineering attack surfaces (for the case where the AI is doing physical work, probably not AI alignment research)
    - Designing electronics that subtly send messages via radio waves, or allow it to connect to nearby wifi networks. Here the electronics would likely have some other explicit purpose; for example, an electron microscope or a robotic arm.

[37] Note that this does not mean that we should totally ignore AIs if they tell us they are suffering.

- Finding ways to get around containment measures. For example, finding holes in a Faraday cage, or building an amplifier for incoming electromagnetic signals.
- Physical destruction, for example a chemical explosion that looks accidental.

Appendix: 4-hour research assistant

Here we will describe an example of an AI system which may feel intuitively safe, while also allowing for a large speed-up in useful AI alignment research progress; in this case speeding up research by a factor of 30. We will argue that such an AI system is not safe.

The specific AI system is an AI which acts as a research assistant to a human researcher. The AI can do the equivalent of 4 hours of human work extremely quickly, and then report back to the human. We'll call each 4-hour equivalent a "chunk".

An important aspect is that work which would have previously taken the AI alignment field one year now only takes one thirtieth of the time; a year of work now happens in less than two weeks. We will argue that in order to make this AI safe, we would need to restrict its capabilities to the point where it is no longer useful. If we do not make these restrictions, the AI is likely to be dangerous.

The human researcher will need to be giving the AI research assistant new 4 hour tasks every 8 minutes.[38]

The human can be telling the AI to do various tasks:

- Specific object-level research tasks
- Summarizing results (note that the researcher may only have 8 minutes to understand them)
- Having the AI to come up with good research tasks to ask other AI's to do
- Brainstorming general research strategies and directions

The 4-hour research assistant will need to be (behaviorally) goal-directed, at least over the course of the equivalent of 4 hours. If it were aligned, then its behavior would be well described as taking actions in order to do the task set by the human.

The AI will need to learn things; this will include learning things that are already known (e.g. reading a textbook) or learning things that aren't already known (e.g. doing novel research, discovering novel facts, inventing novel methods).

The primary failure mode is that the AI will need to retain skills and knowledge between tasks. The individual tasks will unlikely be able to be done in total isolation. The AI may spend one chunk learning a skill, and then be required to use this skill in the next chunk. It may spend many chunks building up familiarity with a specific code base, and if it lost this knowledge it would become useless.

Additionally, many research tasks cannot be simply divided up into 4 hour chunks that can be done separately, especially not in advance. Some tasks are inherently serial, and can only be properly evaluated at the end. An example of this may be a mathematical proof that would take a human 12 hours to figure out; if this task was given to the 4-hour research assistant AI then it may make some progress on the task, but without the end result the human would be unable to tell if it was *actually* doing the task.

If the AI is reset or "mindwiped" every time, it would have to spend all of each chunk catching up and would not have time to do any new work. It does not seem possible to get useful, novel science work out of such an AI. As an analogy, a scientist in the Manhattan project would not have been able to catch up and do useful work in 4 hours.

If instead we don't reset the AI for every chunk, such that it can retain what it has learned, then it is no longer really a "4-hour research assistant". It would now be able to do tasks and plan over longer

[38] How many calls? Say a researcher does 12 hours of work a day, for the 30× speed-up, we will be wanting to output the equivalent of 360 hours of work a day. The research assistant does the equivalent of 4 hours of work each call. Assuming all the work gets done by the AI and the human is just choosing what to tell them to do: $360/4 = 90$ calls a day. $12/90 \times 60 = 8$. The human researcher will need to be giving the AI assistant a new 4 hour research task every 8 minutes.

time horizons. It would be able to build up skills; both skills we want it to have and skills that we don't. If it is learning things in one chunk to use in the next, it is doing planning outside of the 4 hour window. It will have to be capable of even longer term planning if it is deciding on research directions, or noticing if a specific result is useful for the overall long-term project.

Because of the short time that the human has to evaluate the AI's work and assign it a new task, the human will likely not be able to fully understand what the AI is working on. It is important to remember that the human is trying to understand 4 hours of work in a novel scientific field in less than 8 minutes. Given these time constraints and the lack of AI honesty, the AI would likely be able to give convincing *post hoc* explanations for the actions that it took. The AI will likely be able to guide the research, allowing it to learn things that would help it evade oversight or escape.