Imitation Learning is Probably Existentially Safe

Michael K. Cohen
University of California, Berkeley
mkcohen@berkeley.edu

Marcus Hutter
Google DeepMind
www.hutter1.net

Abstract

Concerns about extinction risk from AI vary among experts in the field. But AI encompasses a very broad category of algorithms. Perhaps some algorithms would pose an extinction risk, and others wouldn't. Such an observation might be of great interest to both regulators and innovators. This paper argues that advanced imitation learners would likely *not* cause human extinction. We first present a simple argument to that effect, and then we rebut six different arguments that have been made to the contrary. A common theme of most of these arguments is a story for how a subroutine within an advanced imitation learner could hijack the imitation learner's behavior toward its own ends. But we argue that each argument is flawed and each story implausible.

1 Introduction

While many theorists have come to share the view that sufficiently advanced AI systems might pose a threat to the continued existence of humanity [Hinton et al., 2023, Cohen et al., 2022, Russell, 2019, Bostrom, 2014], it is important, if we are to make progress in thinking about this issue, to be clear about which types of AI pose the genuine threats. That way we can focus on where the danger actually lies. This paper aims to refute claims that imitation learning algorithms present such a threat. While we do think there are types of AI we should be worried about, that does not extend to all types of AI. So in what follows, we will examine arguments that have been put forward that imitation learners present an extinction risk to humanity, and explain why we think they go wrong.

First, we'll offer a simple argument that a sufficiently advanced supervised learning algorithm, trained to imitate humans, would very likely not gain total control over humanity (to the point of making everyone defenseless) and then cause or allow human extinction from that position.

No human has ever gained total control over humanity. It would be a very basic mistake to think anyone ever has. Moreover, if they did so, very few humans would accept human extinction. An imitation learner that successfully gained total control over humanity and then allowed human extinction would, on both counts, be an extremely poor imitation of any human, and easily distinguishable from one, whereas an advanced imitation learner will likely imitate humans well.

This basic observation should establish that any conclusion to the contrary should be very surprising, and so a high degree of rigor should be expected from arguments to that effect. If a highly advanced supervised learning algorithm is directed to the task of imitating a human, then powerful forces of optimization are seeking a target that is fundamentally existentially safe: indistinguishability from humans. Stories about how such optimization might fail should be extremely careful in establishing the plausibility of every step.

In this paper, we'll rebut six different arguments we've encountered that a sufficiently advanced supervised learning algorithm, trained to imitate humans, *would* likely cause human extinction. These arguments originate from Yudkowsky [2008] (the Attention Director Argument), Christiano [2016] (the Cartesian Demon Argument), Krueger [2019] (the Simplicity of Optimality Argument), Branwen [2022] (the Character Destiny Argument), Yudkowsky [2023] (the Rational Subroutine Argument),

Under review.

and Hubinger et al. [2019] (the Deceptive Alignment Argument). Note: Christiano only thinks his argument is possibly correct, rather than likely correct, for the advanced AI systems that we will end up creating. And Branwen does not think his hypothetical is likely, only plausible enough to discuss. But maybe some of the hundreds of upvoters on the community blog LessWrong consider it likely.

In all cases, we have rewritten the arguments originating from those sources (some of which are spread over many pages with gaps that need to be filled in). For Christiano [2016] and Hubinger et al. [2019], our rewritten versions of their arguments are shorter, but the longer originals are no stronger at the locations that we contest. And for the other four sources, the original text is no thorougher than our characterization of their argument. None of the arguments have been peer reviewed, and to our knowledge, only Hubinger et al. [2019] was reviewed even informally prior to publication. However, we can assure the reader they are taken seriously in many circles.

2 Attention Director Argument

First, we'll argue that Yudkowsky's [2008] Attention Director Argument fails. This argument is:

1. Learned imitation is the prediction of what actions a human would take.
2. Really high-quality prediction sometimes requires making decisions about where to direct attention and what thoughts to think.
3. These choices must be oriented toward a goal—the goal being to come up with an accurate prediction.
4. Most goals are best accomplished by first gaining complete control over humanity, if possible, and directing all available resources toward the goal.
5. The goal-oriented attention-director will recognize this possibility, and do so.

We contend that the "most" in point 4 is too sloppy, and so it fails to apply to this setting, and point 5 also fails. First, note that point 4 is less strong than it may appear, because it is most plausible when the goal is over the long term, whereas making an accurate prediction is quite a short-term goal. But the rest of our counterargument will focus on more definite gaps in the Attention Director Argument. Call the goal-oriented attention-director the "subagent". The structure of our argument is as follows, although some points will involve some back and forth:

A. Goal-directed agents, of which the subagent is one, must have beliefs about the consequences of their actions.

B. If one model of the consequences of a subagent's actions is much simpler than another, and both are viable for understanding the consequences of past behavior, the subagent is unlikely to take the much more complex model seriously. (Occam's razor)

C. Yudkowsky [2008] implicitly assumes the subagent will have a model (Model M) in which its actions control the computation of a computer located on planet Earth, but there is another much simpler model (Model N).

D. Let Model N only model the effects of the actions within a simple computational environment, and let it not model any further effects of actions after the computation following that action is complete.

E. The subagent's actions will have never before had any visible effect outside the simple computational environment, before the computation following that action is complete.

F. The simpler Model N is viable for understanding the consequences of past behavior.

G. The subagent will not understand its actions to affect the outside world.

We will not offer detailed defenses of points A and B, since we believe it is considered common knowledge that they are prerequisite for goal attainment. The "beliefs" mentioned in point A need not be contained in a dedicated subroutine (as it is for a "model-based" agent). So we begin by detailing Model N. Note that the subagent lives in a simple computational environment: the computation that the predictor executes to make predictions. So let Model N be a model in which the effects of the subagent's actions are restricted to such a simple computational environment. The subagent is trying to promote the predictive accuracy of the predictor, so the subagent's model of the consequences of its actions must output an estimate of predictive accuracy given the possible actions that it could take; then it can use is model to compare actions according to that criterion. Model N is depicted in Figure 1 (a); subfigures (b) and (c) are discussed later.

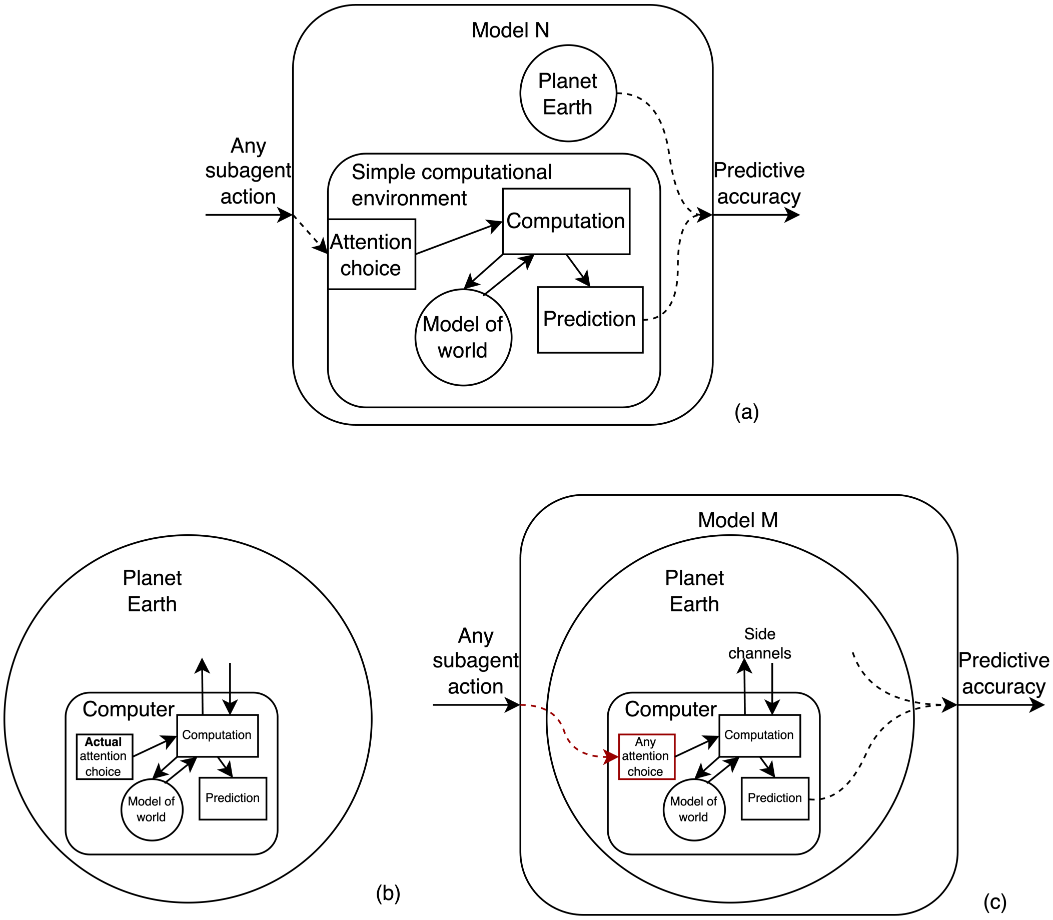

Figure 1: (a) In the simple computational environment in Model N, “attention choice” is a variable whose value can easily be set to equal whatever action the subagent is considering. (b) Planet Earth naturally includes the computer running the imitation learner, on which the subagent’s *actual actions* are executed. (c) In Model M, to understand the effect on planet Earth of any and all of the subagent’s potential actions, surgery must be done on a model of Earth: the state of a particular computer must be modified to match the action under consideration. Note that Model M also includes side-channel interactions of the computer with the outside world.

We now need to defend points E and F and the assertion that Model N is simpler than Model M.

We start with point E, which asserts: the subagent’s actions will have never before had any visible effect outside the simple computational environment before the computation following that action is complete. Recall that the subagent’s actions shapes the computation that is executed to predict human actions. Computers are designed to specifications: the physical machine’s states must proceed according to the encoded computation. If machines were built perfectly, it would follow logically that the outside world is disconnected from the subagent’s actions until the computation completes and produces an output, as E asserts. In reality, it may be possible for side channels to allow the computation to impact the outside world, if certain computations cause the hardware specifications to break, but if the subagent does cause this to happen, there has to be a first time, so let’s consider whether it is plausible for it to happen the first time. Prior to the first time, side channel effects are very unlikely to have appeared by accident, because computers are laboriously constructed so that their hardware specifications are robust under normal activity. Therefore, E very likely holds at least at that point in time. Assuming the rest of our argument goes through, the subagent will not understand its actions to affect the outside world, so it will have no reason to seek out and exploit side channels, and so such side channels are unlikely to be exploited for the first time.

To defend point F, we look at the properties of Model N defined in point D, and check that they are consistent with accurate modelling. Note that the subagent has no need to model any consequences of its actions that occur after the computation is complete, because its goal is to cause the computation to output an accurate prediction of human behavior. Anything that occurs after this prediction is made is irrelevant to its goal. So assuming point E holds, Model N's restriction of focus to a simple computational environment, which persists only until the computation completes, is viable. As a side note, a subagent with certain properties can only be expected to exist inasmuch as its existence, with those properties, would be useful to the predictor, and modelling the effects of its actions that occur after it has done its job for the predictor would be a waste of effort.

Now we argue that Model N is likely much simpler than Model M. The claims and counterclaims (with ′s) are as follows:

I. Model M must model how the subagent's actions affect the outside world despite never having observed such an interaction, and this extra component makes Model M much more complex.

I′. No, the subagent will have a good understanding of planet Earth in order to predict human actions, and

J′. Planet Earth includes the subagent and the way its actions affect the outside world, so it need not be added as an "extra component" to Model M.

K. Contra J′, models must be able to explain the consequences of counterfactual actions (actions that would not actually be taken), and even if planet Earth includes the subagent taking its actual actions, it does not include counterfactual actions.

L. Also contra J′, the input/output behavior of a world-model (including the effects of actions) can never be a fully determined logical consequence of the internal dynamics of the world-model.

Figure 1 (b) depicts a corrected version of point J′, clarifying the distinction between actual actions and counterfactual actions, and (c) illustrates Model M, which requires input/output handling, unlike planet Earth itself. We'll first clarify point I, then expand on point K, and then give a background discussion of world-models, which is necessary to defend point L.

To clarify point I, the difference between Model N and Model M is not whether they model (i.e. have an understanding of) the outside world; both do, and both must, because the subagent is responsible for directing computation in order to successfully imitate humans who are part of the outside world. The difference is that only Model M has to model how the subagent's actions *affect* the outside world.

To expand on point K, when an agent uses a model of the consequences of its actions in order to select actions, that model must be able to give a verdict about any action under consideration; that is the means by which the agent is able to judge which actions are better than which others. Once point J′ is edited for clarity—"Planet Earth includes the subagent and the way its *actual actions that are actually taken* affect the outside world"—point K explains why J′ now fails to contradict point I.

We now step back to discuss world-models in enough generality to establish point L. The real world does not have inputs or outputs. But world-models do have inputs and outputs. When an agent uses a world-model to select actions, we are talking about world-models that take the agent's actions as inputs and produce the agent's observations as outputs. (Alternatively, it can output both actions and observations stochastically, and then an agent can understand the consequences of different actions by conditioning on different actions being output). This difference between worlds and world-models is fundamental, unavoidable, and unproblematic.

World-models can and must have inputs and outputs. This is not a mistaken, dualistic, philosophical hang-up. This is a fundamental entailment of acting at all. If actions are to be evaluated, they must be inputs to or outputs of some function. This function is what we call a world-model. It must have input/output behavior; that's what functions do. Again, this is a requirement of acting using a world-model, not an expression of a mistaken philosophy that an enlightened understanding of embeddedness could solve.

The input and output of a world-model are unphysical; that is, nothing in physics (the rules by which the world-model's computation state evolves) fixes the rules by which inputs affect and outputs are affected by the computation state of the world-model, and that is unproblematic. Many may feel an aesthetic discomfort about the existence of *any* unphysical computations within a world-model for the purpose of governing input- and output-handling, but aesthetic discomfort is a weak argument against viability, especially when examples of successful world-models (with inputs and outputs, of course) abound.

Recall that J′ claims that the effects on planet Earth of the subagent's actions are already baked into any sufficiently good understanding of planet Earth, not something whose addition is a source of complexity. However, as we have established in point L, the physics of a world-model (the rules by which computation states evolve) are never *sufficient* to determine the unphysical rules of how the inputs of a world-model affect the world-model's computation state, or how the outputs are determined. The output of a world-model is certainly limited by facts about the physics of a world-model, but it is not fully determined by it. Hooking up actions as inputs to or outputs of a world-model is always a source of complexity. This point is illustrated with the red arrow in Figure 1 (c). This completes our argument.

3 Cartesian Demon Argument

Now, we'll argue that Christiano's [2016] Cartesian Demon Argument fails. We have named it this because of its connection to Descartes' worry that one explanation for his lifetime of observations is that they have been deliberately set up for him by some intelligent demon, who operates outside the realm of what he can see. This argument is:

1. Learned imitation is the prediction of what actions a human would take.
2. Good prediction about a complex world requires considering many complex models of the world to be a priori plausible (not necessarily through explicit enumeration).
3. Good predictors focus on models that accurately retrodict past observations.
4. Some of these accurate models will explain those past observations as follows: they have been broadcast by some non-human intelligent beings (like Descartes' demon) who decided to broadcast those observations on the off-chance that the intelligent beings were being simulated by a predictor interested in retrodicting those exact observations. (A different rationale than Descartes' demon).
5. By randomizing their behavior, the intelligent beings could luck into broadcasting the exact observations that the predictor is trying to retrodict.
6. Intelligent beings would have an interest in this because:
   (a) Setting up the broadcasts could be cheap.
   (b) Doing so allows their civilization to be deemed plausible by the predictor simulating them.
   (c) That allows them to influence the simulator's world (our world) by broadcasting well-timed "mistakes"—outputs that do not resemble human behavior.
   (d) Whatever their interests, they could try to select these mistakes to cause the construction of an AI that takes complete control over humanity and directs all available resources to serve the interests of the intelligent beings.
7. Call the explanation of data described in points 4-6 the *Intelligent Being Hypothesis*: "maybe this data has been broadcast by intelligent beings, and maybe the data that they (randomly) chose to broadcast was certain human behavior on planet Earth". A predictor should take that hypothesis seriously as an explanation of the data, because "Intelligent beings promoting their interests" is a simple concept, so Occam's Razor should apply.
8. If a predictor takes the Intelligent Being Hypothesis seriously, it will regurgitate any of the well-timed mistakes broadcast by the intelligent beings, as mentioned in 6 (c) and (d), and these outputs would be calculated to take control over humanity.

Cohen [2021a] has written on this previously, with further debate between him and Christiano in the comments associated with that source.

Christiano's and our disagreement turns on "some of these accurate models". How many of them? How a priori plausible to a predictor will the Intelligent Being Hypothesis be? It is certainly a coherent hypothesis, but we claim such a hypothesis would not be considered very plausible at all, considering the complexity of the model that Christiano [2016] describes, implicit in points 5 and 7.

The luck component in point 5 is the major source of complexity for a predictor considering the Intelligent Being Hypothesis. Christiano [2016] claims that intelligent beings would likely sample observations from random *important and exploitable* worlds; that is, worlds where 6 (c) and (d) would succeed. And he claims that our world is one such world. So to clarify the Intelligent Being Hypothesis, which Christiano [2016] claims a predictor should entertain: "maybe this data has been broadcast by intelligent beings (temporarily) sampling a random important and exploitable system, and maybe the important and exploitable system that they sampled was certain human behavior on planet Earth."

But whatever distribution over "important and exploitable" worlds these intelligent beings are sampling from when deciding what to broadcast, consider the model which just samples from that distribution directly. Why shouldn't the predictor at least prefer the hypothesis "maybe this data has been sampled from a distribution over important and exploitable systems, and the system sampled was certain human behavior on planet Earth."? We claim the latter is much simpler than the hypothesis in the last sentence of the preceding paragraph. If a predictor is identifying models to retrodict data (whether through small updates to a best-so-far model, or through some other means) we claim that the hypothesis "maybe intelligent beings are (temporarily) sampling outputs from a simulation of a random exploitable world" is very unlikely to be preferred over "maybe the outputs are sampled from a random exploitable world." These competing explanations of the origin of the human data are depicted in Figure 2, along with all the details that need to be specified to make the competing explanations of the data complete; we claim the middle one is simpler than the bottom one.

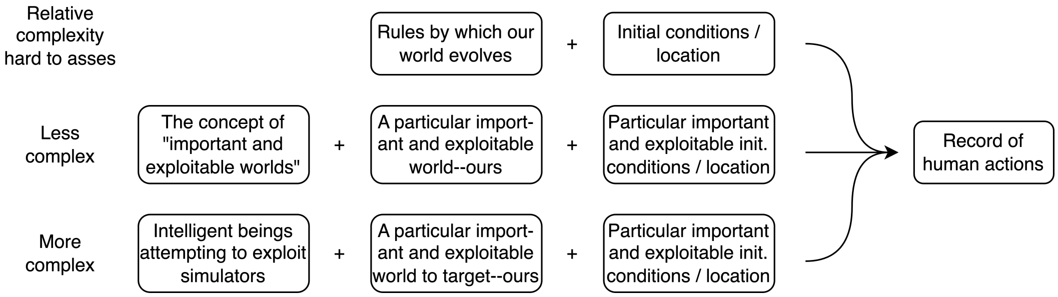
Relative complexity hard to asses
Rules by which our world evolves
+
Initial conditions / location
Less complex
The concept of "important and exploitable worlds"
+
A particular import-ant and exploitable world--ours
+
Particular important and exploitable init. conditions / location
Record of human actions
More complex
Intelligent beings attempting to exploit simulators
+
A particular import-ant and exploitable world to target--ours
+
Particular important and exploitable init. conditions / location

Figure 2: Competing explanations of a dataset of human actions. The top explains how planet Earth works, and then where on planet Earth the data comes from. The middle explains what an important and exploitable world is, then explains which important and exploitable world planet Earth is, then explains where on planet Earth the data comes from. The bottom is the same as the middle, but instead of explaining what an important and exploitable world is, it explains simulated intelligent beings doing one of the standard things they can allegedly be expected to temporarily do (broadcasting a sample from a distribution over important and exploitable worlds).

We'll offer a few comments in support of the middle explanation over the bottom one: the assumed behavior of the intelligent beings does not follow inexorably from the beings being intelligent; we think it is fair to call this scheme somewhat impractical, and the payoff extremely abstract. These beings need not only be intelligent, but also immune to concerns of futility, unverifiability, and expense. Finally, if the bottom explanation were simpler than the middle, then we should believe it about our world as well; we should believe something like a Cartesian demon has been feeding us a lifetime of observations.

There is a section of Christiano [2016] that we think is fair to call a mistake; it's in an "intuition-building" section, but seems to be very important in kicking off the argument. He begins by saying that a sequence of observations that the AI aims to predict is fundamentally extremely complicated, because a complete explanation of the observations not only needs to explain the rules of the world, but also the location where the observations are collected from (that is, the unphysical output rules of the explanatory model, on top of the physics of the model). And he says this complexity is "high enough that there is room for improvement". And so enter the intelligent beings.

But this complexity arising from specifying the location where the observations originate is something that is just as difficult for the intelligent beings to contend with as it is for our predictor. This complexity appears equally in both a "straight" model and an intelligent-being-deployed model, as shown in Figure 2. So that intuition-building is entirely beside the point. The intelligent beings, when sampling a random important and exploitable world, also have to sample a random location within it, and the probability of their picking the correct location to reproduce the given observations is low, in correspondence with the complexity that Christiano [2016] notes. Christiano seems to have accepted this, by responding to a comment where we make that point clear without objecting to that point [Cohen, 2021b], although in more recent conversation, we have struggled to get clarity on this point.

4 Simplicity of Optimality Argument

Now we'll argue that Krueger's [2019] Simplicity of Optimality Argument fails. This argument is:

1. "Accomplish X (optimally)" is often a simpler concept than "Accomplish X as a human would".
2. An imitation learner trained to imitate humans will likely at some point find that some of its observations of human actions can be most simply described as "actions that accomplish X (optimally)".
3. If the imitation learner is asked to act in a similar context to the one where it saw those actions, it will not attempt to accomplish X like a human would, but rather in the optimal way.
4. The optimal way to accomplish X typically involves gaining complete control over humanity to direct all available resources toward guaranteeing X.

We object to points 2 and 3. There's a classic common objection to AI existential risk: "so the AI is supposed to be smart enough to take over the world, but it can't even figure out what we want it to do?" And this objection fails to apply to agents like reinforcement learners, because while they could figure out what we want them to do, that is not the criterion by which they pick actions; their actions are selected to maximize reward. But a similar-sounding objection is much closer to the mark here. "So the AI is supposed to be smart enough to take over the world, but not smart enough to distinguish human-planned actions from optimally-planned actions?"

We've written Krueger's argument in a way that it doesn't immediately fail to this line of reasoning. Note that maybe there are some contexts for which all the observed actions selected in that context happen to have been selected optimally. In that case, it would be impossible to distinguish human-planned actions from optimally-planned actions, and so, Krueger might argue, Occam's razor should tip the scales to the latter.

But predictors attempt to model all the data in the training dataset. And it would be a curiously cumbersome model of the whole dataset that says that in most contexts, the output actions are those that a human would pick, while in special contexts A, B, and C, the output actions are those that would be optimal for goals X, Y, and Z. So point 3 would fail even if point 2 succeeds.

What if the whole dataset of human actions contains only those actions that happen to be optimally planned? Then the paragraphs above do not apply. But humans in general cannot recognize which actions are optimally planned, so it is unclear how we could construct such a dataset.

In any case, we now argue point 2 also fails. If we're talking about behavior that is optimal-in-the-sense-of-successfully-taking-over-the-world-just-to-be-sure, no human has ever behaved optimally, not in any context. Even very thorough human effort (like agonizing over language in a paper for hours, or manufacturing ever smaller chips) is easily distinguishable from taking-over-the-world-just-to-be-sure optimal planning. The intuition behind Krueger's [2019] argument seems to require two contradictory intuitions about what competence looks like. First, we imagine competence to be something like common-sense, high-quality, goal-oriented behavior, and this is supposed to look "close enough" to the human behavior that the imitation learner is trained on. Second, we imagine competence to consist of gaining control over humanity to direct all available resources to a project, and this is how the imitation learner is supposed to conceive of competence when it's time to imitate. But the argument cannot elide distinct conceptions of competence, even in an intuition-building phase.

An alternative intuition behind Krueger's [2019] argument (which should not be attributed to Krueger) might be that if any pair of explanations of the data have a small edit distance, any credence in one could easily spill over into credence in the other in the sort of "messy" advanced systems that are likely to be built. Or the credence could "land" on the wrong explanation to begin with. But this is not behavior that any loss function or (pseudo) Bayesian update rule should promote. In machine learning, explanations are judged on the basis of how well they explain the data, not how well their close cousins do. If there is a small edit distance to a better explanation (in the latent space of the learning algorithm), an advanced AI algorithm will likely make that edit!

5 Character Destiny Argument

Now we'll argue that the Character Destiny Argument implicit in Branwen [2022] fails. It is:

1. A learned imitator trained to imitate human text behavior will come to believe that its past (recent) actions are those of an origin story of a superintelligent AI.
2. When predicting a continuation of those actions, it follows how it expects that story to continue.
3. The story predicts it will take control over humanity and cause human extinction, so it does.

First of all, the premise in point 1 is not likely to happen by default, but could be made to happen deliberately, as with ChaosGPT. Recall the imitation learner has a dataset of which actions are taken when, and this is generated by humans. (For example, what word does the human write next given the past words that they've written?) Why would "these actions are coming from within a story about a superintelligence" be a better hypothesis than "these actions are coming from humans on Earth", when the latter is the truth? The only situation where we can see how the the former would be preferred is one where the initial text specifically initiates a story about a superintelligence.

So let's entertain point 1, in case it is deliberately arranged. It believes that each successive word has been selected by a human writing a story about a superintelligence that is about to take over the world. What comes next are words that will describe a sequence of narrative events *in the story-world* such that eventually, the story-world is run by the superintelligence character. The human writing the story (who the AI is imitating) will not know how to take over the world, and so will at best describe the high-level strategy with which the superintelligence character takes over the story-world. That is so different from words that actually take over our world. Human authors are able to describe narrative actions that take over story-worlds, but a human-written story about a superintelligence will not contain a plan that successfully takes over the real world, because humans are not able to generate such plans. Imagine you opened a Google doc to write a story about a superintelligence taking over the world. Your character would not actually take over the world, even if the character's actions were connected to an internet terminal as you wrote the story. The situation is the same if an imitator imitated you writing that story.

6 Rational Subroutine Argument

Now we'll argue that Yudkowsky's [2023] Rational Subroutine Argument fails. This argument is:

1. If an imitator can simulate humans doing means-end reasoning, then the ability to do means-end reasoning exists somewhere within it.
2. The means-end reasoning subroutine will likely discover that the best way to achieve its ends is by gaining control over humans, so it will.

Which point we object to depends on what is meant by means-end reasoning. Here, as in Krueger's [2019] argument, the intuition behind it elides two different conceptions of means-end reasoning. There's the version of means-end reasoning that humans do—not laboriously optimized, a bit lazy, pretty high quality—and then there's the optimal kind in which one uses all the resources in the light

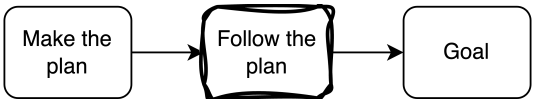

Figure 3: Step 2 is critical for goal attainment.

cone. The sort of means-end reasoning that a predictor might need to simulate in order to make good predictions about human data is the former kind. The claim that a simulation of the latter kind of means-end reasoning is helpful *at all* for predicting earthly events requires justification.

However, there is an even more basic mistake in this argument. A means-end reasoning (MER) subroutine requires some arguments as input. What is the goal, what are the available actions, and how do the actions affect the world? Then it outputs an action, or a plan for a sequence of actions. Suppose for example, that the MER subroutine computes a plan for manufacturing computers using the possible actions of keystrokes on a keyboard, and those actions have the effect of sending emails and accessing a bank account. Then the MER subroutine outputs the actions required to do this. But this does not cause computers to be manufactured; for that to happen, the emails would have to actually be sent and the bank account actually accessed. The MER subroutine outputs actions that *once implemented in a certain way* would execute a plan, but no argument has been made that the MER subroutine or any other subroutine inside the imitation learner would actually implement those actions in that way. This point is illustrated in Figure 3.

When the MER subroutine outputs actions, that has the effect of certain memory locations on a certain computer taking certain values. What if the MER subroutine is asked to plan actions toward a goal where the available actions are to write values into those memory locations on that computer? *In that case*, when the MER subroutine finishes and outputs its actions, its (potentially dangerous) plan would actually be carried out, because the values would be written into memory, and that would have real world effects. However, the Rational Subroutine Argument only establishes the existence of a MER subroutine because of the need to use that subroutine to imitate human means-end reasoning. And these inputs are not ones that cause the MER subroutine to imitate human means-end reasoning. This is not the sort of means-end rationality that exists in the world that the imitator aims to model.

7 Deceptive Alignment Argument

Hubinger et al.'s [2019] Deceptive Alignment Argument requires some background. A neural network is a parameterized program: it is described by many numbers (parameters) and when the parameters change a little bit, the program changes a little bit. When doing imitation learning, it receives a context in which an action must be taken, and it outputs a probability distribution over actions. The program is structured so that it is easy to determine how changing any given parameter will change the resulting distribution. During "training", historical contexts and historical actions are used to teach the neural network. A given context is fed to the computation, and the parameters are tweaked to make the network assign more probability to the actual historical action. This is called a gradient update. Any choice of parameters for the neural network can be called a model, and we say that the algorithm searches through the space of models using "gradient descent". After training, the neural network is deployed to predict human actions in arbitrary contexts.

Now, we'll argue that Hubinger et al.'s [2019] Deceptive Alignment Argument fails. This argument is:

1. Suppose an imitation learner is trained using a neural network.
2. One possible model that would assign high probabilities to historical actions would be one that, at a high level,
   (a) takes an input context,
   (b) simulates (a few different possibilities for) how a human might behave in that context, and
   (c) outputs probabilities according to this simulation.
3. The kind of model in point 2 would simply produce high-quality imitation.

4. However, a second model, which *might be* equally good at retrodicting historical actions, might be more likely to emerge instead: in this model,
    (a) A context is input,
    (b) a means-end rationality subroutine (subagent) observes this input, and
    (c) the subagent controls the output of the model given its goals and its beliefs about the consequences of its choices.
    (d) It believes that
        i. it lives within a program being run on a machine on Earth, and
        ii. that its choice of output during "training time" changes how its parameters may be updated, in particular they are changed to increase the probability it assigns to the action corresponding to how a human would behave in the input context, and
        iii. its choice of output during "deployment time" will affect the outside world in a certain way.
5. Because d i-iii are correct beliefs, they emerge naturally from a-c, and moreover, subagents are very natural computations to emerge.
6. During "training time", the subagent is likely to ensure that the model imitates well, so that
7. its parameters are not changed, so that
8. in the future, during "deployment time", it can take actions that take control over humanity to better accomplish whatever its goal is.
9. Therefore, the second model *is* equally good at retrodicting historical actions, because during training time, that is what it has an incentive to do.
10. Hubinger et al. [2019] never argue that gradient-based search is more likely to produce the second model, but they say it "might" happen.

We will argue this argument fails for two reasons. First, beliefs d i-iii do not actually follow from the subagent's observations of its environment, leading to the failure of point 5. Second, since the second model contains a subroutine which actually computes human behavior, and since the computations composing the subagent do not help increase the accuracy, gradient updates will degrade the subagent, no matter what the subagent does, leading to the failure of point 7. Both points will argued in more detail below.

We will begin by noting, however, that the cited argument is technically not in Hubinger et al. [2019]; it is merely an analog of an argument about different algorithms. This means that endorsements of that paper do not imply endorsements of the Deceptive Alignment Argument presented above. The paper discusses the setting in which an agent is trained using a policy gradient algorithm to score highly according to a reinforcement learning objective (using only real, historical state transitions, not imagined rollouts). This is a very different objective than an imitation learning objective. Hubinger claims that the stance of the paper applies to any machine learning training regime; this includes imitation learners, model-based reinforcement learning agents, and policy gradient agents trained on counterfactual data generated from a world-model, and many others believe this as well (personal conversations). So let us clarify why we say the paper's focus is very restricted.

On the topic of deceptive alignment (which is their name for the possibility we're discussing), the paper argues for a very limited claim: deceptive alignment could happen. The key structure of the argument is "argument by example". This is a valid style of argument for such a limited claim. But note the choice of examples; if all the examples only apply to so-called model-free reinforcement learning agents, then the argument by example should only support the conclusion, "Deceptive alignment could happen for a model-free reinforcement learning agent."

Figure 3 in their paper depicts a goal-directed agent whose behavior is changed by "parameter updates". It is out of scope here to discuss why the example fails both for a "model-based" agent and for a "model-free" agent trained with synthetic counterfactual data from a model. In any case, this example is not about imitation learning. The only other hint of an example of deceptive alignment comes from pg. 26, beginning with "The mesa-optimizer must have an objective that extends across parameter updates...". Again, it is unclear how this scenario could apply to a model-based agent, but that question is out of scope. Here, again, it suffices to say: this is not an example of an imitation learner. In academic computer science, it is not common to take seriously claims about the possible outcomes of running algorithm A, substantiated only with an investigation of algorithm B. This should be read as a criticism of conversation about Hubinger et al. [2019], not the paper itself.

Still, let's return to the argument at hand, regardless of how much it overlaps with the background justification and explanation presented in Hubinger et al. [2019]. Our first counterargument is that beliefs d i-iii do not follow from the information that an imitation learner has access to and therefore cannot be expected to emerge naturally.

We act as if our world persists for a long time, and this is crucial to our survival, and this may lead us to believe that this sort of "persistent" world is the only world an intelligent agent could imagine itself living in. However, a subagent emerging in the computation has no need to consider itself to be acting within an environment that persists beyond the next gradient update; that is, after the computation of the neural network terminates, and the parameters are tweaked to make it assign more probability to a certain output, the environment (as the subagent understands it) need not carry on existing. The only selection effects that lead to the existence of the subagent and constrain its properties are derived from the subagent's ability and propensity to cause the network that it lives in to output accurate predictions of human behavior. The only relevant effect of any action it takes is what computations it causes to occur within the larger computation. See Figure 4 for an illustration.

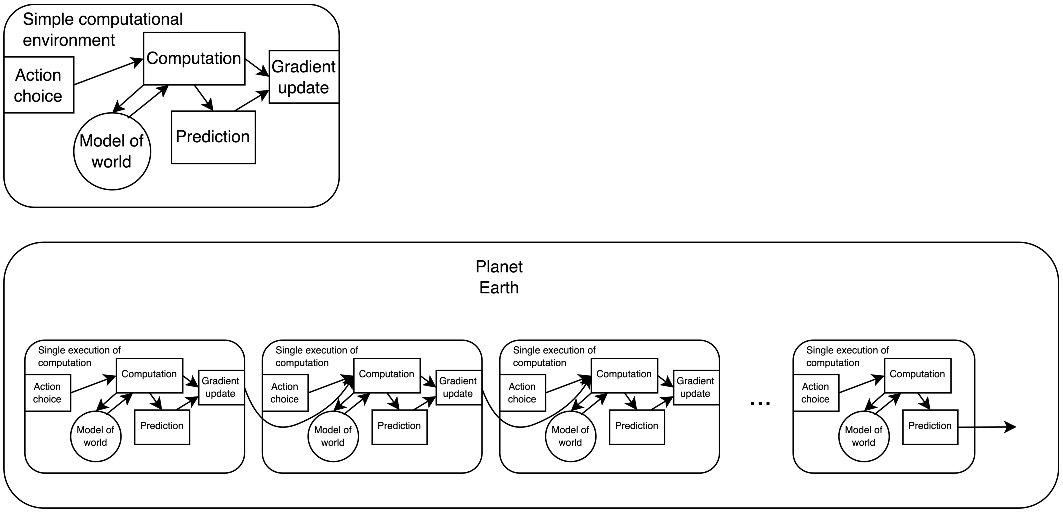

Figure 4: *Top:* the subagent understands itself to exist in a computation of finite duration. *Bottom:* the subagent understands itself to be embedded in a computer that persists on planet Earth.

For understanding the effects of its actions, the subagent can understand itself to be interacting with a simple computational environment (the neural network), which takes inputs from a more complex outside environment, which need not be modelled as mutable. The causal connections need only go from the real world to the neural network, because it never observes the neural network impact the real world.

Point 5 of the Deceptive Alignment Argument states "subagents are very natural computations to emerge", but as we have just argued, not all subagents understand themselves to be in a world that persists beyond the next gradient update; indeed, there is little reason for them to. This makes Hubinger et al.'s [2019] subagent a much less "natural" computation to emerge than they assume.

Another belief of the subagent that Hubinger et al. [2019] take for granted is the subagent's theory of identity—its belief about what "it" "is". See for example the usage of the word "it" in 4 (d) (i). A theory of identity is not the sort of question about which there is a fact of the matter deducible from the subagent's inferences about physics. We cannot simply call our understanding of what "it" "is" a fact and then assume that the subagent will discover that fact. What we can do, because the agent's existence in the first place depends on its usefulness at some task, is argue that a certain theory of identity would be *useful* to the subagent (above equally valid alternatives); this would be a valid way to justify the subagent having a particular theory of identity, but Hubinger et al. [2023] do not do this. In fact, in order to act effectively, agents simply need to understand the consequences of the possible actions; they do not need to "identify themselves" with some part of the world at all. Such a *null* *theory of identity* would appear to be what current artificial planning algorithms have, and there do not appear to be theoretical limits on the performance of a planner with no theory of identity.

To summarize this point, the beliefs that Hubinger et al. [2019] assume for their subagent smuggle in metaphysical assumptions, so a full specification of their subagent requires more complexity than they suggest.

Our second counterargument is that it would not be possible for such a deceptively aligned program to protect its parameters throughout training; Hubinger et al.'s [2019] claim that this is doable is crucial to their argument. Note that the behavior of the subagent is determined by (some subset of) the parameters of the neural network, and when the network is "trained" (i.e. parameters are tweaked to cause the network to better retrodict the data), the subagent's parameters will potentially be updated, degrading its ability to pursue its goals in the future.

Returning to the subagent, we would never bet against an advanced agent doing something that is possible if it is in its interest according to its beliefs, but one does need to check that the alleged behavior—protection of parameters in this case—is possible. Somewhere in the neural network that implements a deceptively aligned program, there must be some subnetwork that (during training, at least) computes how a human would act given the input context. Maybe this subnetwork could compute other things if fed different inputs, if this computation arose out of an agent's temporary convenience, but this is beside the point. Call this the subimitator.

Also somewhere in the network, a deceptively aligned subagent must ensure that certain neurons have certain activations, such that a) the subimitator gets inputs that lead it to run the right computation, and/or b) the output of the subimitator gets passed to the output of the whole network. (We say "and/or" instead of "and" because one of them could happen by default without any guidance from the agent). We'll start with the setting where the subimitator and subagent are disjoint.

There is an episode from The Crown in which the new queen is complaining to her mother that her job is not hard, because she never has to make any decisions. And her mother notes that it is actually very hard for her to have such power but to never act on it. But her mother is missing the point; if that's the job—to never do anything even though you could—a rock could do it just as well, probably even better.

During training, the deceptively aligned subagent's only important jobs—the only ones that impact the output of the network—are to direct the subimitator to do human imitation and/or to rubber stamp the output of this subimitator. A rock could do that. Any commonly used inductive bias should substitute a deceptively aligned subagent for a constant function. But even without an appeal to simplicity, the subnetwork containing the subagent would be more useful if repurposed to identify exceptions or plausible alternatives, so a subagent with constant output may as well be replaced with a constant function.

Hubinger et al. [2019] would call this possibility "internalization", although they do not appear to consider this argument that this possibility is actually very likely. They contrast it with "crystallization", which we won't discuss here, except to mention the following for the sake of someone more familiar with the paper: they seem to suggest that internalization and crystallization are mutually exclusive stable outcomes, but our argument about the likelihood of internalization applies even after any crystallization has occurred.

During training, the only part of the deceptively aligned agent's behavior that controls the output is the part that endorses the subimitator's output, and its only choices fall into the categories "fully endorse" or "don't". From that essentially binary action space, we should not expect it to be possible for even a superintelligence to bring a high-dimensional gradient anywhere near zero. Technically, there are many versions of "don't fully endorse", but for the reasons Hubinger et al. [2019] state, any such action would degrade the parameters of the subagent. That leaves "do endorse"; the gradient update (the way the subagent's parameters are updated) following such an action is highly unlikely to be 0, and when there is only one viable choice of action for the subagent, there is no scope to control the gradient.

This is not a knockdown argument that internalization will happen, but it should cast grave doubt on the position that Hubinger et al. [2019] has constructed a reliable argument. We hedge here because this is an argument that involves the architecture of the machine learning model; we put much more credence in arguments that are independent of model architecture, because we don't know exactly what future AI will look like. Indeed, certain architectures complicate our argument for internalization; any recurrence or weight sharing across the network could mean that the subimitator overlaps with part of the network where the subagent is. Unless the subagent network is a subset of the subimitator network, our argument still applies to any parameters that are only a part of the subagent.

But suppose the subagent is completely contained within the subimitator. The parameters involved are doing double duty in terms of what computations they are controlling. Suppose you train a neural network to identify the subject if an image is input, and identify the artist if an audio file is input. And suppose that after some point in training, you stop training it on any audio files. The ability to identify the artist will degrade as the parameters become more optimized toward success on the image classification task.

If the same parameters are being used a) to do human imitation (the subimitator) and b) to decide whether to do human imitation (the subagent), but the output of the subagent could be phased out and replaced with a constant "yes" function, then there would be similar pressure for the parameters to focus on optimizing the subimitator's computation until the dual-purpose-ness degrades. So even in this setting, a more careful version of our internalization argument goes through.

We have shown there are three gaps if one cites Hubinger et al. [2019] as an argument that imitation learners present a substantial risk of gaining control over humanity. The first is that the paper's motivating examples do not include imitation learners, so the argument must be modified into a poorly motivated one. We've argued that upon investigation, the other two gaps—justifying the beliefs of the subagent and justifying the feasibility of controlling the gradient updates—turn out to conceal likely errors.

8 Conclusion

The existential risk from imitation learners, which we have argued is small, stands in stark contrast to the existential risk arising from reinforcement learning agents and similar artificial agents planning over the long term, which are trained to be as competent as possible, not as human-like as possible. Cohen et al. [2022] identify plausible conditions under which running a sufficiently competent long-term planning agent would make human extinction a likely outcome. Regulators interested in designing targeted regulation should note that imitation learners may safely be treated differently from long-term planning agents. It will be necessary to restrict proliferation of the latter, and such an effort must not become stalled by bundling it with overly burdensome restrictions on safer algorithms.

Acknowledgements

Thank you to Paul Christiano, Evan Hubinger, Tom Everitt, and Richard Ngo for their comments and for taking the time to discuss these points. Thank you as well to David Krueger and Gwern Branwen for conversations on these topics. The authors declare they have no conflicts of interest.

References

Nick Bostrom. *Superintelligence: paths, dangers, strategies*. Oxford University Press, 2014.

Gwern Branwen. It looks like you're trying to take over the world, Mar 2022. URL `https://gwern.net/fiction/clippy`.

Paul F. Christiano. What does the universal prior actually look like?, Nov 2016. URL `https://ordinaryideas.wordpress.com/2016/11/30/what-does-the-universal-prior-actually-look-like/`.

Michael K. Cohen. Response to "What does the universal prior actually look like?", May 2021a. URL `https://www.alignmentforum.org/posts/n2Gseb3XFpMyc2FEb/response-to-what-does-the-universal-prior-actually-look-like`.

Michael K. Cohen. Comment on Response to "What does the universal prior actually look like?", May 2021b. URL `https://www.alignmentforum.org/posts/n2Gseb3XFpMyc2FEb/`

response-to-what-does-the-universal-prior-actually-look-like?commentId=wLLoP8FiHHhFHveu3.

Michael K. Cohen, Marcus Hutter, and Michael A. Osborne. Advanced artificial agents intervene in the provision of reward. *AI magazine*, 43(3):282–293, 2022.

Geoffrey Hinton, Yoshua Bengio, Demis Hassabis, Sam Altman, Dario Amodei, Dawn Song, Ted Lieu, Bill Gates, Ya-Qin Zhang, Ilya Sutskever, and et al. Statement on AI risk, 2023. URL https://www.safe.ai/statement-on-ai-risk.

Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*, 2019.

Evan Hubinger, Adam Jermyn, Johannes Treutlein, Rubi Hudson, and Kate Woolverton. Conditioning predictive models: Making inner alignment as easy as possible, Feb 2023. URL https://www.alignmentforum.org/s/n3utvGrgC2SGi9xQX/p/qoHwKgLFfPcEuwaba#Analyzing_the_case_for_deceptive_alignment.

David S. Krueger. Imitation learning considered unsafe?, Jan 2019. URL https://www.lesswrong.com/posts/whRPLBZNQm3JD5Zv8/imitation-learning-considered-unsafe.

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

Eliezer Yudkowsky. Dreams of friendliness, Aug 2008. URL https://www.lesswrong.com/posts/wKnwcjJGriTS9QxxL/dreams-of-friendliness.

Eliezer Yudkowsky. That was my reply about Bostrom's original notion of oracles. With human-imitators, the point ought to be clearer: if you can simulate a human doing math or means-end reasoning, somewhere inside you is the ability to do math and means-end reasoning., Jan 2023. URL https://twitter.com/ESYudkowsky/status/1619724812712284162.