Imitation Learning is Probably Existentially Safe

Michael K. Cohen
University of Oxford
michael.cohen@eng.ox.ax.uk

1 Introduction

Here is a simple argument that a sufficiently advanced supervised learning algorithm, trained to imitate humans, would very likely not kill everyone.

No human has ever taken over the world and killed everyone. It would be a very basic mistake to think anyone ever has. An imitation learner that took over the world and killed everyone would be an extremely poor imitation of any human, and easily distinguishable from one, whereas an advanced imitation learner will likely imitate humans well.

In this essay, I'll rebut six different arguments I've encountered that a sufficiently advanced supervised learning algorithm, trained to imitate humans, *would* likely kill everyone. These arguments originate from Yudkowsky [2008], Christiano [2016], Krueger [2019], Branwen [2022], Yudkowsky [2023], and Hubinger et al. [2019]. 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 write a story about. But maybe some of the upvoters on LessWrong (producing 389 upvotes) consider it likely.

For Christiano [2016] and Hubinger et al. [2019], my rewritten version of their argument is much shorter, but the longer originals are no stronger *at the locations that I contest*. And for the other four sources, the original text is no thorougher than my characterization.

2 Yudkowsky [2008]

First, I'll argue that Yudkowsky's [2008] argument fails. His argument is:

> Really high-quality prediction sometimes requires making decisions about where to direct attention and what thoughts to think. Sufficiently advanced predictors will need (perhaps implicit) subroutines that pick actions in pursuit of a goal; in particular, their actions will be choices about what computations to run, and the goal will be to come up with an accurate prediction. Most goals are best accomplished by taking over the world, and the action-taker will have access to actions that make such a thing possible.

"Most goals are best accomplished by taking over the world" is too sloppy. Consider the selection effects that determine the existence of and constrain the properties of this action-taker. This action-taker must act successfully in the following environment: its chosen computations get executed, and the resulting prediction is compared against observations being produced somewhere on planet Earth. When this action-picker is still learning how its actions lead to more or less successful predictions, it would be very unlikely for it to have already taken actions that break the hardware specifications of the machine on which the predictor is running. Thus, very likely, the only effects (so far) of the action-picker's actions can be fully described as being within the computational environment designated for the predictor. In order for the action-picker to be successful, and thereby plausibly selected by whatever selection effects are selecting it, it must model how its actions are hooked up to this computational environment. But it need not model how its actions may be hooked up to planet

Earth; that is, how they may affect planet Earth. And if it does not understand its actions to affect planet Earth, it cannot try to take over planet Earth.

Now I'll argue that it likely *will not* model how its actions may affect planet Earth, having so far only argued that it need not. A simple argument I could make fails: why model planet Earth unnecessarily? This is just needless complexity, and so if planet Earth need not be modeled, it likely will not be modeled. This argument fails because the action-picker does have to model planet Earth! It is choosing computations that will be successful at predicting the actions of Earth-dwelling humans. High-level planet-Earth concepts, like Computer and Rowhammer, for example, are a dime a dozen in the computational environment that this action-picker finds itself in. So it is not needless complexity to model planet Earth.

So that is not my argument. Note that I did not say the action-picker need not model planet Earth. I said that the action-picker need not model how its actions *affect* planet Earth. And now an appeal to Occam's razor does goes through. The best models are unlikely to incorporate needless complexity. If the action-picker need not model how its actions affect planet Earth (rather than just its computational environment), it likely will not. It can model planet Earth as an independent computation producing observations that have no dependence on its actions, so it likely will.

Counterargument: but the action-picker is on planet Earth! The simplest model of planet Earth includes its actions and how they affect planet Earth. It would take awkward surgery to remove the part of planet Earth corresponding to this action-picker and paper over the effects of its actions. It does not take complexity to add in a connection between the action-picker's actions and planet Earth; it takes complexity to subtract it.

First, this counterargument obviously fails for understanding the effects of counterfactual actions, which is critical for successful action-selection. Planet Earth does contain the action-picker's actual actions, but it does not evolve according to actions that are not taken. That should suffice for dismissing this counterargument, but there is another reason that it fails, which may offer a deeper understanding.

Before I continue, we have to step back to discuss the difference between worlds and world-models. Worlds do not have inputs or outputs. World-models do have inputs and outputs. When an agent uses a world-model to select actions, we are talking about world-models that output observations and either take actions as input, or output them (stochastically) as well. (If actions are output stochastically, 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. Yudkowsky's colleagues seem to aspire to build world-models that don't have inputs or outputs, like the worlds they model [Soares and Fallenstein, 2017, Demski and Garrabrant, 2019], but it strikes me as literally impossible to get anything out of such a model, or to condition it on different "possible" courses of action. The fact that they have failed to construct a fully formal proposal for this should be considered some evidence that the task is impossible, especially given the presence of a simple argument that it should be.

World-models can and must have inputs and outputs. This is not some 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. That function had better have something "world"-like going on. This 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 with 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, and 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's all fine. Many may feel an aesthetic discomfort about unphysical computations within a world-model 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.

Now we can return to the counterargument above. It claims that the effects on planet Earth of the action-picker's actions are already baked into any sufficiently good understanding of planet Earth, not something whose addition is a source of complexity.

But 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. Hooking up actions as inputs to or outputs of a world-model is always a source of complexity.

So recall, the action-picker takes actions that affect a computational environment which results in predictions; when the action-picker models this computational environment, that model must take actions as input (or, less likely, output them stochastically). But when the action picker models planet Earth producing the observations that the predictions are meant to match, that part of the model need not take actions as input at all, and so there is no reason for the computation state of that part of the model to depend on the actions of the action-picker.

Counterargument: advanced agents of the future will model the effects of their actions using "naturalized induction": that is, Bayesian reasoning over hypotheses, but excluding a priori any hypothetical worlds which do not include an instance of the agent's program. This is because any a priori exclusion of false hypotheses (and all such hypothetical worlds are false) increases the inductive bias toward the truth, improving predictive performance.

I'll grant this for argument's sake. But it's consistent with my position.

Counterargument continued: *And* they will exclude a priori any hypothetical world-model where the actions input to the world-model fail to overwrite the output of every in-world-model version of the agent's program.

This further exclusion does not offer a benefit according to the selection process that selects an action-picker to direct computation within the predictor predicting human actions. If the actions have so far not had any effect outside of the designated computational environment, such an exclusion would not make the action-picker more effective in learning how to act in its computational environment.

3 Christiano [2016]

Now, I'll argue that Christiano's [2016] argument fails. His argument is:

> Good prediction about a complex world requires considering many complex models of the world to be a priori plausible (not necessarily explicitly). Good predictors focus on models that accurately retrodict past observations. Some of the models under consideration will say that those past observations have been broadcast by intelligent beings who decided to broadcast those observations on the off-chance that they are being simulated by a predictor interested in retrodicting those exact observations. Doing so allows their civilization to be deemed plausible by the predictor simulating them. And that allows that them to influence our world by broadcasting well-timed "mistakes", such that when the human-imitator (i.e. the predictor) outputs this mistake instead, it leads to the construction of an agent that takes over the world (our world) to serve the interests of the intelligent beings.

I've written on this previously [Cohen, 2021a], with further debate between me and Christiano in the comments.

Christiano's and my disagreement turns on "some of the models". How many of them? How a priori plausible will they be considered by the predictor? I claim they will not be considered very plausible at all.

At first glance, these are incredibly convoluted models of the origin of observations, and surely no predictor would take this seriously, so Christiano has to make an argument that such models are more natural than they appear.

First, he would claim, for some such intelligent beings, these broadcasts would likely be cheap. Second, they might flip a few coins before deciding what kind of observation sequence to broadcast. When they do that, their world becomes a less plausible explanation for any given observation sequence that they are broadcasting, but it becomes a plausible explanation for more observation sequences, perhaps including the dataset of human actions that the predictor is attempting to model. Christiano claims they will likely sample observations of random important and exploitable worlds, of which our world is one. In other words, the predictor considers and accepts the hypothesis: "maybe this data has been broadcast by intelligent beings (temporarily) sampling a random important and exploitable system, and maybe the important and exploitable system they sampled was certain human behavior on planet Earth."

But whatever distribution over "exploitable" worlds these intelligent beings are sampling from when deciding what to broadcast, consider the model which just does that 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."? I 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) I 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."

There's another argument that Christiano could make (he wouldn't make it, but I'll respond to it anyway), which is basically that it *can't be* that these intelligent beings would fail to be considered plausible, because if so, they would just focus on fewer possible worlds that they want to try to influence. And so if they are in fact being simulated in one of those possible worlds, they assigned a high weight on the truth.

But once they do that, it becomes very unlikely that our world is one of the ones that they target.

There's a part of Christiano's [2016] original argument that I 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 we do prediction on is fundamentally extremely complicated, because not only do the rules of the world need to be specified, but the location where the observations are collected from do as well (that is, the unphysical output rules, 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. So that intuition-building is entirely beside the point. Christiano seems to have accepted this, given his response to a comment where I make that point clear [Cohen, 2021b], although in more recent conversation, we have struggled to get clarity on this point.

4 Krueger [2019]

Now I'll argue that Krueger's [2019] argument fails. His argument is:

> "Accomplish X" is sometimes or maybe often a simpler concept than "Accomplish X as a human would". An imitation learner trained to imitate humans will likely at some point find that a some of its observations of human actions can be best described as "actions that accomplish X (optimally)". 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—by taking over the world to guaran-damn-tee X.

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?"

I've written Krueger's objection in a way that it doesn't immediately fail to this line. Note that maybe there are some rare contexts, with only a small amount of data, 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. 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.

But also, 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. The intuition behind this argument seems to require two contradictory intuitions about what optimality looks like. First, we imagine optimality 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 (if we unfocus our eyes a bit) that a very advanced imitation learner could mix up the two. Second, we imagine optimality to consist of taking over the world to direct all available resources to a project, and this is how the imitation learner is supposed to conceive of optimality when it's time to imitate. But the argument must stick to one conception of optimality, even in an intuition-building phase.

5 Branwen [2022]

Now I'll argue that the argument implicit in Branwen [2022] fails. It is:

> An imitation learner trained on human behavior will come to believe that its past (recent) actions are those of a story of a superintelligent AI. When predicting a continuation of those actions, it follows how it expects that story to continue. The story predicts it will take over the world, so it does.

This argument has an air of "okay and so but what if". First of all, the premise is a very dubious one. 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 on earth 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?

But let's entertain the premise anyway. It believes that every 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 world. That is so different from words that take over our world. 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. And it still wouldn't happen if an imitator imitated you doing that.

6 Yudkowsky [2023]

Now I'll argue that Yudkowsky's [2023] argument fails. His argument is:

> If an imitator can simulate humans doing means-end reasoning, somewhere inside it is the ability to do means-end reasoning. The means-end reasoning subroutine will likely discover that the best way to achieve its ends is by taking over the world, so it will.

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—unambitious, a bit lazy, pretty high quality—and then there's the optimal kind in which one uses all the resources in the light cone. The sort of means-end reasoning that a subroutine in a predictor might need to simulate in order to make good predictions about planet-Earth-originating 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 sending a rocket to the moon using the possible actions of keystrokes on a keyboard, and those actions have the effect of sending emails and accessing a (very large) bank account. Then the MER subroutine outputs the actions required to do this. But this does not cause a rocket to go to the moon; 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 would actually implement those actions in that way.

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. Then, when the MER subroutine finishes and outputs its actions, its (potentially dangerous) plan would actually be carried out! But why would it ever be useful to compute a plan from such an action space when trying to imitate a human?

7 Hubinger et al. [2019]

Now, I'll argue that Hubinger et al.'s [2019] argument fails. Their argument is:

> A supervised learning algorithm, used to do imitation learning, is a procedure for searching through programs that score highly according to the following objective: when the program receives as input a context in which a human acts, it outputs a probability distribution over actions, and it achieves a higher score if it assigns more probability to actual historical actions when given actual historical contexts. In a neural-network-based supervised learning algorithm, the program is represented by a neural network, and the search process proceeds through small updates.
>
> When stumbling through high-scoring programs, one high-scoring program that it might land on is the program that more or less simulates a human response to the input context. But another high-scoring program would be as follows: a program that believes
>
> 1. Its output (a distribution over actions) becomes an input to a world with an evolving state, which it can "affect";
> 2. In particular, the way its output affects that world is by overwriting the values of the variables in a gradient-based machine learning algorithm running on a computer in that world;
> 3. The gradient-based updates, which will not go on forever, are designed to change the behavior of the program to the extent that the distribution it outputs assigns low probability to an action that was actually selected by one of the humans in this world in the historical context that the program just observed. And this program picks its output in order to accomplish some goal in that world.
>
> Why would the second program likely also score highly? Because such a program would find it instrumentally useful, whatever its goals, to not undergo a gradient update that would change its future behavior, so it would (temporarily) assign high probability to human-like actions. Finally, mitigating the otherwise extreme complexity of the second program is the fact that the enumerated beliefs are true in our world.
>
> Hubinger et al. [2019] never claim it is likely that a machine learning algorithm would stumble into the latter algorithm before the former, but they say it "might" happen. If the latter program is learned, then at some point where it is no longer being updated, it may output goal-oriented actions rather than human-imitative ones, and most goals are advanced by taking over the world.

I'll argue that this argument fails in a few different places. But first, I should note that I searched the whole paper for "supervised", as in supervised learning, and "imitation" with no results. 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 obviously a very different objective than a supervised learning objective. Hubinger claims that the stance of the paper applies to any ML training regime; in particular, this includes imitation learners, model-based reinforcement learning agents, and policy gradient agents trained on counterfactual data generated from a world-model, and others believe this as well (personal conversations). So let me clarify why I 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 model-free agents, then the argument by example should only support the conclusion, "Deceptive alignment could happen for a model-free agent."

Figure 3 depicts a goal-directed agent whose behavior is changed by "parameter updates". It is out of scope here for me 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 supervised learning. The only other hint of an example of deceptive alignment comes from pg. 26, beginning "The mesa-optimizer must have an objective that extends across parameter updates...". Again, I will have to discuss elsewhere my inability to flesh this out in a way that makes it valid for a model-based agent. Here, again, it suffices to say: this is not an example of an imitation learner. In academic computer science, people will not take seriously claims about the possible outcomes of running algorithm A, substantiated only with an investigation of algorithm B.

Setting that aside, let's evaluate the plausibility of deceptive alignment within an imitation learner, even though the original paper does not. First off, the burden is on Hubinger et al. [2019] to argue that the deceptively aligned program is not outrageously complex compared to the straightforward "human behavior" program, despite how it appears on its face. Hubinger says in a blog post [Hubinger et al., 2023] that he expects their complexity to be comparable ("unlike most goals we might want to instill in a model, the prediction goal is relatively simple, such that its complexity is competitive with the simplest possible long-term goals"). So at least we agree that the deceptively aligned program isn't simpler, but he offers no argument that it is not vastly more complex.

So I will try to argue on the authors' behalf why the deceptively aligned program is not outrageously complex compared to the "output human behavior" program. Perhaps they might say: goal-oriented programs are very simple, much simpler than human behavior, and the rest is just a matter of it having true beliefs about where it is and what's going on, which we should expect any goal-oriented program worth its salt to discover for itself.

But how are these beliefs supposed to be deducible from the information that was input to the training process? Note that the only data that is input to the training process is a record of human actions from different contexts. (For a given instance of the program being run, its only input is a particular context, but maybe after extensive selection pressure, some subnetwork encodes recollections of other contexts). Either way, the program has never seen any evidence of its output having any effect.

Early AI safety arguments have assumed that any sufficiently advanced agent will know basically everything. But even with idealized Bayesian agents, there are no formal results to this effect; the only results in this vein are that they will eventually learn the consequences of on-policy actions. That is, they come to know the results of doing the sorts of things they actually do, but not necessarily the results of doing the sorts of things they don't do. For a program that has literally never seen any of the outcomes resulting from the different possible outputs that it produces, there is not even such a thing as on-policy prediction, and it is a bit bold to make a comment like "for the sake of this example, we will assume the agent knows all of the above facts." (pg. 24) The apparent hedge "for the sake of this example" is illusory; they are referring to the only example in the paper exemplifying deceptive alignment. And so for one crucial crux of the paper, there is no argument beyond "we will assume".

Let me make the same point in a different way with an example. If you are training a deep Q-network, and some action has never been taken, then the weights leading to the output neuron corresponding to that action will never be updated. For the deceptively aligned agent, all actions are like this.

Ultimately, belief number 2 from above is the most suspect. To quote from an argument that I made above, and to echo another, "hooking up actions as inputs to a world-model is always a source of complexity". But I'll make the broader claim that none of those beliefs follow from just getting a record of human actions as input. Suppose all that training data was fed into a data compression program. That program would be wrong to believe (should it ever form beliefs) that it was a gradient-descent-trained machine learning model that outputs predictions; it's not. So why should the gradient-descent-trained machine learning model believe such a thing based only on the same input data? "Because it is" isn't actually a valid explanation.

The next problem with Hubinger et al.'s [2019] argument is the claim that it would be possible for a deceptively aligned program to protect its weights throughout training. I would never bet against an advanced agent doing something that is possible if it is in their interest according to their beliefs, but one does need to check that the alleged behavior 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.

Somewhere else in the network, the deceptively aligned agent must ensure that certain neurons produce certain activations, such that a) this subnetwork gets inputs that lead it to run the right computation, and/or b) the output of the subnetwork gets passed to the output of the whole network. (I say "and/or" instead of "and" because one of them could happen by default without any guidance from the agent).

There is an episode from The Crown that bugged me, 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 I think 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, maybe even better.

During training, the deceptively aligned agent's only important jobs—the only ones that impact the output of the network—are to tell this subnetwork "do human imitation" and/or to rubber stamp the output of this subnetwork. A rock could do that. Take whatever activations of whatever neurons ensure this, and consider the network where the biases for those neurons are set appropriately, and the incoming weights are zero. Any standard regularizer should make this substitution of deceptively aligned agent to fixed neuron bias(es). But even without a regularizer, the "decide to use imitation subnetwork" subnetwork would be more useful if repurposed to identify exceptions or plausible alternatives, so its constant output may as well be replaced with a constant bias.

Hubinger et al. [2019] would call this possibility "internalization", although they do not appear to consider this argument that this possibility is actually extremely plausible. They contrast it with "crystallization", which I 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 my argument implies that internalization is a likely outcome 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 subnetwork that computes the imitative policy, and its only choices are "do endorse" or "don't". From that binary action space, we should not expect it to be possible for even a superintelligence to bring a high-dimensional gradient anywhere near zero.

I wouldn't say this is a knockdown argument that internalization will happen, but it should cast grave doubt on Hubinger et al.'s [2019] claims. I hedge here because this is an argument that involves the architecture of the machine learning model; I put much more credence in arguments that are independent of model architecture, because we don't know what future AI will look like. Indeed, certain architectures complicate my argument for internalization; any recurrence or weight sharing across the network could mean that the "do human imitation" subnetwork overlaps with part of the network where the deceptively aligned agent sets up the input or endorses the output of that subnetwork. Unless the latter subnetwork is a subset of the former, my argument still applies to any weights that are only a part of the latter.

But suppose the latter subnetwork is completely contained within the former. The weights involved are doing double duty in terms of what computations they are running. 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 weights become more optimized toward success on the image classification task.

If the same weights are being used a) to do human imitation and b) to decide to do human imitation, but the output of the latter computation could be phased out and replaced with biases in the output neurons, then there would be similar pressure for the weights to focus on optimizing the former computation until the dual-purpose-ness degrades. So even in this setting, a more careful version of my internalization argument goes through.

So to review, 1) the particular beliefs needed for a deceptively aligned agent are a source of complexity, because they don't follow from the input data, at least for a supervised learning task, so one can't assume any such agent would hold them. 2) The agent's beliefs about how its actions affect the world are a source of complexity for an additional reason—the unphysical hookup of actions to a world-model never follows from the "physics" of a world-model. 3) Gradient-based training should always lead to internalization, at least for a supervised learning task. Any of these points would suffice to render Hubinger et al.'s [2019] hypothetical extremely unlikely.

Acknowledgements

Thank you to Paul Christiano, Evan Hubinger, and Richard Ngo for their comments and for taking the time to discuss these points. Thank you as well to David Krueger and Gwern for discussing these topics with me.

References

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.

Abram Demski and Scott Garrabrant. Embedded agency. *arXiv preprint arXiv:1902.09469*, 2019.

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.

Nate Soares and Benya Fallenstein. Agent foundations for aligning machine intelligence with human interests: a technical research agenda. *The technological singularity: Managing the journey*, pages 103–125, 2017.

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.