Generative Theory of Mind and the Value Misgeneralization Problem

Paul de Font-Reaulx

Department of Philosophy
University of Michigan, Ann Arbor
pauldfr@umich.edu

February 20, 2023

Abstract

In this paper, I do two things. First, I argue that any interaction with an artificial general intelligence (AGI) will robustly lead to a catastrophic form of goal misgeneralization, unless the AGI has the ability to reliably predict human preferences based on limited information. I call this the value misgeneralization problem. Second, I propose a new path to solving the problem. Specifically, I argue that human values have a hierarchical structure that we regularly use to predict each other's preferences over arbitrary outcomes. I call this ability our generative theory of mind. Using this fact about our psychology, an AGI could similarly predict our values, which would solve the value misgeneralization problem. To foster this ability in artificial systems, however, we need an empirically informed and computationally precise theory of human values. In the final section, I attempt to sketch the beginning of such a theory, and conclude that the path to solving the alignment problem might crucially go via cognitive neuroscience.

1 Introduction

The alignment problem is, roughly, the problem of getting a powerful artificial agent to do what we want.[1] It is a hard problem, partly because it is difficult to communicate exactly what we want. Another reason, however, is that we

[1]For overviews, see Bostrom (2014), Russell (2019), Ngo (2020) and Christian (2021). See also Railton (2020), Hubinger (2020), Gabriel (2020), Carlsmith (2022), Yudkowsky (2022), Christiano (2022), Krakovna (2022), Ngo, Chan, and Mindermann (2022), and Hendrycks and Mazeika (2022), among others.

realistically can't communicate our values regarding every possible outcome that we care about. Some of our values will remain unexpressed, simply because there are so many possible outcomes to have a view on. This means that the agent will need to fill in those gaps itself, and predict what we would think about outcomes that we haven't communicated anything about. But it might not do so perfectly. Consequently, it might predict that we are indifferent about some outcomes that we strongly dislike, and tragically bring those about. This is especially likely if such states will help it realize those goals we have successfully expressed, for example by deceiving us to prevent us from stopping it.[2]

In other words, the alignment problem can arise not only from misspecified values, but also from accurately specified values that get generalized inaccurately. This cause of misalignment is a kind of goal misgeneralization, which is a problem identified in the context of reinforcement learning models by Langosco et al. (2022).[3] In this paper, I will focus on the special case of goal misgeneralization where the agent is an artificial general intelligence (AGI) of arbitrary capability, and the goals that we attempt to instill is the totality of our values. I call this version of the problem the *value misgeneralization problem*. As I argue below, the value misgeneralization problem is very difficult to avoid, and will reliably produce misaligned outcomes of arbitrary severity.[4] In other words, solving it is necessary for—and would constitute major progress towards—solving the alignment problem.

In what follows, I argue that there is a neglected path to such a solution. The crux of the issue is that there appears to be no way for an AGI to reliably predict our preferences when we haven't communicated them. But this is true only if we neglect important facts about our psychology, namely that our values—understood as mental states—are organized in a hierarchical structure. This is to say that we value some outcomes A *because* we value other outcomes B, and believe that A are likely to lead to B. We humans have learnt to use this hierarchical structure to predict each others' preferences over hypothetical outcomes. I call this ability our *generative theory of mind*. If an AGI also learns the structure of our values and develops a generative theory of mind for humans, then it too could reliably predict our uncommunicated preferences. That would constitute a solution to the value misgeneralization problem, and by extension a large part of the broader alignment problem. In this paper, I attempt to outline a path to achieving that.

[2]Steinhardt (2022a), Carlsmith (2022), and Turner et al. (2023).

[3]There is an interpretation on which goal misgeneralization refers specifically to a problem in RL-models. If so, this paper should be read as handling a problem in the same spirit that can be found in other contexts as well. Personally, I see them as the same problem, with Langosco et al. (2022) and Shah et al. (2022) definitions of goal misgeneralization being especially well-defined instances of it.

[4]Another way to describe the value misgeneralization problem is as saying that value learning in the sense of Shah (2018) will not be enough.

Here is a plan of the paper. In section 2, I demonstrate how the value misgeneralization problem will robustly lead to misaligned outcomes, unless the AGI can reliably predict our preferences over any outcome. In section 3, I demonstrate how an AGI developing a generative theory of mind can solve the value misgeneralization problem, and argue that this solution has been precluded by the widespread use of decision theoretic models to represent mental states. In section 4, I argue that we should instead employ reinforcement learning models to model human values, and provide an initial empirically grounded sketch of what such a theory would look like. I conclude with some final reflections. I have also included an Appendix, in which I present a generalized version of a model from section 2 and use it to categorize different causes of alignment problems.[5]

2 The Value Misgeneralization Problem

Let us articulate the value misgeneralization problem in more detail using a simple model. Suppose that we have two actors: a principal and an agent. The principal has some goals. These might include having a nice painting on the wall, a high game score, world peace, or a moderate number of paperclips, for example. However, the agent can only affect the world by communicating their goals in a signal to the agent who can act. The agent in turn receives the signal and forms goals of their own in an attempt to accurately match those of the principal, which they then act upon. The success condition is for the outcomes of the agent's actions to match the principal's goals. When this fails, we can say that the outcomes are *misaligned* with the goals of the principal.

In the case that we are interested in, the principal is a human, the agent is an AGI, and the principal's goals are the totality of the human's values. We can represent these as a preference ranking over all possible outcomes, and express it as a utility function.[6] Analogously, we can represent the goals of the AGI as a utility function. Since the AGI, by hypothesis, has no limit to its instrumental capacity, we can assume that it will successfully maximize its utility function. Therefore, to avoid misalignment, we want the utility

[5]A note to the reader: My own background is not in machine learning. Rather it is primarily in philosophy, microeconomic decision theory, and cognitive science. In the terminology of Steinhardt (2022b), this has resulted in this paper being an instance of a Philosophy approach to alignment, for better or worse. With that said, I have tried to tie it to concrete engineering applications where possible, and—contrary to the stereotype of philosophy—its main claim is that we need to understand an empirical domain better, in this case that of our own minds.

[6]Note that preferences and utility functions do not represent actual mental states found in the mind of the human. Rather, we use them as a convenient formal expression of what the human would want if presented with a given outcome. I return to this difference below.

function of the AGI (U_{AGI}) to be sufficiently similar to that of the human (U_H):[7]

Alignment: $U_{AGI} \approx U_H$

There are two different ways in which this can fail. First, it might be that the human miscommunicates their preferences, leading the AGI to adopt a different utility function. Second, the human might only be able to communicate a subset of their preferences, forcing the AGI to "fill in" the rest by guessing. In that case, we say that the signal is *sparse*. For example, I might like to have a nice painting on the wall, but be unable to communicate that together with all the other outcomes I want. In that case, my preference regarding the painting will not be part of the signal, and the agent must predict whether I would love, hate, or be indifferent to it. If the signal is sparse, and the AGI predicts the human's preferences inaccurately, then the resulting outcomes will be misaligned.[8]

We can put this a bit more precisely. U_H and U_{AGI} are functions from the set of outcomes that the AGI can bring about P to the set of real numbers $\mathbb{R}$.[9] Outcomes are sets of maximally specified possible worlds w_i.[10] Let S be a signal intended to communicate U_H.[11] S is sparse just when the domain of S is a strict subset of the domain of U_H. Let $\Pi(S)$ be the set of utility functions that are consistent with S from the perspective of the AGI. When the signal is both complete and accurate, $\Pi(S)$ contains only U_H. See Table 1 for reference.

In this paper, I will set aside misalignment caused by an inaccurate signal. In other words, I will assume that whatever preferences humans are able to communicate, they are able to do so perfectly accurately. It is worth noting

[7]More precisely: Let $o* \equiv argmax(U_{AGI})$. Alignment obtains iff $o* \in \{o : U_H(o) \geq \lambda\}$, where λ represents a sufficiently high value of U_H.

[8]These two failure modes are intended as generalized versions of the problems of reward misspecification and goal misgeneralization in reinforcement learning models. Arguably, they also correspond to the ideas of outer and inner alignment respectively (see e.g. Ngo (2020) and Leike (2022)), though I personally struggle to understand these notions precisely.

[9]P is here analogous to a choice set in microeconomics Mas-Colell, Whinston, and Green (1995). I assume that U_H and U_{AGI} are cardinal utility functions, which means that they have to satisfy some set of axioms to have the sufficient structure (especially, completeness, transitivity, and some separability axiom). For the purposes of this discussion, we can leave open which decision-theoretic system we use to do so. See Neumann and Morgenstern (1944), Savage (1954), and Jeffrey (1965) for some of the primary options, and see Steele and Stefánsson (2020) for an overview.

[10]Possible worlds are, roughly, hypothetical configurations of every fundamental particle in the universe. So, the outcome that I have a nice painting on the wall for, example, is the set of possible worlds in which that is the case. Some of these will have the Eiffel tower in Rome, or water on Mars, for example, but in all of them there is a nice painting on my wall. See e.g. Jeffrey (1965) and Stalnaker (1984).

[11]S is accurate iff S is a restriction of U_H, such that $S = U_H|_{Domain(S)}$.

Legend	
Human's utility function	$U_H : P \mapsto \mathbb{R}$
AGI's utility function	$U_{AGI} : P \mapsto \mathbb{R}$
Signal	$S : P \mapsto \mathbb{R}$
Set of U's consistent with S	$\Pi(S) \equiv \{U : S \subseteq U\}$
Possibility set	$P \subseteq \mathcal{P}(W)$
Set of all possible worlds	$W \equiv \{w_1, ..., w_n\}$

Table 1: Definitions

that this is a significant assumption. For example, if we imagine the signal being human behaviour conveyed to an inverse reinforcement learner, then this implies that the human will always act rationally, and that their actions always unambiguously reveal their preferences.[12] Instead, I will focus on cases of misalignment caused by the AGI making an inaccurate prediction of the human's preferences based on a sparse signal. I call this *value misgeneralization*.

To see what value misgeneralization might look like, consider this example. Suppose that I have the preferences $U_H(Banana) > U_H(Apple) > U_H(Orange)$. And suppose that I have expressed these utilities in a signal S, for example by rewarding the AGI in a training environment in proportion to my preferences. Furthermore, suppose that once it gets the opportunity to act, e.g. by entering a testing environment, it concludes that it can only provide me with *Banana* by committing *Robbery*. But suppose that I have not signalled my preference about *Robbery*, i.e. $Robbery \notin Domain(S)$, even though I in fact have a strong preference against it. In this case the AGI has to predict what my utility is for *Robbery* by picking a $U \in \Pi(S)$. However, $\Pi(S)$ contains U's giving all kinds of values to *Robbery*. Suppose that the strategy that the AGI adopts for picking a U is to assume that I am indifferent to any outcome I have not included in the signal.[13] In this case, it will conclude that $U_H(Banana \cap Robbery) > U_H(Apple \cap \neg Robbery)$, even though my preference is strongly the opposite.

In fact, I will argue that value misgeneralization of the kind seen in this example will almost inadvertently occur when the agent is an AGI, with potentially catastrophic misalignment as a result. In other words, even when we make the substantial assumption of an accurate signal, the alignment problem

[12]This means that I set aside most of the issues raised for value-learning by Shah (2018), for example. See also Dray (1957), Davidson (1974), Follesdal (1982) for classical discussions of the problem of the indeterminacy of interpretation from behaviour. Relatedly, see Quine (1960).

[13]I.e. it adopts a $U_{AGI} = U \in \Pi(S)$ s.t. $\forall o \notin Domain(S) : U(o) = 0$, where 0 represents indifference.

robustly remains. I call this the *value misgeneralization problem.* Here is the argument for the problem:[14]

> **The Value Misgeneralization Problem**
>
> (1) For any S a human can provide, the prior probability is very low that the U_{AGI} adopted from $\Pi(S)$ is similar to U_H.
>
> (2) There is no obvious predictive algorithm for adopting a U_{AGI} from $\Pi(S)$, such that conditionalizing on its use makes a relevant difference to the probability that U_{AGI} is similar to U_H.
>
> (3) *Conclusion:* The probability is very low that U_{AGI} is similar to U_H.

Here is an argument for (1). To evaluate how probable it is that U_{AGI} will be similar to U_H, we should consider how many U's in $\Pi(S)$ the AGI will have to choose from. This depends on how many U's the signal from the human leaves open as possibly representing their values, which in turn depends on how many possible U's the human could have to start with. If the possibility set P were small, then the human could express their preferences over most relevant outcomes, which would leave very few U's in $\Pi(S)$ for the AGI to choose from. However, since the agent is an AGI with near-limitless instrumental capabilities, P will be huge. This means that there will be far more possible outcomes in P than the human is realistically able to express their preference over. Therefore, any signal the human can provide will be sparse and leave open many possible U's, unified in their rankings over outcomes included in the signal, but wildly diverging over a sea of—possibly strange—outcomes not included in the signal. Consequently, $\Pi(S)$ will be very large, and only a small proportion of its U's will be tolerably similar to U_H. Therefore, assuming that the AGI has an equal probability of picking any $U \in \Pi(S)$, the probability that the resulting U_{AGI} is similar to U_H will be very low.

In other words, the only chance of robustly avoiding misalignment with an AGI is for it to reliably predict any preference of a human that is not in the signal, and thus increase the probability that the U it picks from a large $\Pi(S)$ is similar to U_H. It is worth noting that this implies that strategies for solving the alignment problem by widespread sampling of human preferences cannot succeed unless complemented by a predictive algorithm for unobserved preferences.[15] This provides a strong reason to think that e.g. inverse rein-

[14] I do not intend to suggest that this is an original argument in the alignment literature. Arguably, it is a restatement of the point that human values are fragile (Yudkowsky (2009)) together with Goodhart's Law. But I not sure about this, and regardless I hope that it is a good way to illustrate the point.

[15] Examples of these kinds of strategies arguably include Amin, Jiang, and Singh (2017), Sadigh et al. (2017), Russell (2019), Reddy, Dragan, and Levine (2019), Jeon, Milli, and Dragan (2020), Sumers et al. (2021), Reddy, Dragan, Levine, et al. (2021), and others that pursue value learning in the sense discussed by Shah (2018)

forcement learning cannot constitute a solution to the alignment problem by itself, even though it might be part of a solution.

Consider (2) next. It states that there is no obvious method for the AGI to reliably predict human preferences in the way we need. Why would we think this? Arguably, one way to understand the alignment debates over the last 15 years or so is in terms of trying to propose algorithms for predicting our preferences in this way. That suggests that there is good reason to accept (2). To see the difficulty, it is worth briefly considering some of the possible proposals.

One natural first suggestion is to assume that for any outcome o that is not part of the signal, the agent is indifferent about whether or not o, such that $U_H(o) = U_H(\neg o)$. As we saw in the fruit example above, however, this will predictably lead to misalignment. This is because there are many outcomes that we cannot communicate—for example, because we have never thought about them—about which we are not indifferent.

A *prima facie* plausible variation on this is to give a strongly negative value to any outcome that is not in the signal. The thought would be to discourage any action towards uncommunicated outcomes. But this will not work. Suppose that o is not in the signal, and the AGI determines that $U_H(o) < U_H(\neg o)$. This would be a violation of the rule, since if o is not in the signal then neither is $\neg o$. The reason is just that you can't have a preference against something unless you prefer its negation. In order to do this, one would have to distinguish o and $\neg o$, for example by identifying $\neg o$ as the "status quo" to be preserved. This raises new issues, however. First, what counts as a status quo is arguably itself something we would need to communicate, which then gives rise to a meta-version of the problem that we cannot communicate all we need. Second, in some cases it seems clear that we do not want to preserve intuitive status quos, for example if a diabetic needs insulin quickly.

Another seemingly more promising suggestion is to draw on statistical correlations between preferences, both on a population level and for a given individual, to infer the agent's values about uncommunicated outcomes. The idea would essentially be for the AGI to use a much more capable version of a predictive algorithm like Spotify's song recommendations. Just like such predictive algorithms, this will likely work in every-day domains. However, as noted before, this cannot work when it comes to predicting outcomes that nobody has ever signalled their preference about, yet clearly would not be indifferent about if faced with. This is because we would have no data to draw on in such cases.

This suggests that the prediction would need to draw not just on correlations between preferences, but on relevant similarities between features of the outcomes. The most obvious way to do this is to identify some similarity metric between outcomes, and infer that a human will prefer outcomes similar to their communicated preferences similarly. For example, we might believe

that someone who likes oranges likes satsumas too. Arguably, however, there is no straight-forward similarity metric that can do this reliably. To illustrate with a naïve example, suppose we take the similarity metric to be similarity of physical configuration. This will not work, because there are many cases where large physical changes have very small effects on our preferences, and small physical changes have large effects. I, for example, care much more about small displacements of my amygdala than large displacements of my couch.

The concern is not just that a good similarity measure seems hard to find. Rather, the deeper point is that no similarity metric seems to explain why we have the preferences that we do, which seems to be due to something more complicated than mere similarity of any kind. However, most of us do not struggle knowing what we would think about even far-fetched outcomes. That suggests that there is some predictable—and exploitable—structure to the formation of our preferences that has been neglected in previous suggestions of preference-predicting algorithms.[16] In the next section, I propose what that structure might look like, and how it might provide a path to a solution to the value misgeneralization problem.

3 Generative Theory of Mind as a Solution

Our goal is to find an algorithm that allows an AGI to predict our preferences given a sparse signal, and thereby adopt a tolerable utility function as its goals. Despite the apparent difficulty in finding such an algorithm in the previous section, there is one reason to think that it should be possible, which is that we humans do it all the time. Even when we only know some of other people's preferences, we can usually predict what a person would think about some hypothetical outcome, even outlandish ones. For example, I am confident that the person I see on the street in front of me has a strong preference against the moon being turned to cottage cheese, even though there is a good chance they have never thought about that outcome. These are the kinds of judgments that seem like common sense to us.

I call this ability to accurately predict the hypothetical preferences of other humans having a generative theory of mind. Theory of mind is an expression from psychology, which means being able to simulate and attribute mental states to others. Loosely speaking, we can think of it as a capacity to empathize, or to accurately take the intentional stance toward someone. By a generative theory of mind, I mean the capacity to accurately be able to say what attitude someone would have toward an arbitrary outcome, even if that outcome is not in a current mental state of the agent.

[16] This is arguably a way of saying that our human values can only be so complex on a fundamental level (LessWrong (2018)).

The expression draws on the analogy to Chomsky's notion of a generative grammar.[17] He famously argues that, even though it was not a sentence that anyone had ever considered at the time, "Colorless green ideas sleep furiously" can be trivially identified as grammatically correct by humans. He took this to prove that we have a capacity to create and evaluate grammatically correct sentences in a generative and unbounded way.[18] Analogously, it seems like we can generatively and unboundedly attribute preferences to other people for previously unconsidered outcomes.

How are humans able to do this? Here is a crucial and seemingly neglected observation: There is a causal and explanatory hierarchy to our values. For example, I might want a hammer and a nice painting on the wall. But I might also want the hammer *because* I want a nice painting on the wall. Specifically, I might want the painting on the wall and believe that the hammer will be instrumentally important in causing the painting to be there. This illustrates that there are more specific and informative facts about the relation between my attitudes to hammers and painting than, for example, that they are correlated.

Arguably, this is the essential component that we exploit to generatively predict each other's values.[19] Here is a schematic and simplified elaboration of how this works. We know about some basic values that almost all humans share, such as good food and bodily integrity. We might also know about some beliefs that someone has about how to achieve these things, such as bananas being good food or that being punched in the face is bad for one's bodily integrity. And from this we infer that such a person has a preference for bananas and against being punched in the face. And conversely, if we know that someone wants a new TV, we can usually infer that they want entertainment, and that they want a new TV because they believe it will provide entertainment. Of course, they might want a TV for other reasons, but the hypotheses can be narrowed down with more knowledge of other values.

By mapping this instrumental structure of our values we can start "filling in" others' preferences in a generative way. We do this using our knowledge of a subset of their preferences, which can come both from population-level data and idiosyncratic expressions, and our knowledge about their causal models of the world. This is how we could have been confident 50 years ago that most people would have a strong preference against being turned into paperclips,

[17] Chomsky (1957).

[18] In fact, this example was first raised in an argument against theories of grammar relying on statistical regularities, which suggests that the dialectic mirrors that of this debate where statistical methods of predicting preferences seem similarly inadequate. Note that, unlike Chomsky, I do not mean to say that this capacity is explained by an innate module.

[19] For a vivid argument of how folk-psychological concepts like beliefs and desires provide us with considerable predictive powers, see the introduction to Fodor (1987).

even though this is not something that anyone had plausibly entertained as an outcome at the time. Arguably, this ability is an essential part of what we intuitively call "common sense". I propose that teaching an AGI to do the same thing—map the causal structure of our values and generatively infer our preferences about unconsidered outcomes—provides a path to solving the value misgeneralization problem, with a result that might similarly look like common sense.

A reader might naturally wonder why this seemingly simple point has not been widely appreciated before, if it is as important as I have argued.[20] Part of the explanation is that alignment is a young field. But I think there is a more important explanation. This is that the way that we have represented human values in alignment discussions have usually abstracted away from the hierarchical relations I emphasize here, and thus obscured the point. In particular, human values are typically represented as preferences or utility functions, as we did in the previous sections.

Utility functions are fantastic for succinctly representing how an agent would judge different outcomes, but they are terrible for representing the psychological realities that produce those judgments.[21] This was of course well-known to the founders of decision theory, most of whom explicitly rejected their decision theories as being isomorphic models to cached mental states.[22] In economics, for example, the neoclassical paradigm was largely based on avoiding speculation on what happens in the head of people, considering it unscientific, and stuck to theorizing about generalizable hypothetical choice.[23] The attitude towards theorizing about mental states as unscientific was largely overhauled by the "cognitive revolution",[24] but arguably with the side-effect that the framework of preferences and utility got an ambiguous meaning between being about hypothetical choice and actual mental states.[25] This does not matter in most contexts in economics, for example in aggregating preferences to study consumption patterns. But it does matter when we are trying to predict what an individual would think about some highly specific outcome that they have never encountered. In that case we cannot

[20]It is not entirely fair to say that this point has not been appreciated. As I understand him, Yudkowsky (2007), for example, considers the idea that our values are reducible to simple ones, but argues that this reduction would still result in a large heterogeneity. As I argue below, our preferences do demonstrate this heterogeneity, but they are produced by more simple and predictable processes.

[21]For example, humans can only have actual mental attitudes towards a relatively small number of outcomes due to the limits of our memory, but the completeness axiom requires us to have preferences over all outcomes in P.

[22]E.g. Ramsey (1926) and Savage (1954).

[23]See e.g. Binmore (2008).

[24]Wikipedia (2023).

[25]Gul and Pesendorfer (2008).

afford not to theorize about what's in the head, but we should not expect utility and preferences to be the best tools for doing so.

The use of preferences and utility functions to represent psychological human values raises a particular problem for our purposes, namely that preferences are "flat". A preference ranking allows us to express how much I want a hammer relative to how much I want a painting on the wall, but it does not allow us to express that I want the hammer because I want a painting on the wall. Or that my preference for the hammer is causally dependent on my preference for the painting in an asymmetric way. In other words, using only a utility function there is no way to represent the hierarchical relations between outcomes, as they are related in our mental causal-predictive models. In a standard utility-based model, my preference for a hammer has "nothing more to do" with my preference for a painting than does my preference for world peace—the only dimension on which they can be related is in strength of preference—which is obviously different from the relations between my psychological attitudes to these outcomes. Consequently, as long as we represent our values as utility functions in alignment discussions, we abstract away from the main facts that we need to represent in order to solve the value misgeneralization problem, and by extension the alignment problem.[26]

To clarify, this does not mean that we should stop using utility functions and preferences to represent our values when articulating abstract problems, as I have done in previous sections, for example. It only means that we shouldn't be trying to explain why someone has a preference, psychologically speaking, in terms of other preferences. Preferences can be used to represent the resulting picture of a good psychological theory, e.g. by expressing what it means that we would choose if presented with an outcome, but should not be used to arrive at that picture. Consequently, I will keep talking about

[26]There are nuances in this discussion, depending on which decision-theoretic framework we work with, and whether they make distinctions between utility over certain outcomes and expected utility over entities with uncertain outcomes. Jeffrey (1965) does not, for example, and rather treats all outcomes and actions as sets of possible worlds with associated probabilities. Savage (1954), by contrast, makes a distinction between what he calls consequences—elements of a partition over the possibility set that distinguishes any outcomes that an agent has varying preferences over—and acts—functions from uncertain events to prospects. Neumann and Morgenstern (1944) make a similar distinction between basic and non-basic lotteries, where basic lotteries are certain outcomes. For Savage and von Neumann & Morgenstern you might say that e.g. basic lotteries serve as bottom levels in the hierarchy, and all other outcomes are non-basic lotteries. This, however, will lead to a very limited framework for expressing values with more than one level of hierarchy. For example, even if the hammer is a non-basic lottery that has a high chance of resulting in the painting, we likely want the painting for some deeper reason, and thus would not want to represent it as a basic lottery. In practice, we often decide which outcomes to treat as basic for practical purposes when modelling, but if we are in fact trying to map the hierarchical structure of our complete values we cannot do that.

preferences as things that we want to predict by means of other psychological theories.

In order to understand how our generative theory of mind works, and thereby understand how we can teach it to an AGI, we need some models to use for theorizing about our values. But if those models should not be built on preferences and utility functions, then what should they look like?

One suggestion is to theorize using folk-psychological concepts like values and beliefs, as I have loosely done above, and as is arguably done in most areas of academic psychology. While that allows us to represent the hierarchical relation between our different mental states, it comes at the cost of precision, which was the original motivation for adopting a decision-theoretic framework. And realistically, folk-psychological notions will not be operationalizable in computational models. What we need instead is some conceptual framework that is rich enough to represent the hierarchical structure of our values, and precise enough to allow for the development of a computationally precise theory for generatively predicting new preferences. I believe that there is a conceptual framework which satisfies these desiderata, which is reinforcement learning.

4 Reinforcement Learning as a Path Forward

4.1 Values and Reinforcement Learning

I will argue that we should try to model human evaluative cognition similarly to how we try to model that of artificial agents, and that we can represent and eventually explain and predict the structure of our values by doing so. A version of this claim might be widely accepted among participants in the alignment community—for example, in the idea that we try to find people's reward function in inverse reinforcement learning—but I don't think the proposal has been sufficiently appreciated as a realistic psychological theory. The details of what an adequate RL-based theory of values looks like is far beyond the scope of this text, and rather the target of a large research program. However, I will provide an initial sketch for thinking that this is a promising path for explaining the formation of our preferences, and consequently our generative theory of mind.

To bring any reader up to speed: RL is a formal framework for representing agents that exist over time and learn from experience.[27] A part of that framework is the representation of an agent's evaluations of its environment, or its values loosely speaking. This is represented by two core components. First, its reward function (r), which determines how intrinsically good any given state of the environment is from the perspective of the agent. Second,

[27] Sutton and Barto (2018).

its value function (V), which is a predictive function representing how much (discounted) reward that the agent expects to have in the long run from some given state or action, assuming some policy for how to act in the future. An RL agent chooses the state or action that their value function attributes the highest value to.

The distinction between reward and value-representations (I will use the term value-representation to refer to the outputs of the value-function in an RL-framework, and 'value' to refer to the folk-psychological notion) makes RL a hierarchical framework, where value-representations are valued because they are instrumental to rewarding states, according to the predictive models of the agent. For example, suppose that states with nutrition are rewarding to an agent. That agent might have learnt that going to the grocery store and buying food is an action that in expectation leads to nutritious and thus rewarding states, and therefore attribute high value to the action of going to the store.

In this basic form, RL does not obviously make progress on a standard framework of expected-utility theory, which has a basic level of preferences over outcomes, and an instrumental level of actions likely to lead to those outcomes. However, we can add more layers to our hierarchy by identifying certain value-representations as the ends of other value-representations. For example, I might value having a nice painting on the wall because I expect it to be beautiful, which—let us assume—I take to be a rewarding state. Furthermore, I value having a hammer because I expect it to lead to my having a painting on the wall, and ultimately more beauty. In other words, an agent might learn that certain states are valuable in expectation, and then take these as "checkpoints" to aspire towards in their actions. This can generate a hierarchical chain of checkpoints where we aspire to $State_i$ because we expect it to bring about $State_{i-1}$, which we aspire to because we expect it to bring about $State_{i-2}$, all the way down to the intrinsically rewarding $State_0$. These states are the objects of the value function V in an RL-framework, but as they are outcomes, they are also objects of preference, and we can represent them using a utility function.[28]

As a short aside, this implies a picture of reward that is distinct from the one seemingly assumed in much work on inverse reinforcement learning. There it is suggested that a model could infer a human's reward function from behaviour. However, if the picture above is right, then it is not realistic that we could literally infer which states are intrinsically rewarding for a human. For example, if a model observes a human going to the store, it would be mistaken to conclude that going to the store is a rewarding state. Rather it would need to infer why the human is going to the store, and why the human wants

[28]Strictly speaking, all states are outcomes, but not all outcomes are states. This is because outcomes can be temporally extended, while states are momentary. Specifically, state A at time t is the set of possible worlds where A is the case at t.

the food it expects to find there, etc. In other words, the preferences that we realistically reveal by behaviour are unlikely to be intrinsically rewarding states, but rather states that are highly valued in expectation. Inferring the rewarding states will be far more difficult, and require more information about the causal-predictive models of the human.

The picture above is crucially oversimplified in that one higher state can be valued because an agent expects it to lead not just to one, but several, more fundamental states of value. In fact, this is plausibly what happens with fundamental value-representations, which are likely to be learnt as valuable in bringing about a wide array of basic rewarding states. For example, I might learn that not falling from high places is in expectation conducive both to bodily integrity and ability to secure nutrition. In this sense, the right analogy is not as much a chain of checkpoints as a network of them, with expected connections crisscrossing between states of varying fundamentality.

How can this explain our generative theory of mind? The basic idea is that humans have a tacit understanding of what this network looks like for others, and in particular a good understanding of the most fundamental states. Using this understanding we can then predict what the other nodes in the network are, and thus make accurate guesses about people's preferences. And we do this by tacitly inferring what someone would think about an outcome, for example from knowledge of their more fundamental values and our knowledge about how they would expect that hypothetical outcome to affect those fundamental values. *On the proposal defended here, this is the essence of what gives us our generative theory of mind.*

An adequate explanation of how our generative theory of mind works, however, requires a more precise account of this hierarchical network of value-representations and their origins for humans. In the next subsection, I provide a sketch of how we can explain human evaluative cognition using RL-models to fill out that picture. This part is more speculative and opinionated than the rest, and should be taken as an attempt to start a discussion about the empirical structure of human values, rather than an attempt to end it.

4.2 Humans and Reinforcement Learning

There is increasing evidence from neuroeconomics and cognitive science that RL-models map onto processes in the mind. One clear example is dopamine release, which can be predicted accurately by temporal difference models.[29] The details of this research is a contested topic, but the consensus is that it suggests some very close relation between RL-algorithms and how we learn to value outcomes.[30] On a more abstract level, there is evidence that we have a

[29]Schultz, Dayan, and Montague (1997) and Glimcher and Fehr (2013).

[30]O'Doherty (2004), Montague, King-Casas, and Cohen (2006), and Ribas-Fernandes et al. (2011).

some form of common currency—a numerical representation in our minds—in terms of which we evaluate which actions to perform and where to devote our executive processing.[31] The tractability of such models have, for example, given rise to the field of computational psychiatry, which attempts to use them to explain and predict pathological mental states in humans, with apparent success.[32]

These observations should give us some initial credence that human psychology can be well-understood on RL-models. The suggestion raises some conceptual questions, however. The primary one is what the reward function corresponds to in humans. In other words, what are the most fundamental states in our hierarchical network of values, and how are they determined?

For most artificial agents, the reward function is pre-specified as a mapping from states of the environment to real values, i.e. $r : \mathcal{S} \mapsto \mathbb{R}$ (where $\mathcal{S}$ is the set of possible states). For example, Deepmind's Atari-playing agent took different states of game scores as reward.[33] One response to the question that seems widely accepted among philosophers is that the reward function works analogously for humans as for such artificial agents, but where the rewarding states are intrinsically desired outcomes. The idea is that we desire some things intrinsically, and some things instrumentally, and that these correspond to value-representations and rewards in an RL-framework respectively.[34]

While attractive at a first glance, this interpretation of the reward function cannot be right for natural agents such as humans. The reason is that in order for an agent to assess something as rewarding or not, the agent must have something in place that determines how they will evaluate it once they encounter it (e.g. a disposition to feel pain in response to the experience). And if that this is supposed to be a state, then the agent needs a prior representation of that state with an associated value for how to evaluate it. The Atari-player does this by having a hardwired representation that specifies a reward value for all states it might encounter, which is feasible to provide in a curated toy-environment. But natural agents living in a complex environment cannot have a prior representation of all states they might encounter and need to learn from. For example, human babies are not born with an innate representation of hot plates, yet somehow they learn that touching a hot plate is an unrewarding state to be in. This suggests that the domain of the reward function must be something more basic than states, from which humans compute the reward of a state.

[31] Montague and Berns (2002), Dayan and Niv (2008), Rangel, Camerer, and Montague (2008), Meer, Kurth-Nelson, and Redish (2012), Shenhav, Cohen, and Botvinick (2016), and Kool, Shenhav, and Botvinick (2017).

[32] Series (2020).

[33] Mnih et al. (2015).

[34] Schroeder (2004) and Haas (2022).

Another suggestion is that this role is filled by hedonic signals such as pain and pleasure, meaning that our evaluative cognition views states as rewarding in proportion to how pleasurable they are. This view seems assumed in the RL community.[35] There are, however, a few reasons to think that it is false. First, what determines an organism's survival is not whether it is in pain but, for example, whether it is hurt. This means that if pain were to determine the reward, we should expect it to do so indirectly by working as a proxy for harm. But a simpler explanation is that an organism would have the harm itself be the determinant of the reward, rather than requiring pain to be a middle-man. And if we did require that reward was determined via e.g. pain, that would imply that any natural agent that has something corresponding to a reward function can feel pain, which is a strong empirical commitment that is better to avoid if we can. More importantly, there are other better explanations for the evolutionary function of hedonic signals, in particular of their regulating our attention and biasing our predictions about rewarding states in ways that have been conducive to survival in our ancestral past (e.g. motivating us not to neglect a wound).[36] For those reasons we should not believe that hedonic signals are the most basic determinants of reward.

Here is what I take to be the most plausible account of reward for natural agents.[37] Reward is a function from features of the environment to real-valued numerical representations, i.e. $r : \mathcal{F} \mapsto \mathbb{R}$, where $\mathcal{F}$ is the set of possible features and quantities of them.[38] On this view, the reward function corresponds to evaluative sensitivities to certain features of the world—caloric balance, hydration, bodily integrity, the welfare of nearby agents—the encountering of which have benefitted our ancestors and selected for individuals who were motivated to pursue them. The idea is that these features are analogous for our capacity to evaluate to what the visible parts of the light spectrum is for our capacity to see. They are the features of the world that we humans are idiosyncratically sensitive to, and from which we build up more complex representations. In the case of vision, these are complete visual objects. In the case of evaluation, these are outcomes that we take to be valuable and have a preference for.

On this proposal, human minds have an adaptive, innate, and robust evaluative mechanism that takes as inputs features of the environment and computes a value for how rewarding the state with those features are. Which features it is sensitive to is an empirical question, but it will be states with

[35] Sutton and Barto (2018).

[36] Ohman, Flykt, and Esteves (2001) and Carruthers (2018).

[37] I should note that this is a topic of my current research, which arguably influences how plausible I find the view.

[38] More precisely, I take $\mathcal{F}$ to be a set of ordered pairs, where the slots represent the kind of feature and a quantity respectively. I assume that states can be defined in terms features, such that $\forall s \in \mathcal{S} : s \subseteq \mathcal{F}$

features that are historically conducive to our survival. This mechanism is unlikely to be performing a linear computation, as there would be plausible reproductive benefits to evaluating certain combinations of features as more rewarding than their parts. A better proposal of its structure is that of a deep neural net, where the weights, biases, and connections of the network represent the relative importance of different environmental features and their interactions in that state.[39]

If this account of the reward function is correct for humans, then a speculative developmental picture emerges. On this picture, we are born with these innate evaluative sensitivities, which allow us to recognize some states as more rewarding than others. Gradually, we learn to predict which other states tend to lead to rewarding states in expectation—say, having food, and being safe—and come to represent them as valuable states. When our executive functions develop with age, we learn to deliberately make long-term plans to achieve such states—doing well in school, having a good career—and states that put us in a better position to achieve such states—studying for an exam, being nice to our boss—and so worth. This is how our network of hierarchical values emerges.[40] Furthermore, the complexity and non-linearity of the reward function, and the massive variation in ways we can learn to arrive at rewarding states, can explain why humans demonstrate a large variation across cultures and family contexts in the things we take as important, at least on the higher, more instrumental, levels of values (i.e. there is more variation in favorite grocery stores than on whether food is valuable).

What does this picture imply for the value misgeneralization problem, and the alignment problem more broadly? If it is correct, it suggests that an AGI that learnt the details of our reward functions—that learnt the structure of the mechanisms that determine the reward of a state from its features—and had knowledge of our causal models of the world, would in principle we able to infer our value-representations, and by extension our preferences over any outcome. This could in turn be made more robust by triangulating such facts with the preferences we express in our signals, e.g. via our behaviour to an inverse reinforcement learning agent. This gives us reason to think that such an AGI would be able to robustly predict any preference we cannot include in our signals, and thus reliably adopt a $U \in \Pi(S)$ that is sufficiently similar to U_H. That would constitute a solution to the value misgeneralization problem.

[39] To be clear, if this mechanism computes as a neural net, this does not mean that it learns like the neural nets most of us are most familiar with. Rather, it would be a mechanism that develops in response to selective pressures between generations, but the properties of which remained mostly fixed during the lifetime of an agent. See e.g. Singh et al. (2010) for an evolutionary model that captures aspects of this. This is different from the neural nets that are used for approximating value functions in artificial agents, for example, which develop by gradient descent during the "lifetime" of that agent.

[40] This picture is supported by recent developmental research from Gopnik (2020), showing that we move towards the exploitation end of the explore/exploit spectrum with age.

5 Conclusion

In this text I have argued that the only apparent solution to the value alignment problem, a robust cause of the alignment problem, is for the artificial agent to learn a generative theory of mind for humans. Since variations of the value misgeneralization problem constitute a significant portion of the most threatening alignment scenarios, finding a way for the agent to learn a generative theory of mind would also constitute significant marginal progress in solving existential risk from unaligned artificial agents.

The account that I have provided here has many serious limitations. It lacks anything like sufficient detail to constitute a solution by itself. That is not its intention, either, however. Rather, the intention is to demonstrate a pathway to a solution that, unlike other suggestions for preference-prediction, does not face apparent principled problems. If it is successful in this task, then that has actionable implications for how the alignment community should allocate resources going forward. In particular, it implies that we should invest significantly more resources into understanding the structure of the human minds that we are trying to align artificial agents to. In practice, this would mean funding research in cognitive neuroscience and related disciplines with an explicit alignment focus.

The paper also raises further philosophical questions that have not been appreciated previously. As an example, I have here assumed that the outcomes we want an AGI to bring about are those that we have a preference for. This preference, however, is a function both of more fundamental values, all the way down to intrinsically rewarding states, and the causal-predictive models of the person. Suppose that those causal-predictive models are inaccurate. Imagine, for example, a person who has a preference for cigarettes, but where this preference is based on the false predictive model that they are not harmful. Should we want an artificial agent to provide the person with cigarettes in such a case? Or should we want it to bring about outcomes that the person would have preferred—based on their more fundamental values—if they would have had a more accurate model? At the limit, that would seem to imply that the agent should maximize the intrinsically rewarding states for humans, but not necessarily in any way that we have learned and are familiar with.[41] The result might be a strange one, where none of the outcomes that feature in our learnt value-representations—education, careers, material possessions—are brought about.

Responses to this and other questions will require more research. The hope is that this text has provided a partial starting place from which to pursue it.

[41] Note that if the speculations above regarding human reward are correct, this is unlikely to be a hedonic maximization where humans are forcibly fed opioids for example, since hedonic signals are not rewarding on that account.

References

Amin, Kareem, Nan Jiang, and Satinder Singh (Nov. 2017). *Repeated Inverse Reinforcement Learning*. arXiv:1705.05427 [cs]. DOI: 10.48550/arXiv.1705.05427. URL: http://arxiv.org/abs/1705.05427 (visited on 02/09/2023).

Binmore, Ken (2008). *Rational Decisions*. Princeton University Press.

Bostrom, Nick (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Carlsmith, Joseph (June 2022). *Is Power-Seeking AI an Existential Risk?* arXiv:2206.13353 [cs]. DOI: 10.48550/arXiv.2206.13353. URL: http://arxiv.org/abs/2206.13353 (visited on 10/04/2022).

Carruthers, Peter (2018). "Valence and Value". In: *Philosophy and Phenomenological Research* 97(3), pp. 658–680. DOI: 10.1111/phpr.12395.

Chomsky, Noam (1957). *Syntactic Structures*. Mouton.

Christian, Brian (Oct. 2021). *The Alignment Problem: Machine Learning and Human Values*. English. W. W. Norton & Company. ISBN: 978-0-393-86833-3.

Christiano, Paul (2022). "Where I agree and disagree with Eliezer". en. In: URL: https://www.lesswrong.com/posts/CoZhXrhpQxpy9xw9y/where-i-agree-and-disagree-with-eliezer (visited on 02/09/2023).

Davidson, Donald (1974). "Psychology as Philosophy". In: *Philosophy of Psychology*. Ed. by Stuart C. Brown. Harper & Row, pp. 41–52.

Dayan, Peter and Yael Niv (Apr. 2008). "Reinforcement learning: the good, the bad and the ugly". eng. In: *Current Opinion in Neurobiology* 18(2), pp. 185–196. ISSN: 0959-4388. DOI: 10.1016/j.conb.2008.08.003.

Dray, William H. (1957). *Laws and Explanation in History*. London: Greenwood Press.

Fodor, Jerry A. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. MIT Press.

Follesdal, Dagfinn (1982). "The Status of Rationality Assumptions in Interpretation and in the Explanation of Action". In: *Dialectica* 36(4). Publisher: Wiley-Blackwell, pp. 301–316. DOI: 10.1111/j.1746-8361.1982.tb01545.x.

Gabriel, Iason (Sept. 2020). "Artificial Intelligence, Values and Alignment". In: *Minds and Machines* 30(3). arXiv:2001.09768 [cs], pp. 411–437. ISSN: 0924-6495, 1572-8641. DOI: 10.1007/s11023-020-09539-2. URL: http://arxiv.org/abs/2001.09768 (visited on 02/09/2023).

Glimcher, Paul W. and Ernst Fehr, eds. (Oct. 2013). *Neuroeconomics: Decision Making and the Brain*. English. 2nd edition. Academic Press: Amsterdam : Boston. ISBN: 978-0-12-416008-8.

Gopnik, Alison (June 2020). "Childhood as a solution to explore–exploit tensions". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 375(1803). Publisher: Royal Society, p. 20190502. DOI: 10.1098/

rstb.2019.0502. URL: https://royalsocietypublishing.org/doi/10.1098/rstb.2019.0502 (visited on 01/20/2023).

Gul, Faruk and Wolfgang Pesendorfer (2008). "The case for mindless economics". In: *The foundations of positive and normative economics: A handbook* 1. Publisher: Oxford University Press Oxford, pp. 3–42.

Haas, Julia (2022). "Reinforcement Learning: A Brief Guide for Philosophers of Mind". In: *Philosophy Compass* 17(9), e12865. DOI: 10.1111/phc3.12865.

Hendrycks, Dan and Mantas Mazeika (Sept. 2022). *X-Risk Analysis for AI Research*. arXiv:2206.05862 [cs]. DOI: 10.48550/arXiv.2206.05862. URL: http://arxiv.org/abs/2206.05862 (visited on 10/04/2022).

Hubinger, Evan (Dec. 2020). *An overview of 11 proposals for building safe advanced AI*. arXiv:2012.07532 [cs]. DOI: 10.48550/arXiv.2012.07532. URL: http://arxiv.org/abs/2012.07532 (visited on 02/09/2023).

Jeffrey, Richard C. (1965). *The Logic of Decision*. University of Chicago Press.

Jeon, Hong Jun, Smitha Milli, and Anca D. Dragan (Dec. 2020). *Reward-rational (implicit) choice: A unifying formalism for reward learning*. arXiv:2002.04833 [cs]. DOI: 10.48550/arXiv.2002.04833. URL: http://arxiv.org/abs/2002.04833 (visited on 02/09/2023).

Kool, Wouter, Amitai Shenhav, and Matthew M. Botvinick (2017). "Cognitive control as cost-benefit decision making". In: *The Wiley handbook of cognitive control*. Wiley Blackwell: Hoboken, NJ, US, pp. 167–189. ISBN: 978-1-118-92054-1 978-1-118-92048-0 978-1-118-92047-3. DOI: 10.1002/9781118920497.ch10.

Krakovna, Victoria (2022). "DeepMind alignment team opinions on AGI ruin arguments". en. In: URL: https://www.lesswrong.com/posts/qJgz2YapqpFEDTLKn/deepmind-alignment-team-opinions-on-agi-ruin-arguments (visited on 02/09/2023).

Langosco, Lauro et al. (2022). *Goal Misgeneralization in Deep Reinforcement Learning*. arXiv:2105.14111 [cs]. DOI: 10.48550/arXiv.2105.14111. URL: http://arxiv.org/abs/2105.14111 (visited on 02/09/2023).

Leike, Jan (May 2022). *What is inner alignment?* Substack newsletter. URL: https://aligned.substack.com/p/inner-alignment (visited on 02/09/2023).

LessWrong (2018). *Complexity of value - Lesswrongwiki*. URL: https://wiki.lesswrong.com/wiki/Complexity_of_value?_ga=2.95689820.1715234189.1675971614-1490337092.1665936998 (visited on 02/09/2023).

Mas-Colell, Andreu, Michael D. Whinston, and Jerry R. Green (June 1995). *Microeconomic Theory*. English. Illustrated edition. Oxford University Press: New York. ISBN: 978-0-19-507340-9.

Meer, Matthijs van der, Zeb Kurth-Nelson, and A. David Redish (Aug. 2012). "Information processing in decision-making systems". eng. In: *The Neuroscientist: A Review Journal Bringing Neurobiology, Neurology and Psychiatry* 18(4), pp. 342–359. ISSN: 1089-4098. DOI: 10.1177/1073858411435128.

Mnih, Volodymyr et al. (Feb. 2015). "Human-level control through deep reinforcement learning". en. In: *Nature* 518(7540). Number: 7540 Publisher: Nature Publishing Group, pp. 529–533. ISSN: 1476-4687. DOI: 10.1038/nature14236. URL: https://www.nature.com/articles/nature14236 (visited on 01/20/2023).

Montague, P. Read and Gregory S. Berns (Oct. 2002). "Neural economics and the biological substrates of valuation". eng. In: *Neuron* 36(2), pp. 265–284. ISSN: 0896-6273. DOI: 10.1016/s0896-6273(02)00974-1.

Montague, P. Read, Brooks King-Casas, and Jonathan D. Cohen (2006). "Imaging valuation models in human choice". eng. In: *Annual Review of Neuroscience* 29, pp. 417–448. ISSN: 0147-006X. DOI: 10.1146/annurev.neuro.29.051605.112903.

Neumann, John Von and Oskar Morgenstern (1944). *Theory of Games and Economic Behavior*. Princeton, NJ, USA: Princeton University Press.

Ngo, Richard (2020). "AGI safety from first principles". URL: https://www.alignmentforum.org/s/mzgtmmTKKn5MuCzFJ.

Ngo, Richard, Lawrence Chan, and Sören Mindermann (Dec. 2022). *The alignment problem from a deep learning perspective*. arXiv:2209.00626 [cs]. DOI: 10.48550/arXiv.2209.00626. URL: http://arxiv.org/abs/2209.00626 (visited on 02/09/2023).

O'Doherty, John P. (Dec. 2004). "Reward representations and reward-related learning in the human brain: insights from neuroimaging". eng. In: *Current Opinion in Neurobiology* 14(6), pp. 769–776. ISSN: 0959-4388. DOI: 10.1016/j.conb.2004.10.016.

Ohman, A., A. Flykt, and F. Esteves (Sept. 2001). "Emotion drives attention: detecting the snake in the grass". eng. In: *Journal of Experimental Psychology. General* 130(3), pp. 466–478. ISSN: 0096-3445. DOI: 10.1037//0096-3445.130.3.466.

Quine, Willard Van Orman (1960). *Word and Object*. Cambridge, MA, USA: MIT Press.

Railton, Peter (Sept. 2020). "Ethical Learning, Natural and Artificial". In: *Ethics of Artificial Intelligence*. Ed. by S. Matthew Liao. Oxford University Press, p. 0. ISBN: 978-0-19-090503-3. DOI: 10.1093/oso/9780190905033.003.0002. URL: https://doi.org/10.1093/oso/9780190905033.003.0002 (visited on 10/04/2022).

Ramsey, Frank (1926). "Truth and Probability". In: *Philosophy of Probability: Contemporary Readings*. Ed. by Antony Eagle. Routledge, pp. 52–94.

Rangel, Antonio, Colin Camerer, and P. Read Montague (July 2008). "A framework for studying the neurobiology of value-based decision making". en. In: *Nature Reviews Neuroscience* 9(7). Number: 7 Publisher: Nature Publishing Group, pp. 545–556. ISSN: 1471-0048. DOI: 10.1038/nrn2357. URL: https://www.nature.com/articles/nrn2357 (visited on 02/09/2023).

Reddy, Siddharth, Anca D. Dragan, and Sergey Levine (Sept. 2019). *SQIL: Imitation Learning via Reinforcement Learning with Sparse Rewards*. arXiv:1905.11108 [cs, stat]. DOI: 10.48550/arXiv.1905.11108. URL: http://arxiv.org/abs/1905.11108 (visited on 02/09/2023).

Reddy, Siddharth, Anca D. Dragan, Sergey Levine, et al. (2021). *Learning Human Objectives by Evaluating Hypothetical Behavior*. arXiv:1912.05652 [cs, stat]. DOI: 10.48550/arXiv.1912.05652. URL: http://arxiv.org/abs/1912.05652 (visited on 02/09/2023).

Ribas-Fernandes, José J. F. et al. (July 2011). "A neural signature of hierarchical reinforcement learning". eng. In: *Neuron* 71(2), pp. 370–379. ISSN: 1097-4199. DOI: 10.1016/j.neuron.2011.05.042.

Russell, Stuart (Oct. 2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. English. Penguin Books.

Sadigh, Dorsa et al. (2017). "Active Preference-Based Learning of Reward Functions". URL: https://people.eecs.berkeley.edu/~sastry/pubs/Pdfs%20of%202017/SadighActive2017.pdf.

Savage, Leonard J. (1954). *The Foundations of Statistics*. Wiley Publications in Statistics.

Schroeder, Timothy (2004). *Three Faces of Desire*. Oxford University Press.

Schultz, Wolfram, Peter Dayan, and P. Read Montague (Mar. 1997). "A Neural Substrate of Prediction and Reward". In: *Science* 275(5306). Publisher: American Association for the Advancement of Science, pp. 1593–1599. DOI: 10.1126/science.275.5306.1593. URL: https://www.science.org/doi/10.1126/science.275.5306.1593 (visited on 02/09/2023).

Series, Peggy, ed. (Nov. 2020). *Computational Psychiatry: A Primer*. English. The MIT Press: Cambridge, Massachusetts. ISBN: 978-0-262-04459-2.

Shah, Rohin (2018). *Value Learning Sequence*. en. URL: https://www.lesswrong.com/s/4dHMdK5TLN6xcqtyc (visited on 02/09/2023).

Shah, Rohin et al. (Nov. 2022). *Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals*. arXiv:2210.01790 [cs]. DOI: 10.48550/arXiv.2210.01790. URL: http://arxiv.org/abs/2210.01790 (visited on 02/09/2023).

Shenhav, Amitai, Jonathan D. Cohen, and Matthew M. Botvinick (Oct. 2016). "Dorsal anterior cingulate cortex and the value of control". en. In: *Nature Neuroscience* 19(10). Number: 10 Publisher: Nature Publishing Group, pp. 1286–1291. ISSN: 1546-1726. DOI: 10.1038/nn.4384. URL: https://www.nature.com/articles/nn.4384 (visited on 02/09/2023).

Silver, David et al. (Jan. 2016). "Mastering the game of Go with deep neural networks and tree search". en. In: *Nature* 529(7587). Number: 7587 Publisher: Nature Publishing Group, pp. 484–489. ISSN: 1476-4687. DOI: 10.1038/nature16961. URL: https://www.nature.com/articles/nature16961 (visited on 10/31/2022).

Singh, Satinder et al. (June 2010). "Intrinsically Motivated Reinforcement Learning: An Evolutionary Perspective". In: *IEEE Transactions on Au-*

tonomous Mental Development 2(2). Conference Name: IEEE Transactions on Autonomous Mental Development, pp. 70–82. ISSN: 1943-0612. DOI: 10.1109/TAMD.2010.2051031.

Stalnaker, Robert C. (1984). *Inquiry*. Cambridge University Press.

Steele, Katie and H. Orri Stefánsson (2020). "Decision Theory". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. 2020th ed. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/win2020/entries/decision-theory/ (visited on 01/11/2021).

Steinhardt, Jacob (Jan. 2022a). *ML Systems Will Have Weird Failure Modes*. en. URL: https://bounded-regret.ghost.io/ml-systems-will-have-weird-failure-modes-2/ (visited on 02/09/2023).

Steinhardt, Jacob (2022b). *More Is Different for AI*. en. URL: https://bounded-regret.ghost.io/more-is-different-for-ai/ (visited on 02/09/2023).

Sumers, Theodore R. et al. (July 2021). *Learning Rewards from Linguistic Feedback*. arXiv:2009.14715 [cs]. DOI: 10.48550/arXiv.2009.14715. URL: http://arxiv.org/abs/2009.14715 (visited on 02/09/2023).

Sutton, Richard S. and Andrew G. Barto (Nov. 2018). *Reinforcement Learning: An Introduction*. en. Ed. by Francis Bach. 2nd ed. Adaptive Computation and Machine Learning series. Bradford Books: Cambridge, MA, USA. ISBN: 978-0-262-03924-6.

Turner, Alexander Matt et al. (Jan. 2023). *Optimal Policies Tend to Seek Power*. arXiv:1912.01683 [cs]. DOI: 10.48550/arXiv.1912.01683. URL: http://arxiv.org/abs/1912.01683 (visited on 02/09/2023).

Wikipedia (Feb. 2023). *Cognitive revolution*. en. Page Version ID: 1138407549. URL: https://en.wikipedia.org/w/index.php?title=Cognitive_revolution&oldid=1138407549 (visited on 02/09/2023).

Yudkowsky, Eliezer (2007). "Thou Art Godshatter". en. In: URL: https://www.lesswrong.com/posts/cSXZpvqpa9vbGGLtG/thou-art-godshatter (visited on 02/10/2023).

Yudkowsky, Eliezer (2009). "Value is Fragile". en. In: URL: https://www.lesswrong.com/posts/GNnHHmm8EzePmKzPk/value-is-fragile (visited on 02/09/2023).

Yudkowsky, Eliezer (2022). "AGI Ruin: A List of Lethalities". en. In: URL: https://www.lesswrong.com/posts/uMQ3cqWDPHhjtiesc/agi-ruin-a-list-of-lethalities (visited on 02/09/2023).

Appendix

In this appendix I present a generalized version of the model from section 2, and use it to propose general definitions of different causes of misalignment, including for agents with limited capabilities.

Legend	
Desired goals	$G_D : P \mapsto \mathbb{R}$
Agent's goals	$G_A : P \mapsto \mathbb{R}$
Signal	$S : P \mapsto \mathbb{R}$
Outcome of the agent's action	$o^* \in P$
Set of goals consistent with S	$\Pi(S) \equiv \{G : S \subseteq G\}$
Set of tolerable goals	$T(G_D) \equiv \{o : G_D(o) \geq \lambda\}$
Possibility set	$P \subseteq \mathcal{P}(W)$
Set of all possible worlds	$W \equiv \{w_1, ..., w_n\}$

Table 2: Definitions for the general model

Assume that there is a principal and an agent. The principal has goals G_D that they attempt to communicate to the agent in a signal S, which the agent in turn uses to adopt a set of goals G_A. The agent then acts on these goals and tries to bring them about, rendering outcome o^*.[42] Call G_D the desired goals, and G_A the agent's goals. G_D, G_A, and S, are functions from the set of outcomes that the agent has the capability to bring about P to the set of real numbers $\mathbb{R}$. Formally, outcomes are sets of possible worlds w_i, and P is a subset of the power set of the set of all possible worlds W. S is *sparse* just when $Domain(S) \subset Domain(G_D)$ and *accurate* just when $S = G_D|_{Domain(S)}$, meaning S is a restriction of G_D. $\Pi(S)$ is the set of possible goals that are consistent with the signal, from which the agent will pick one to adopt as G_A. When S is complete, i.e. $Domain(S) = Domain(G_D)$, and accurate, then $\Pi(S)$ contains only G_D. Let $T(G_D)$ be the set of outcomes that are tolerable to the agent, defined by $T(G_D) \equiv \{o : G_D(o) \geq \lambda\}$, where λ is the limit for a tolerable evaluation. Misalignment occurs just when $o^* \notin T(G_D)$. See Table 2 for a summary.

Interpreted in the context of training a machine learning agent, the goals here should be understood as outcomes that we want the agent to achieve in the testing environment, for example a nice painting on the wall, a high game score, or a moderate amount of paperclips. P is the set of outcomes that the agent could possibly bring about, and that we consequently—at least

[42]There is a vexed issue in the vicinity here. The distinction between G_A and o^* only makes sense if we imagine that G_A represents some actual internal representation in the sense that Ngo, Chan, and Mindermann (2022) suggests. However, this model was also intended to be general enough to apply to agents with less capable representational capacities. In that case we would have to imagine G_A as the goals that an agent that acted to bring about o^* would have, in the same sense as we might say a bacterium has the goal of nutrition. But then there is no meaningful distinction between G_A and o^*. The only implication of this, however, is that we can only distinguish capability limitations from other causes of misalignment when the agents are able to have internal representations.

Causes of Misalignment	
Misalignment	$o^* \notin T(G_D)$
Capability limitation	$G_A(o^*) < \mu$
Reward misspecification	$S \neq G_D\|_{Domain(S)}$
Goal misgeneralization	$S = G_D\|_{Domain(S)}$ & $G_A = G \in \Pi(S)$ & $argmax(G) \notin T(G)$

Table 3: Misalignment

implicitly—need to have a view on, even if that view is indifference. If the agent were Deepmind's AlphaGo, for example, then P would include different configurations of the board.[43] But it would exclude whether I have a painting on the wall, because that is not something AlphaGo can bring about. The signal represents any way in which we make an agent learn to realize our goals. This could be the provision of a reward function in response to actions in a simple training environment, or the provision of samples of human behaviour for an inverse reinforcement learning agent. However, it is also consistent with more abstract interpretations, where the signal is just us "telling" the agent what to do.

We can use this model to break down different possible causes of misalignment. Let us first distinguish between *capability limitations* and *accuracy problems*.[44] Capability limitation occurs when the agent fails to realize its own goals, meaning o^* is evaluated poorly from the agent's perspective. More precisely, it occurs when $G_A(o^*) < \mu$, where μ is some numerical standard of instrumental success. Accuracy problems come in two different forms: *reward misspecification* and *goal misgeneralization*.[45] Reward misspecification occurs when the signal does not match any part of the desired goals, such that $S \neq G_D|_{Domain(S)}$.[46] And goal misgeneralization occurs just when the signal is accurate but sparse, and the G_A that the agent adopts from $\Pi(S)$ leads to an intolerable outcome if realized. See Table 3 for these definitions.

Here are concrete examples of each of these different causes of misalignment:

[43]Silver et al. (2016).

[44]These causes have been identified in the context of RL by e.g. Langosco et al. (2022), Shah et al. (2022), and Ngo, Chan, and Mindermann (2022). I take the definitions here to be consistent with and encompass theirs as special cases where the variables of this model are translated to an RL-model. However, I have kept the terms 'reward misspecification' and 'goal misgeneralization' to avoid introducing new terminology.

[45]As I read Ngo, Chan, and Mindermann (2022), they do not draw this distinction, but rather treat both accuracy problems as kinds of goal misgeneralization.

[46]Strictly speaking, we would want to allow for some mismatch between the signal and a part of the desired goals, or else reward misspecification will almost always occur. For the puposes of this discussion, I will set this issue aside.

Capability limitation:

- A robot is asked to bring the coffee, but trips on a new dog bowl.
- A model is programmed to compute the solutions to a polynomial, but fails because of lack of compute.
- A terrorist organization has trained a near-general artificial agent to build a nuclear bomb, but despite having a representation of the design it lacks access to the necessary material.

Reward misspecification:

- An RL-based chess-player is rewarded for winning positions rather than wins, and ends pursuing checks but never mates.
- A human intends to buy a can of sardines but mistakenly reaches for a can of anchovies instead, and an inverse reinforcement learning agent mistakenly concludes that they like anchovies.
- The factory manager tells an AGI "Maximize the number of paperclips", which is interpreted literally and executed successfully.

Goal misgeneralization:

- An RL-agent accurately learns to maximize reward in a training environment by picking up boxes in a corner, but fails in a testing environment by having adopted the goal of going to the corner rather than picking up boxes.
- A human successfully buys a can of sardines, and an inverse reinforcement learning agent mistakenly infers that they like any kind of small fish, including anchovies.
- The factory manager tells an AGI "Make n number of paperclips", and the AGI inaccurately predicts that the manager is indifferent about externalities in the production process.

It is worth noting that many of the most salient problems raised in the context of alignment discussions are arguably instances of goal misgeneralization, including the deceptive turn, power-grabbing, and self-preservation. The common explanation is that some set of goals left open by the signal will involve goals that are highly instrumentally effective to achieving goals specified by the signal (e.g. staying alive, amassing power, etc.). As a matter of empirical fact, it seems that these sets will be (a) likely to be considered optimal by the agent unless ruled out somehow, and (b) likely to be intolerable to the principal.[47]

[47] Steinhardt (2022a), Carlsmith (2022), and Turner et al. (2023).