Alignment as a Dynamic Process

Paul de Font-Reaulx
Department of Philosophy
University of Michigan, Ann Arbor
pauldfr@umich.edu

Abstract

Most learning AIs today have exogenously given and fixed aims which they gradually learn to optimize for. It has been an assumption in alignment research that artificial general intelligences of the kind that could pose an X-risk would too. On this assumption, value alignment becomes the task of finding the right set of aims before we allow the agent to act. However, an agent can also have aims that fundamentally change during their lifetime. The task of aligning such agents is not one of specifying a set of aims, but of designing a meta-function that guides the agent's developing aims to an equilibrium that produces behaviour aligned with our human values. If artificial general intelligences would possess such dynamic aims, then this has significant implications for the kind of alignment research we should conduct today. In this paper, I argue that there is a substantial probability that artificial general intelligences would have such dynamic aims, and in response I articulate an agenda for dynamic alignment research.

1 Questioning a Static View of Alignment

The alignment problem is commonly framed as a problem of precisely fixing the right aims to an extremely capable artificial agent—what I will, following convention, call an artificial general intelligence (AGI)—before we allow it to act. The challenge is to specify those aims so that when the agent attempts to achieve them, the result will be one that is in line with our human values. This framing involves two phases. First, we specify fundamental aims to be given to the agent. On a reinforcement learning (RL) architecture, these are represented by the agent's reward function. Second, we let the agent act and attempt to achieve those aims. The aims themselves, however, remain fixed during the second phase. Call this a *static* view of alignment.

Imagine, by contrast, that the aims of the agent changed over its lifetime. In an RL-framework framework, this would be an agent with a reward function that changed over time as a function of its actions. For such an agent, it would not make sense to adopt the static view of alignment, because the aims that we meticulously specified in the first state would just change as the agent was allowed to act in the second stage. For such agents, we must instead conceive of alignment as a dynamic process. On this *dynamic* view of alignment, the alignment problem is not the challenge of specifying a perfect set of aims. Instead, it is the challenge of designing an agent who will themselves develop aims that in equilibrium produce behaviour in line with our values. Note that an agent with such dynamic aims is distinct from an agent that develops static aims by approximating human values via revealed preference,[15] or one that is trained to learn cooperative dynamics.[14]

In some ways, aligning an agent with dynamic aims would be far harder than aligning one with static aims, in virtue of the complexity plausibly involved in the dynamic process. On the other hand, it would also open new avenues for how to approach alignment which might resolve current gridlocks.[4] This makes it unclear whether we should hope that AGIs would have dynamic or static aims, or attempt to design them one way or another. As I argue below, we do in fact have reason to believe that they would have dynamic rather than static aims, and we should diversify our strategies

ML Safety Workshop, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

and invest resources into understanding dynamic alignment better. That would be a significant shift from how alignment research is pursued today, where the static view is widely assumed as the only alternative.

In the rest of this paper, I first argue that humans plausibly have dynamic aims that change over their lifetime. I then present an argument for why this gives us reason to think that AGIs would have dynamic aims as well. Finally, I discuss what this means for alignment research and present an agenda of questions to be answered.

2 Human and Artificial Aims

Consider the aims of some popular artificial agents, both real and hypothetical. For example, consider an artificial agent like Deepmind's Atari playing reinforcement learner.[13] This agent has an exogenously specified reward function in the form of the game score, which represents its fundamental aims. These aims don't change during the lifetime of the agent and can be considered "hardwired" into its system. The same can be said for Nick Bostrom's infamous paperclip maximizer, which has as its fundamental aim the maximization of the number of paperclips in the universe.[2] This agent might adopt many sub-goals to be instrumentally effective at achieving its aims—including the eradication of humanity—but the fundamental aim of paperclip maximizing itself does not change.

By contrast, it is not clear that humans work like this. Rather, at least on a naïve understanding, it seems that human aims sometimes change over our lifetime. In some cases, this seems to happen as a product of socialization, such as when we internalize the norms and values of our surrounding community. In other cases, it seems like we deliberately cause changes our aims by deliberation, such as when we become vegetarians.[1] If these cases involve a change in our most fundamental aims—our reward function—then that would make humans very different kinds of agents from Deepmind's Atari-player, Bostrom's paperclip maximizer, and most other AIs represented in discussions.

One could object and say that fundamental humans aims do not in fact change. On this proposal, humans are born with innate fundamental aims which are fixed throughout their lifetime, and only change in terms of finding better ways to obtain them. A person defending this view might for example argue that when we decide to become vegetarian, we have not had any change of fundamental aims. Rather, we have come to believe that being vegetarian is an instrumentally more effective way to achieve our innate aims (say, of being healthy and somewhat altruistic).

This view of human aims could well be right, and should be considered a live option. However, it is certainly not obviously right, and it leaves us struggling to explain much human behaviour. Specifically, it is unclear whether there are any given set of basic aims to which the diaspora of human aspirations are reducible. Suppose, for example, that we are born with the innate aim to maximize the welfare of ourselves and our close kin. It is then hard to see how we could end up with humans who are willing to give resources to distant strangers, or who suffer for beauty or science, and who themselves view these actions as successful. Arguably, a more natural explanation would be that their fundamental aims have changed over their lifetime in a way that makes these projects rational by the lights of those new aims. See Appendix A for a short response to the objection that reward corresponds to pleasure and pain.

If this is correct, that means that humans have dynamic aims. If so, then we would not be able to accurately model a human with a static reward function, as we do with the artificial agents above. Rather, if we were to accurately model a human in an RL-framework we would need to model the reward function as itself determined by some higher-order meta-function. This meta-function would govern how the reward function would develop over time. There are some precedents for such a meta-function,[3] but it is generally not deeply explored in the literature.[2]

For all that we have said here, this meta-function could take many different forms. A simple example, however, might be a one that produces adaptive rewards, analogous to the notion of adaptive preferences.[8] Such a function would increase the reward for outcomes that the agent encounters more frequently, thus making it value its environment higher. Even this very simple suggestion might allow us to make sense of phenomena like the internalization of values prevalent in our close

[1] In the relevant philosophical literature, this is known as *self-constitution*.[12]

[2] An exception is Kleiman-Weiner and his colleagues' work on using hierarchical Bayesian models to capture aims that develop in response to a social environment.[11]

social context. Whether we need more complex meta-functions to model humans, for example, is a question for future research.

3 Would AGIs have Dynamic Aims?

If we can model humans in the hierarchical and dynamic fashion suggestion above, then presumably we can also design effective artificial agents with dynamic aims as well. For the purposes of AI-alignment, however, we are primarily concerned with AGIs in particular, and whether they would have dynamic aims. It might be that this is up to the designers, such that AGIs could equally well be designed either with static or dynamic aims. In other words, it might be that whether an artificial agent has static or dynamic aims makes no difference to the probability of it being generally intelligent. Call this claim *Orthogonality*.[3]

If Orthogonality is true, then we would have reason to aspire to design AGIs with whichever kind of aims—static or dynamic—that seems most feasible to align. To determine that question, we need to understand the alternative of dynamic alignment better. In other words, if Orthogonality is true, then we should plausibly invest resources into dynamic alignment research.

However, I think there is reason to believe that Orthogonality is false. Specifically, I think that on our current evidence it's more likely that AGIs would have dynamic aims than that they would have static aims. But if so, then dynamic alignment research would be even more important than if Orthogonality was false, because it would be the primary kind of alignment we should expect to be required to do. Here is a schematic argument for the claim that general intelligence correlates with dynamic aims in artificial agents:

> (1) Humans have changing aims.
>
> (2) The cognitive processes of humans provide good evidence about what the cognitive processes of an artificial agent would need to look like to be generally intelligent.
>
> (3) If the processes that underlie a capacity to change aims also underlie an agent's general intelligence, then that agent could not easily be generally intelligent without having changing aims.
>
> (4) The processes that underlie humans' capacity for changing aims also underlie humans' general intelligence.
>
> (C) AIs could not easily be generally intelligent without having changing aims.

I will briefly articulate the components of the argument a bit more. (1) is defended in the previous section. On (2), human cognition is our primary existence proof for the possibility of advanced general intelligence. All else equal, it would therefore be very surprising if we managed to design an agent that demonstrated a capacity for general intelligence, given realistic computational constraints, that bore no resemblance to the computational architecture of human cognition. How closely artificial cognition would plausibly need to resemble human cognition is an open question. The upshot, however, is that if we learn that some process or component is essential to human general intelligence, that gives us defeasible evidence that something analogous would be necessary for artificial general intelligence. However, see Appendix B for some criticism.

On (3), if an agent has a capacity to have their aims changed, then there are some computational processes which underlie and cause this change. However, if those processes are also involved in other capacities, such as the agent's capacity to instrumentally bring about its aims, then we should not expect to see one capacity without the other. For example, in humans executive function processes like inhibitory control are involved both in prudential planning (e.g. not burning your savings on a night out) and social behaviour (e.g. being polite to someone whom you would prefer to tell off).[7] This means that if we see someone who is myopic in their decisions, we should also expect them to be more antisocial, all else equal.

[3] This is different from what Nick Bostrom calls 'the orthogonality thesis'.[1] His is a claim about the relation between capacities and aims, namely that any level of instrumental rationality can be combined with any aim. Orthogonality as considered here is a claim about the relation of two capacities, namely the capacity to change aims and be instrumentally rational.

(4) is a highly speculative premise. As noted in the previous section, it is an open question whether human aims are fundamentally dynamic, and there is basically no research on the processes that would underlie the change of our aims, if they were dynamic. What we can say is that it would at least be surprising if the processes underlying our putative aim-changes and our instrumental rationality were completely disconnected. For example, this would imply that the processes by which we internalize a norm would not be involved in deliberation on whether we should follow that norm for instrumental reasons. For instance, the process by which we might internalize the claim that stealing is wrong would be disconnected from the process by which we see that stealing might get us into trouble. It seems plausible, however, that instrumental considerations might be part of the internalization process, i.e. that prudentially beneficial norms are easier to "swallow" than highly self-sacrificial norms.[16] Against this backdrop, we should at the very least not assume that we can easily have human general intelligence without human dynamic aims.

If we accept these claims, then we should conclude that Orthogonality is false, and that AGIs are more likely to have dynamic aims than static aims. I want to stress, however, that I do not take this to be a decisive argument. It is rather likely—if I had to bet money on it, I would give it around 65% probability[4]—that at least one of the premises above is false. However, given that these claims are based on a dearth of research, I expect significant value in further research into whether they are true.

4 A Plan for Dynamic Alignment Research

I have argued that AGIs could be designed with dynamic aims, and that aligning such an AGI would require a significantly different kind of approach to what I have called the static view of alignment, which is the dominant conception of alignment today. I have also argued that we have reason to think that AGIs not only *could* have dynamic aims, but also that they relatively likely *would*.

This raises the question of what we should do now. Regardless of whether AGIs would need to have dynamic aims, it seems we have reason to invest resources into understanding dynamic alignment. Assuming an RL framework, this will mean research into how we can design meta-functions that will produce reward functions for the AGI conducive to our values, which constitute equilibria in the environment that it will share with us. For example, this might mean finding a meta-function which will make an AGI develop cooperative and altruistic reward functions robustly across different settings.

However, in order to understand how these meta-functions could work, and in order to gauge how seriously we need to consider AGIs with dynamic aims, we need to better understand the questions underlying the argument that I have made here. Since the possibility of dynamic aims is largely given credence by the hypothesis that humans have them, I expect there to be significant value in better understanding human aims and the cognitive processes that underlie their formation. This leaves us with a set of questions that I believe to constitute a possible research program. Here are some of them:

- Do humans have fundamentally dynamic aims?
- What are the cognitive processes by which human aims change, if they do, and how do they relate to different domains of instrumental rationality?
- How do we best model such dynamic processes in a computational framework?
- How do we align an agent's actions with our values on this dynamic model?

If the argument made here is sound, then research into these questions has significant expected value. Ideally, this should be conducted in close collaboration with researchers in other disciplines, especially neuroscientists and cognitive scientists who can contribute to the development of precise models of human cognition, so that these can be used to guide the development of safe artificial agents. That is likely to contribute not only to dynamic alignment, but also to solving moral decision-making and value clarification.[10]

[4] My credences for the truth of the premises are currently roughly: $P(1) = .7$, $P(2) = .9$, $P(3) = .9$, and $P(4) = .6$. This gives a probability of 34.02% that all premises are true.

References

[1] Nick Bostrom. The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents. *Minds and Machines*, 22(2):71–85, 2012. Publisher: Springer Verlag.

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

[3] Jeshua Bratman, Satinder Singh, Jonathan Sorg, and Richard Lewis. Strong mitigation: nesting search for good policies within search for good reward. pages 407–414, June 2012.

[4] Joseph Carlsmith. Is Power-Seeking AI an Existential Risk?, June 2022. arXiv:2206.13353 [cs].

[5] Peter Carruthers. Valence and Value. *Philosophy and Phenomenological Research*, 97(3):658–680, 2018.

[6] Richard Dawkins. *Flights of Fancy: Defying Gravity by Design and Evolution*. Apollo, May 2022.

[7] Adele Diamond. Executive Functions. *Annual review of psychology*, 64:135–168, 2013.

[8] Jon Elster. Utilitarianism and the Genesis of Wants. In Amartya Kumar Sen and Bernard Arthur Owen Williams, editors, *Utilitarianism and Beyond*, pages 219–238. Cambridge University Press, 1982.

[9] Peter Godfrey-Smith. *Other Minds: The Octopus, the Sea, and the Deep Origins of Consciousness*. Farrar, Straus and Giroux, New York, first edition edition, December 2016.

[10] Dan Hendrycks and Mantas Mazeika. X-Risk Analysis for AI Research, September 2022. arXiv:2206.05862 [cs].

[11] Max Kleiman-Weiner, Rebecca Saxe, and Joshua B. Tenenbaum. Learning a commonsense moral theory. *Cognition*, 167:107–123, October 2017.

[12] Christine M. Korsgaard. *Self-Constitution: Agency, Identity, and Integrity*. Oxford University Press, Oxford ; New York, 1st edition edition, June 2009.

[13] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing Atari with Deep Reinforcement Learning, December 2013. arXiv:1312.5602 [cs].

[14] Peter Railton. Ethical Learning, Natural and Artificial. In S. Matthew Liao, editor, *Ethics of Artificial Intelligence*, pages 45–78. Oxford University Press, September 2020.

[15] Stuart Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin Books, October 2019.

[16] Tadeusz Wieslaw Zawidzki. *Mindshaping: A New Framework for Understanding Human Social Cognition*. Bradford, 2013.

Appendix A: Reward as Pleasure and Pain

One popular suggestion is that the reward function corresponds in natural agents to the presence of pleasure and absence of pain. This suggestion is implausible. One reason is the problem of reduction raised in section 2, i.e. that few goals can be reduced to an expected good balance of pain and pleasure. For example, the parent who sacrifices their life for their child does not expect much pleasure for themselves in the future. Another additional reason, however, is that this gets the function of hedonic signals wrong. Events are not good for us because they are pleasurable. Rather, it's more accurate to say that they are pleasurable because they are good for us, even though this as well distorts the picture somewhat.[5]

Appendix B: Will AGI Cognition Resemble Human Cognition?

There are plenty of concerns one can raise against the suggestion that AGI would resemble human cognition. Firstly, it is not clear that humans are the only existence proof of general intelligence. Octopodes in particular offer an example of a seemingly impressive intelligence that is evolutionarily very distant from us.[9] However, while we need to study such animals in greater detail to be certain, they do not seem to provide real alternative examples to the level of capacity that we are interested in here. Secondly, technologies like heavier than air flight were not invented by attempting to mimic birds, which might raise doubts about the need to mimic human cognition to create AI. We should note two things in response. First, general intelligence is far less ubiquitous a capacity than flying—which we see instantiated in many different ways in birds, insects, and bats—so “hitting” that capacity by sheer luck seems far less likely, suggesting we would need to aim for it more precisely. Second, in the case of flight, humans arguably did learn something about the principles of flight from animals, but found better ways to instantiate them given the constraints of heavy materials.[6]