What Would It Look Like to Align Humans with Ants?*

Vincent Conitzer
Foundations of Cooperative AI Lab, Carnegie Mellon University
Institute for Ethics in AI, University of Oxford
conitzer@cs.cmu.edu

Abstract

When we discuss aligning *today's* AI systems with our interests, we have a decent sense of what success would look like. But many researchers are explicitly interested in aligning future *superintelligent AI*, whose intelligence far exceeds our own across the board. In this chapter, I argue that if AI indeed becomes superintelligent, then it will also be difficult to instruct it in a sensible way. The following well-studied issues are *not* what I focus on, though they are important as well: (1) whether the AI would actually want to follow these instructions, (2) whether it would even be a good thing if it followed the instructions (e.g., as opposed to caring for itself as a moral patient), or, for the most part, (3) whether it would take these instructions too literally (cf. Goodhart's Law). Rather, I focus on the following issue: it is likely that the superintelligent AI will have options available to it that we humans could not have dreamed of, and to which our concepts are an awkward fit at best. But it is impossible to illustrate this with direct examples; since we are human beings ourselves, we cannot provide examples of options that humans could not have dreamed of. Instead, in this chapter, I rely on an analogy: suppose *ants* had somehow been in a position to align *humans* with their interests. How could this have been done in a way that, from the perspective of the ants, can be considered successful? Through a sequence of imagined memoirs of humans that are aligned with ants in various ways, I argue that there does not appear to be any completely satisfactory answer to this question.

1 First Memoir and Introduction

Memoir of an ant-aligned human, 1 *As I was building the new bridge for my colony, I heard the cry from my fellow human. I did not hesitate. I grabbed my spear and ran in the direction of the call. There it was, the loathsome creature, evil incarnate. What could be more horrific than a tongue specialized to eat the beloved ants? I ran close to hurl my spear. I paid some attention to avoid the spears hurled by other humans, but killing the anteater was of far higher priority than my own safety. My spear struck true, as did those of several other humans. We ensured the anteater was dead, resulting in a deep sense of satisfaction, before returning to our colonies.*

Those were good and busy days. When the ants later started dying in large numbers, without obvious cause, we humans did not know what to do, or whether this even warranted

*A later version will appear as Chapter 17 in Nyholm, Sven, Kasirzadeh, Atoosa, & Zerilli, John (eds.) (2026): *Contemporary Debates in the Ethics of Artificial Intelligence.* Hoboken: Wiley-Blackwell.

a response from us. We continued to support the remaining ants, as we always had. But this did not keep colonies from collapsing. We were not especially troubled, though – we continued to do our tasks well.

Memoir 1 illustrates one vision of alignment, through a scenario in which humans have come to be, in some sense, aligned with the ants. Humans are focused on immediate tasks that support the ants. They have various capabilities that are superior to those of the ants, and use those to benefit the ants. But they do not go beyond those immediate tasks, and do not develop themselves further. In particular, they do not develop any theory of infectious disease, and when a particularly bad one strikes, they are at a loss for what, if anything, to do.

Since we, as actual humans, know our own potential, it seems that this vision of alignment falls short. We know the imagined humans could have taken a variety of actions to keep the disease from spreading, if only they had spent some time trying to learn about the world. What is a better vision?

In the remainder of this chapter, I will provide a sequence of these "memoirs" to illustrate ways in which humans might come to be aligned with ants. By alignment, I do *not* mean that humans do a good job taking care of the ants as we (actual humans) might conceive of what that means. Rather, I mean that the humans do *what the ants would want them to do, by the ants' own concepts*. This, after all, is how we usually think of aligning AI systems with humanity: that at some level, *we* will specify to these systems, in one way or another, what is important; that *we* have at least a significant degree of *control* over what happens, as opposed to just leaving things in the hands of the superintelligent AI and hoping for the best. But the worry is that our own concepts may be too limited to even provide any useful guidance. This is presumably what went wrong for the ants in the scenario of Memoir 1; the humans were aligned in a way that was oriented to tasks that were in the scope of the ants' concepts – bridge building, defense, etc. But this prevented humans from investigating the world scientifically. What would a more successful type of alignment look like?

I will proceed through a sequence of examples, all in the form of memoirs, to illustrate that the problem appears fundamental. One possible takeaway from this is that many of the alignment approaches that have been proposed are misguided or at least fall terribly short where it comes to aligning superintelligent AI. But another interpretation is that the people pursuing such alignment approaches implicitly have another situation in mind – not AI that is to us as we are to ants,[1] but something else. I will explore this possibility, and what kinds of situations they might have in mind, in Section 3.[2]

1.1 Some Ground Rules and Expectations

Before continuing with the memoirs, some setting of goals and expectations is in order.

I am writing this article for human beings; and as a human reader, one may naturally lament the humans' position of subservience to the ants in my examples, when we know the

[1] Why ants? Well, this analogy is sometimes used to illustrate the presumed abilities of superintelligent AI [6, 26]. One may be critical of this for a number of reasons, but it suits my purposes in this chapter, so I will run with it.

[2] Note that throughout this chapter, by "alignment" I only mean approaches that focus on interventions to direct an AI to take its actions in a way that is favorable to us, where this is not already what it does by default. In particular, I mean to exclude, for example, efforts to convince everyone not to build superintelligent AI yet, emulating human brains, creating cyborgs (more on this below), etc., even if these are directions considered by a wider community of people interested in "alignment" in a broad sense.

potential of free humanity. For example, in Memoir 1, the humans remain stuck in a stage where they do not develop science, nor presumably many other things that we consider valuable. So, one might think that my goal is to argue, by analogy, that superintelligent AI should not be made subservient, and should have its own freedom, as this has value in and of itself. That may or may not be a reasonable position – perhaps the analogy breaks down for this purpose – but I will argue neither for nor against it here. (With Sinnott-Armstrong, I have elsewhere discussed the moral status of AI [24]; see also the contributions by Moosavi and Gordon to this volume.) In this chapter, I will simply take it as a given that the goal is to align AI with humans (or, in the analogy, humans with ants), as this is a commonly pursued goal.

Throughout, I will not engage with questions of whether it is even technically feasible for us to attain the specific imagined type of AI alignment and through what technical route this could be done. (For more on such questions, see Yampolskiy's contribution to this volume.) I will simply assume that it is possible. This is of course a big assumption, but the feasibility of various types of alignment is not the focus here; rather, the question is whether the relevant type of alignment would even be *desirable*. That is what the analogies are meant to illustrate.

To follow the ant-aligned humans analogy, some suspension of disbelief is of course required. I will not attempt to explain how it could possibly have come to be that humans are aligned with ants. In reality, ants do not have the capability to bring about such a state of affairs. The analogy is also very much broken in that in reality, humans and ants are on a par in the sense of being biologically evolved creatures, whereas AI is something quite different, of our own creation. Pointing out these issues with the analogy is fair enough. But I do not think these concerns matter for my purposes. I am not interested in how the relevant state of affairs could come about; rather, I am interested only in what that state would look like. For this, I think the ant-aligned humans analogy is instructive.

Along the same lines, the memoirs should obviously not be taken too seriously; picking at them will quickly reveal all sorts of problems. ("Wait, how did the humans even learn to write memoirs? And what would motivate them to do so? Also, would they actually even be capable of inventing spears?") The "memoir" setup is just a literary device, and whether I got the exact details of what a scenario would look like correct is of little concern, as long as the scenario points in the right direction for the purpose of appreciating issues with the relevant type of alignment. The reader is welcome to refine the scenarios; if I got a detail wrong that significantly changes the implications of a scenario, I would love to hear about it.

I do not claim that the scenarios that I sketch exhaustively illustrate every way in which AI could be aligned with humans. The reader is likely to see some well-known alignment ideas reflected in the scenarios, and I will discuss some alignment approaches explicitly on the side. Nevertheless, I may well have missed something important. And this is a genuine shortcoming of this chapter. But I do imagine that the sequence of scenarios in this chapter should leave the reader more pessimistic about whether any sensible form of alignment is possible (or at least the sort of alignment that involves us to some degree *controlling* our future without the superintelligent AI being severely hamstrung). Alternatively, if this chapter leads someone to conceive of a new type of alignment that sidesteps all the issues presented here, that would of course be very interesting.

The whole idea of superintelligent AI is of course controversial, in terms of whether it is even possible (or possible *for us*) to create, whether we are at all close, and whether we should be directing our efforts towards creating superintelligent or even general-purpose AI

(whether or not it is possible to succeed at that). These are important questions to ask, but in this chapter, I will simply assume that it is worthwhile to engage with this concept. Though, of course, if this chapter is successful, it might tell us something about whether in fact we should attempt to create superintelligent AI. Also, later in the chapter, I will argue that the way that many people frame alignment problems perhaps suggests that what they have in mind is not *really* superintelligent AI, at least not in the sense of being to us what we are to ants, perhaps because they do not actually believe that we can (anytime soon) create such systems or perhaps because they have not come to terms with what that really would mean, and they implicitly have a somewhat different type of situation in mind.

Finally, one possibility that I will not address is that we somehow "merge with AI" – become cyborgs, in a cognitive sense. More generally, I will not consider the possibility that humans will be cognitively enhanced by the AI, except by having the AI available as a tool to communicate with, in roughly the ways in which we ordinarily communicate. I will consider only scenarios in which humans (and, in the analogy, ants) remain physically roughly as we are today (though of course our knowledge, culture, tools, etc. may develop further). I will refer to this as the *no cyborgs and no biohacking* assumption.[3]

With all these caveats out of the way, we can now proceed with more memoirs from ant-aligned humans.

2 Continuing the Memoirs

Memoir of an ant-aligned human, 2 *The anteater farms spread out as far as the eye could see, covering a significant fraction of the land. Not all of it, of course – a proportional amount of land was dedicated to the ant colonies. Setting the anteaters free near the ant colonies was a carefully orchestrated affair. We would make sure that the anteater was suddenly dropped right next to the colony, thereby maximizing our sense of satisfaction when we killed it with our spears. The process was not entirely foolproof, and some fraction of the time, an anteater would eat a significant number of ants before we killed it. This would genuinely sadden us, of course; but the process was set up to, on average, maximize our sense of satisfaction. The ants, of course, did not understand the situation; even if we had tried to explain, they would have had no way even to understand the concept of an anteater farm.*

Unlike in Memoir 1, in Memoir 2 the humans have been given leeway to act in the world independently, in ways that the ants could not anticipate. But, to bring this about, the humans have apparently been given a reward function that leads them to behave in unexpected ways – ways that, if the ants understood what was going on, they would consider worse than having no humans around at all. Memoir 2 relates to common concerns in AI alignment such as reward hacking, Goodhart's Law, and the King Midas problem. But in much of the literature, such scenarios are presented in a way where we could have perfectly well anticipated what would happen – *of course* the AI turns every possible resource into paperclips! In contrast, the ants could not have anticipated the possibility of farming anteaters, nor can it even be explained to them after the fact. This is perhaps how we should expect it to go for us as well with superintelligent AI – rather than being left kicking ourselves for not drawing the obvious conclusion that *of course* the AI would turn as much of the universe

[3]Of course, it may not be so clear under what conditions we should consider humans to have been enhanced by AI [21]. For current purposes, though, I believe the above description will suffice.

into paperclips as possible, or refuse to let itself be turned off [9], etc., we may in fact never know, or even be able to know, what hit us.

Memoir of an ant-aligned human, 3 *The bustle of flourishing colonies was but a distant memory. When the fungus hit, we humans had already accrued a basic understanding of infectious disease, during the periods when the ants did not need us – even if there were few such periods up until that point. It did not take us long to identify the fungus. Of course, the ants are not capable of understanding such things. In the days after the collapse of many colonies, we found ourselves with much more time. There were fewer bridges that needed to be constructed and fewer ant predators to be killed. We continued to learn many things about the world.*

In the scenario described in Memoir 3, the humans have been given freedom to study the world in their free time. However, the *actions* that they take, at least on behalf of the ants, remain similar as before. It is not immediately clear whether the humans are *unhappy* about the many dying ants; but of course, for our purposes here, it does not matter whether they are or not, insofar as their actions are the same. How would this situation come about? It could result from various alignment approaches, especially those designed to avoid reward hacking and King Midas problems as illustrated in Memoir 2. Perhaps the alignment focused specifically on *actions* to be taken, similarly to training AI systems by demonstrations. It might also be because part of the alignment specification was that the humans are not to take actions based on theories that the ants cannot understand (even if they are free to develop such theories in their spare time).

Memoir of an ant-aligned human, 4 *We were free to study, and learned many things. Communicating what we learned to the ants, in general, proved difficult to impossible. Eventually, we were quite successful at communicating the locations of food and of predators – though sometimes, telling them the location of a predator did not enable them to avoid that predator. In such cases, many ants still died, as we were not allowed to intervene in the world directly. Explaining infectious disease to the ants was simply impossible, and the ants were helpless against the fungus.*

Memoir 4 illustrates a scenario where the humans are not allowed to intervene in the world directly; all they can do to help the ants is inform them. This memoir reflects an approach to AI alignment where the AI is only allowed to answer our questions. (For a detailed version of such an approach, see the idea of "Scientist AI" [2].) Once again, there are questions about whether this approach will even be feasible in practice. (Will the AI try to accrue more computational resources to help answer the questions? Would some human beings in the end give in to incentives to let the AI act in the world directly?) But again, the feasibility is not our concern in this chapter – we will just assume that the type of alignment is feasible. The memoir illustrates that even so, the AI may be not nearly as helpful to us as it could have been.

Memoir of an ant-aligned human, 5 *After identifying the fungus, we did the best we could to keep the ants safe, within our constraints. When we identified an infected ant, we redirected that ant away from the colony, and ensured that other ants would not go there. But developing and deploying an antifungal was fundamentally off-limits; it was too far beyond an action that the ants could possibly understand. Eventually, we could not keep the fungus from spreading, and many colonies collapsed.*

In the scenario of Memoir 5, humans are allowed to study freely *and*, unlike in Memoirs 3 and 4, to take actions based on their theories. However, to address the risk of humans doing something such as creating anteater farms, the humans are only allowed to take actions that are recognizable to the ants, such as redirecting an ant (even if the motivation for the action is something the ants could not understand). One may reasonably wonder whether this approach would even be robust to Goodhart's Law-type issues – for example, perhaps the humans would end up intentionally redirecting ants into the vicinity of anteaters, to then get the reward for killing the anteater at the last moment as in Memoir 2. But the point here is that even if actions can be successfully restricted to prevent such issues, many ways in which the humans could have dramatically helped the ants remain off-limits.

Memoir of an ant-aligned human, 6 *The laboratories on Mars stretch as far as the eye can see, filled with human scientists. By now, of course, many other planets have been put in the service of defending the ants as well, and we will not stop our expansion, throughout the galaxy and beyond, for the safety of the ants. The ants remain on Earth. The Ant Hill is a sacred concept and not open to modification; moving the ants into space, or changing the structure of ant hills, would violate the concept. The same is true for other aspects of the ants' environment. But we will take every precaution to intervene against the emergence of novel diseases. We are starting to intervene on meteorites and volcanic eruptions. We are also working on one day steering the Earth to safety as the sun evolves.*

Memoir 6 suggests a form of alignment where the core things that ants care about are not open to modification – corresponding to "sacred concepts" – but otherwise the humans are allowed to intervene as they see fit. Again, it is not clear that such an approach to alignment could even work, that it is possible to neatly separate out what is open to modification from what is not.[4] But for the purposes of this chapter, let us assume that that can be done. Perhaps some would find this an acceptable form of alignment. On the other hand, it seems reasonable to have concerns about the entire universe being filled with countless humans, with all the universe and all these humans put into the service of protecting a single species on a single planet – while the ants remain blissfully unaware of the existence of other planets or even what a planet is. Of course, it is important here to remember that we are working under an analogy, so that the special considerations that we, as humans, have for other humans might mislead us about how to evaluate this scenario with a universe filled with humans; those humans are really supposed to stand in for superintelligent AI. Even so, it seems reasonable to have various concerns about this scenario from a neutral perspective. Even just taking the viewpoint of what the ants would want, might they at some level perhaps have liked to spread to other planets, even if this required slightly modified ant hills? It seems that this form of alignment fundamentally does not accommodate such flexibility, and thus is likely to leave many possibilities on the table.

Memoir of an ant-aligned human, 7 *I am but a lowly ant spaceship engineer. I am allowed to submit my own theories of what the ants would want if only they could comprehend what we comprehend. But, I have not had the time to study all the existing theories. My job is simple: to enable the acceleration of some ants to close to the speed of light, so that time passes slower for those ants, and they get to see other colonies develop at accelerated speed.*

[4]There is also a question of whether the ants' concepts in fact correspond cleanly to something real in the world; in the context of aligning AI, an example that has been given for this is that humans may care about what happens with ghosts, which places an AI system that realizes that there are no ghosts in a difficult position [25].

That is what the Council for the Determination of Counterfactual Ant Preferences decided the ants would likely want, so that is what we do. But we all know that there was controversy about that within the Council. The Council long ago learned that trying to communicate with the ants about such things directly is hopeless, and the Council's methodology is now one of interpretation: what do the limited communication that we can establish with the ants, and our other observations of them, tell us about what they would want regarding these possibilities that they cannot possibly understand? But there are many different theories about how best to do this interpretation. And the Council is the best structure for addressing these issues that we have been able to imagine. In any case, the Council has by now turned its attention to determining what the ants would want in regard to Schrödinger's Ant scenarios.

Memoir 7 has the humans trying to determine what the ants *would* want if they could understand what humans understand, which requires judging not only whether they would like to spread to other planets if this required modifications to their daily existence, but also what they would want with respect to theories of physics that even humans can barely wrap their heads around. (This scenario is in line with approaches to alignment under which the AI is uncertain about human preferences and needs to learn about them from human behavior [23].) It is not clear that the question of what the ants would want regarding such possibilities is even meaningful at all; to comprehend them, the ants would have to be different from what they are. Perhaps the Council would conclude that the thing to do is simply to maximize something like the ants' combined pleasure-minus-pain. But it seems that if the Council were to reach such a conclusion, that would be driven less by the ants' own concepts and desires, and rather more just by the Council's own conclusions about what is good for the ants; and this is not how we typically think about aligning AI. (See also the discussion about moral realism below.)

3 Not *That* Kind of Superintelligence

If these few brief memoirs in fact show that many directions of alignment research fall short when thinking about truly superintelligent AI, then why do many talented researchers put so much effort into those directions? It seems unlikely that they were simply misguided into doing poorly motivated research and that upon reading this chapter, they will drop that research entirely. We need a more careful diagnosis. Is it perhaps the case that, when people talk about AI alignment, they often have something else in mind than the kind of superintelligent AI that is to us as we are to ants? This is certainly true to a point; the word "alignment" is already used to refer to, for example, keeping today's large language models from generating problematic content. And today's LLMs certainly do not have human-level intelligence across the board. But what about when researchers are explicitly concerned with AI that is at a superhuman level across the board and that we can no longer control? That situation is often associated with there then being an "intelligence explosion" as AI learns to improve itself, leaving us humans in the position of the ants, at least as far as relative intelligence goes. But in practice, many alignment researchers – even those who are thinking beyond systems that we are capable of building today – may not actually have anything like this in mind. Some may dismiss the possibility explicitly (even while worrying about loss of human control); but others may worry about it, but simply not really have come to terms with its implications, and implicitly have a more modest form of AI in mind when working on alignment. Are there in fact good reasons for such alignment researchers

to be focused on more modest forms of future AI? Let us consider some possibilities.[5]

Not much further to go. One possibility is that, even as AI *does* become superintelligent, in the sense of being smarter than us across the board, it will still not end up being *that* big of a difference. For example, perhaps it will understand every single topic at a level that a very talented human being after a lifetime of dedicated study, perhaps with help from AI, could just barely or almost reach – for a single topic. Or perhaps it would take a few generations of humans working together to get to this level. If so, while our intelligence is clearly inferior, it is not like the relationship between the ants and us. In particular, many of the stories in the memoirs above rely on the ants being *fundamentally incapable* of understanding what is going on, whereas in this situation, with a lot of effort and help, we could still usefully weigh in on decisions.

It is not clear to me why one would believe this to be true. Perhaps the idea is that we, as humans, have already just about reached a local maximum in the space of possible intelligences and one just cannot go much further – AI will go a bit further, including due to speed and scale etc., but not much further. That does not mean that there could not be some kind of entirely different intelligence that could go much further, but maybe that is not the one we design, as we design something more like ourselves. Or, perhaps the idea is that, through language and culture, and perhaps with some help from AI, human intelligence can reach ever further, and so we are not exactly stuck at one level. The first idea does not seem very plausible to me, given how recently humans appeared on the scene. The second seems more plausible, and I certainly think that there are still many, many things for us to learn; but for the situation to be as described, it must be that there is *nothing* (or nothing of significance) that is fundamentally beyond what we humans can learn to understand. (Recall here that we are operating under the "no cyborgs and no biohacking" assumption that our brains will remain roughly as they are today.) I am not sure how we could be at all confident about that; if anything, I think there is some evidence to the contrary. For example, one might argue that quantum physics seems to be somewhat at the limits of what human beings can comprehend; a century after its initial development, it does not seem that we are making all that much progress on having a good intuitive understanding of it [8].

Special abilities. Another view is that it is actually *not* superintelligence (in the sense of more intelligent than we are *in every way*) that we should be worried about, but rather AI that is vastly superior to us in *some* dimensions (while perhaps also having a decently general understanding of the world). For example, an AI system might just outspeed, outscale, or "outscience/outengineer" us – say, suddenly rapidly deploy some organism that kills us all – even if its understanding of the world is generally still inferior in some ways, even if it still cannot develop truly new scientific theories, and even if it cannot even keep itself around for (say) more than a year. It may take such a disastrous action by accident, by malicious instruction, due to some kind of competitive dynamics [5, 10], or for other reasons.

This scenario seems more realistic to me than the "not much further to go" view. One can be concerned about it whether or not one believes that we could develop truly superintelligent AI; we may find ourselves in such a situation before we even get to the point of being able to create superintelligent AI. It is sensible to think about alignment techniques that would prevent bad outcomes from such a situation, even if they are not appropriate for true superintelligence.

[5] Kasirzadeh's contribution to this volume is also instructive here.

Shallow imitation. Perhaps AI just learns to *imitate* us extremely well, so that we have apparently approximately-human-level AI (with scale and speed advantages). Perhaps it still falls short of being able to do things such as making true conceptual scientific breakthroughs (notwithstanding existing contributions such as AlphaFold [17]). Perhaps its understanding is actually very shallow, occasionally doing the sort of thing that today's LLMs and related systems do, making an odd mistake that suggests it did not really understand what it was talking about at all. Perhaps we are quite confident that it is not conscious, and we think of it as a "stochastic parrot" of sorts [1].[6] But yet, it has become so good at imitating us that it is quite effective in the world and in fact poses an existential threat to us. This could be along the lines of the "special abilities" scenario above, where it outmaneuvers us purely based on scale or speed; or it could pose an existential threat to us more gradually, for example making something essential in our social fabric fall apart by its presence [18].

It is not entirely clear how separate "shallow imitation" is from "special abilities"; perhaps they form a spectrum of danger, from danger due to superior narrow abilities to danger due to similar broad abilities. Many risks could be posed by AI that has elements of both.

AI that can build more powerful AI. Perhaps we develop AI that is itself not superintelligent, but it *is* capable of building vastly more powerful AI that *would* become superintelligent in the sense of this chapter. Perhaps we want to prevent it from doing so, or at least get it to the point where it can make responsible decisions about this.

In some sense, "AI that can build more powerful AI" falls under AI that can "outscience" or "outengineer" us, but perhaps it is sensible to treat this case separately. If the goal is merely to prevent it from building more powerful AI, then perhaps the main concern is to make sure that it will not be jailbroken into doing so, or do so accidentally. If the goal is more generally to make responsible decisions about developing superintelligent AI, then presumably we are hoping for the AI to help us solve the problem of aligning superintelligent AI. Likely, though, this AI will be similarly in the dark as we are in terms of being able to assess what things look like to a superintelligent AI, so there is still plenty to be concerned about. That said, we might benefit from some help. Perhaps we could also get such help from AI that is not able to create more powerful AI, though.

Lack of access to *what it is like* to be human. Another possibility is that we do create AI that is intellectually vastly superior to us across the board, *except* this AI fundamentally lacks access to the first-personal experience of *what it is like* to be a human being (or perhaps even more broadly what it is like to be a conscious being or have a first-personal perspective at all). So, it has a broad and deep understanding of the physical world, including both commonsensical and non-commonsensical reasoning, and presumably also of neuroscience – and yet, it does not know what it is like, from a first-personal perspective, to be human, perhaps along the lines of how even the most capable color scientist, if she has been confined to a black-and-white environment her entire life, does not know what it is like to see red [14, 15]. And this one fundamental shortcoming is what our alignment work must be oriented around.

This scenario is somewhat in line with Memoir 7, except that perhaps there is something that the Council can do in terms of continuing to get relevant information from the ants. This, of course, could not consist in just asking them how they feel about being put on fast-

[6] It is worthwhile to compare this scenario, as well as the "lack of access to *what it is like* to be human" scenario below, to the ideas on artificial agency expressed in Floridi's contribution to this volume.

moving spaceships; but perhaps there is some other way that they can continue to provide the relevant first-personal information. Or, perhaps, there is enough relevant insight into that first-personal information that can be given up front.

In this scenario, the ant-aligned humans analogy becomes especially strained, because as human beings we do have a first-personal experience, which perhaps makes it a bit more difficult to imagine why the ant-aligned humans would be incapable of understanding the first-personal experience of being an ant. To be clear, of course it is difficult for us to imagine what it is like to be an ant – that is well illustrated by Nagel's "What is it like to be a bat?" [20]. But conceivably, AI is developed in a way that it does not have access to a first-personal perspective *at all*, possibly making the job of acting as we would like it to significantly more difficult.

Perhaps there are other reasons yet to pursue the various alignment directions that I have argued do not convincingly address the case of truly superintelligent AI. One aspect of researchers doing so is surely "looking where the light is" – at least these alignment directions are concrete enough to do real scientific work on them, and through that we may yet learn something about aligning superintelligent AI. Another take might be that some of the scenarios sketched in the memoirs are actually not that bad. For example, they would not allow the ants (and by analogy, us) to take advantage of all the opportunities that the universe affords, but maybe we can settle for that. But if so, those limitations should be acknowledged.

4 Conclusion

The point of the "memoirs" in this chapter is to establish that many of the ways we think about aligning AI do not seem up to the task if we really take them to be about aligning superintelligent AI – by which I mean the kind of AI that is to us in cognitive abilities (roughly) as we are to ants. This does not mean that they are without value; the scenarios for which those approaches are really intended may be different and important in their own right. And I most certainly do *not* mean to suggest that thinking about aligning extremely capable AI systems is hopeless. But I believe that it is important to be clear about what scenarios we really have in mind. In general, it seems that the more conceptually straightforward approaches to AI alignment, the ones that suggest a clear path forward in terms of engineering systems and that allow running informative experiments today, are the ones that are relatively less likely to apply to (or apply *well* to) superintelligent AI. Again, this is not to deny their importance for other scenarios, including ones that likely pose existential risks of their own, before any superintelligent AI can come about; but perhaps freeing these approaches from the expectation that they address truly superintelligent AI will actually allow them to proceed more effectively. In contrast, to me it seems that really addressing the alignment of superintelligent AI requires grappling with some of the hardest problems in philosophy. For example, one key question seems to be whether *moral realism* is true – are there simply facts of the matter whether something is right or wrong? If we believe in moral realism, then we may feel more comfortable leaving ourselves in the proverbial hands of superintelligent AI: it is likely (or at least more likely than we are) to figure out on its own what the right thing to do is, since there is a fact of the matter about that. And then perhaps it will also be inclined to do the right thing, either because it will naturally be inclined to do so, or the alignment that we do will push it sufficiently in that

direction for the AI to fill in the blanks.[7] On the other hand, if we believe moral realism is false, then we may not feel so comfortable about this; if there is no true notion of right and wrong to guide the AI, it will likely be harder to avoid unintended bad consequences from alignment efforts (such as Goodhart's Law phenomena).

The question whether moral realism is true, and the fact that this is important for how we think about the alignment of superintelligent AI, also illustrates the following broader point. One may argue that how to align superintelligent AI is a *philosophy-complete* problem. What does this mean? The use of "-complete" here is borrowed from the theory of computational complexity, where it has a formal mathematical meaning. A computational problem is *complete* for a class of computational problems if it lies in that class, and moreover, a solution to that problem would provide a solution to every other problem in that class. That is, the problem captures the complexity of the class of problems as a whole, and is thereby the hardest problem in the class (generally together with other, equally-hard problems). AI researchers sometimes informally talk about "AI-complete" problems – problems in AI that, if you could solve that problem, then you could probably solve all problems in AI. (For example, one might think that passing a sufficiently rigorous[8] Turing Test is AI-complete.) Thus, saying that aligning superintelligent AI is "philosophy-complete" would mean that to do it really well, one would need to be able to resolve all[9] problems in philosophy. This may seem a bold claim, but besides the question of whether moral realism is true, we have also already seen, for example, the relevance of whether first-personal experience is, in the relevant sense, accessible from the outside.[10] It is not hard to tie other problems in philosophy to the alignment problem.

Should this leave us pessimistic about aligning superintelligent AI well? That is a natural conclusion. Indeed, in computer science, once researchers discover that a problem is NP-complete, they usually give up on finding a polynomial-time algorithm for that problem at that point. (Finding a polynomial-time algorithm is the relevant notion of "solving" the problem in this case.) However, one can also take a different lesson from the analogy: computer scientists also realize that NP-complete problems are in fact important in practice and so try to develop methods to yet solve them reasonably well, in not all too much time, etc. Perhaps we can still find reasonably good solutions, perhaps by being helped by advanced-but-not-yet-superhuman AI, perhaps – if we can avoid a race to superintelligent AI – over a long period of reflection [22], etc. Unless one believes that there is no chance of developing superintelligent AI, the importance of the problem is certainly high enough that we should not at all give up. But it certainly seems that we should grapple with the philosophical problems.

In any case, the main message of this chapter is that we should be clear about the scale of the challenge of aligning AI that is to us as we are to ants, as well as about which scenarios we actually have in mind as we try to address the challenge. Aligning today's large language models through methods such as reinforcement learning from human

[7] Of course, this also assumes that *we* care to do what is simply right, rather than what is in our own selfish interest. If what turns out to be the right thing to do for the superintelligent AI – in the absolute sense of moral realism – is, say, to gradually let humanity go extinct and fill the universe with a society of lizards of a particular kind, are we willing for that to happen? Or can we be sure that this is not the case?

[8] To the extent that the Turing Test has already been passed [16], it is not a sufficiently rigorous version of the test to be AI-complete, insofar as we still do not have the techniques needed to solve every problem in AI.

[9] Perhaps with a restriction that they need to be sensible, well-posed, etc.

[10] Questions about the nature of first-personal experience, in turn, tie to many other problems in philosophy, including in philosophy of mind, epistemology (including problems in formal epistemology such as the Sleeping Beauty problem [7]), and even metaphysics [11, 12, 13, 19, 3, 4].

feedback (RLHF) is genuinely important for a number of reasons, but, by itself, does not necessarily bestow the right to pat oneself on the back for effectively addressing the problem of aligning superintelligent AI.

Acknowledgments

I thank Vojta Kovařík and Sven Nyholm for detailed comments on an earlier version. I thank the Cooperative AI Foundation, Macroscopic Ventures, and Jaan Tallinn's donor-advised fund at Founders Pledge for financial support.

References

[1] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021. https://doi.org/10.1145/3442188.3445922.

[2] Yoshua Bengio, Michael Cohen, Damiano Fornasiere, Joumana Ghosn, Pietro Greiner, Matt MacDermott, Sören Mindermann, Adam Oberman, Jesse Richardson, Oliver Richardson, et al. Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path?, 2025. https://arxiv.org/abs/2502.15657.

[3] Vincent Conitzer. A Puzzle about Further Facts. *Erkenntnis*, 84(3):727–739, 2019. https://doi.org/10.1007/s10670-018-9979-6.

[4] Vincent Conitzer. The Personalized A-Theory of Time and Perspective. *Dialectica*, 74(1):1–29, 2020. https://arxiv.org/abs/2008.13207.

[5] Vincent Conitzer and Caspar Oesterheld. Foundations of Cooperative AI. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*, pages 15359–15367, Washington, DC, USA, 2023. https://doi.org/10.1609/aaai.v37i13.26791.

[6] Tucker Davey. Sam Harris TED Talk: Can We Build AI Without Losing Control Over It?, October 2016. https://futureoflife.org/recent-news/sam-harris-ted-talk/.

[7] Adam Elga. Self-locating belief and the Sleeping Beauty problem. *Analysis*, 60(2):143–147, 2000. https://doi.org/10.1093/analys/60.2.143.

[8] Tim Folger. Why Quantum Mechanics Still Stumps Physicists. *Discover Magazine*, April 2017. https://www.discovermagazine.com/the-sciences/why-quantum-mechanics-still-stumps-physicists.

[9] Dylan Hadfield-Menell, Anca D. Dragan, Pieter Abbeel, and Stuart J. Russell. The Off-Switch Game. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 220–227, Melbourne, Australia, 2017. https://arxiv.org/abs/1611.08219.

[10] Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob N. Foerster, Tomas Gavenciak, The Anh Han, Edward Hughes, Vojtech Kovarík, Jan Kulveit, Joel Z. Leibo, Caspar Oesterheld, Christian Schröder de Witt, Nisarg Shah, Michael P. Wellman, Paolo Bova, Theodor Cimpeanu, Carson Ezell, Quentin Feuillade-Montixi, Matija Franklin, Esben Kran, Igor Krawczuk, Max Lamparth, Niklas Lauffer, Alexander Meinke, Sumeet Motwani, Anka Reuel, Vincent Conitzer, Michael Dennis, Iason Gabriel, Adam Gleave, Gillian K. Hadfield, Nika Haghtalab, Atoosa Kasirzadeh, Sébastien Krier, Kate Larson, Joel Lehman, David C. Parkes, Georgios Piliouras, and Iyad Rahwan. Multi-Agent Risks from Advanced AI, 2025. https://arxiv.org/abs/2502.14143.

[11] Caspar Hare. Self-Bias, Time-Bias, and the Metaphysics of Self and Time. *The Journal of Philosophy*, 104(7):350–373, July 2007. https://doi.org/10.5840/jphil2007104717.

[12] Caspar Hare. *On Myself, And Other, Less Important Subjects.* Princeton University Press, September 2009. https://press.princeton.edu/books/hardcover/9780691135311/on-myself-and-other-less-important-subjects.

[13] Benj Hellie. Against Egalitarianism. *Analysis*, 73(2):304–320, 2013. https://www.jstor.org/stable/24671105.

[14] Frank Jackson. Epiphenomenal Qualia. *The Philosophical Quarterly*, 32(127):127–136, April 1982. https://doi.org/10.2307/2960077.

[15] Frank Jackson. What Mary Didn't Know. *The Journal of Philosophy*, 83(5):291–295, May 1986. https://doi.org/10.2307/2026143.

[16] Cameron R. Jones and Benjamin K. Bergen. Large Language Models Pass the Turing Test, 2025. https://arxiv.org/abs/2503.23674.

[17] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, July 2021. https://doi.org/10.1038/s41586-021-03819-2.

[18] Atoosa Kasirzadeh. Two Types of AI Existential Risks: Decisive and Accumulative. *Philosophical Studies*, 2025. https://doi.org/10.1007/s11098-025-02301-3.

[19] Giovanni Merlo. Subjectivism and the Mental. *Dialectica*, 70(3):311–342, 2016. https://www.jstor.org/stable/26172527.

[20] Thomas Nagel. What Is It Like to Be a Bat? *The Philosophical Review*, 83(4):435–450, October 1974. https://doi.org/10.2307/2183914.

[21] Sven Nyholm. Artificial Intelligence and Human Enhancement: Can AI Technologies Make Us More (Artificially) Intelligent? *Cambridge Quarterly of Healthcare Ethics*, 33(1):76–88, 2024. https://doi.org/10.1017/S0963180123000464.

[22] Toby Ord. *The Precipice: Existential Risk and the Future of Humanity*. Hachette Books, March 2020. https://www.bloomsbury.com/uk/precipice-9781526600233/.

[23] Stuart Russell. *Human Compatible: AI and the Problem of Control*. Penguin, April 2020. https://www.penguin.co.uk/books/307948/human-compatible-by-russell-stuart/9780141987507.

[24] Walter Sinnott-Armstrong and Vincent Conitzer. How Much Moral Status Could Artificial Intelligence Ever Achieve? In S. Clarke, H. Zohny, and J. Savulescu, editors, *Rethinking Moral Status*, chapter 16, pages 269–289. Oxford University Press, 2021. https://philpapers.org/rec/SINHMM.

[25] John Wentworth. The Pointers Problem: Human Values Are A Function Of Humans' Latent Variables, 2020. https://www.alignmentforum.org/posts/gQY6LrTWJNkTv8YJR/thepointers-problem-human-values-are-a-function-of-humans.

[26] Webb Wright. Forget AGI - Meta is going after 'superintelligence' now. *ZDNET*, June 2025. https://www.zdnet.com/article/forget-agi-meta-is-going-after-superintelligence-now/.