Reviews for the Alignment Forum

Suggestions of posts on the AF to review

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

How does one write a good and useful review of a technical post on the Alignment Forum?

I don’t know. Like many people, I tend to comment and give feedback on posts closely related to my own research, or to write down my own ideas when reading the paper. Yet this is quite different from the quality peer-review that you can get (if you’re lucky) in more established fields. And from experience, such quality reviews can improve the research dramatically, give some prestige to it, and help people navigate the field.

In an attempt to understand what makes a good review for the Alignment Forum, Joe Collman, Jérémy Perret (Gyrodiot on LW) and me are launching a project to review many posts in depth. The goal is to actually write reviews of various posts, get feedback on their usefulness from authors and readers alike, and try to extract from them some knowledge about how to go about doing such reviews for the field. We hope to have enough insights to eventually write some guidelines that could be used in an official AF review process.

On that note, despite the support of members of the LW team, this project isn’t official. It’s just the three of us trying out something.

Now, the reason for the existence of this post (and why it is a question) is that we’re looking for posts to review. We already have some in mind, but they are necessarily biased towards what we’re more comfortable about. This is where you come in, to suggest a more varied range of posts.

Anything posted on the AF goes, although we will not take into account things that are clearly not “research outputs” (like transcripts of podcasts or pointers to surveys). This means that posts about specific risks, about timelines, about deconfusion, about alignment schemes, and more, are all welcome.

We would definitely appreciate it if you add a reason to your suggestion, to help us decide whether to include the post on our selection. Here is a (non-exhaustive) list of possible reasons:

- This post is one of the few studying this very important question
- This is my post and I want some feedback
- This post was interesting but I cannot decide what to make of it
- This post is very representative of a way to do AI Alignment research
- This post is very different from most of AI Alignment research
- …

Thanks in advance, and we’re excited about reading your suggestions!

Review of "Fun with +12 OOMs of Compute"

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Introduction

This review is part of a project with Joe Collman and Jérémy Perret to try to get as close as possible to peer review when giving feedback on the Alignment Forum. Our reasons behind this endeavor are detailed in our original post asking for suggestions of works to review; but the gist is that we hope to bring further clarity to the following questions:

- How many low-hanging fruits in terms of feedback can be plucked by getting into a review mindset and seeing the review as part of one’s job?
- Given the disparate state of research in AI Alignment, is it possible for any researcher to give useful feedback on any other research work in the field?
- What sort of reviews are useful for AI Alignment research?

Instead of thinking about these questions in the abstract, we simply make the best review we can, which answers some and gives evidence for others.

In this post, we review Fun with +12 OOMs of Compute by Daniel Kokotajlo. We start by summarizing the work, to ensure that we got it right. Then we review the value of the post in itself -- that it, by admitting its hypotheses. We follow by examining the relevance of the work to the field, which hinges on the hypotheses it uses. The last two sections respectively propose follow-up work that we think would be particularly helpful, and discuss how this work fits into the framing of AI Alignment research proposed here by one of us.

This post was written by Adam; as such, even if both Joe and Jérémy approve of its content, it’s bound to be slightly biased towards Adam’s perspective.

Summary

The post attempts to operationalize debates around timelines for Transformative AI using current ML techniques in two ways: by proposing a quantity of resources (compute, memory, bandwidth, everything used in computing) for which these techniques should create TAI with high probability (the +12 OOMs of the title), and by giving concrete scenarios of how the use of these resources could lead to TAI.

The operational number comes from Ajeya Cotra’s report on TAI timelines, and isn’t really examined or debated in the post. What is expanded upon are the scenarios proposed for leveraging these added resources.

- OmegaStar, a bigger version of AlphaStar trained on every game in the Steam library, as well as on language-related “games” like “predict the next word in a web page" or "make a user engaged as a chatbot". Note that these language-related games are played at the scale of the entire internet.

- Amp(GPT-7), a scaled-up GPT model that is then amplified to be able to decompose any tasks into subtasks it can delegate to copies of itself, recursively.
- Crystal Nights, a simulation of the whole of evolution, but done smartly enough (for example by selecting for abstract intelligence) to reduce the compute needed to fit within the budget of +12 OOMs
- Skunkworks, a STEM AI system that uses incredibly detailed simulations to iterate over millions of design variations and test them, all without having to build the prototypes.
- Neuromorph, a learning process to train brain-like models, starting with running the best brain model available right now at the correct biological scale, and then filling in the blanks and iterating on this process using standard ML techniques (like SGD).

In light of these scenarios, Daniel then argues that if one takes them seriously and considers their results as likely to be TAI, one should put the bulk of its probability mass about when TAI will happen at the point where such increase of compute is reached, or before. In comparison, Ajeya's model from her report puts the median of her distribution at this point, which results in having quite a lot of probability mass after this point.

Daniel thus concludes that an important crux of timeline debates is how people think about scenarios like the ones he presented, and asks for proponents of long timelines to defend their positions along this line for a more productive discussion.

Does the post succeed on its own terms?

This work relies on one big hypothesis: we can get +12 OOMs of compute and other relevant resources in a short enough time frame to warrant the label "short timelines" to the scenario developed here. We found issues with this hypothesis, at least with how it is currently stated and defended. But before detailing those, we start by admitting this assumption and examining how the post fares in that context.

All three of us found Daniel's scenarios worrying, in terms of potential for TAI. We also broadly agree with Daniel's global point that the risk of TAI from these scenarios probably implies a shift in the probability mass that lies after this point to somewhere closer to this point (with the caveat that none of us actually studied Ajeya's report in detail).

Thus the work makes its point successfully: why that amount of additional compute and resources might be enough for TAI, and what it should imply for the most detailed model that we have about TAI timelines.

That being said, we feel that each scenario isn't as fleshed out as it could be for maximum convincingness. They tend to feel like "if you can apply this technique to a bigger model with every dataset in existence, it would become transformative". Although that's an intuition we are sympathetic to, there are many caveats that -- ironically -- the references deal with but Daniel doesn't discuss explicitly or in detail.

Some examples:

- With OmegaStar, one of us thought he remembered that AlphaStar's reward function was hand shaped, and so humans might prove a bottleneck. A bit more research revealed that AlphaStar used imitation learning to learn a reward function from human games -- an approach that solves at least some of the problems with scaling to "all games in the steam" library.
  Since the issue of humans as bottlenecks in training is pretty relevant, it would have been helpful to describe this line of thought in the post.
- With Amp(GPT-7), we wondered why GPT-7 and not GPT-8 or GPT-9. More concretely, why should we expect progress on the tasks that are vital for Daniel's scenario? We don't have convincing arguments (as far as we know) for arguing that GPT-N will be good at a task for which GPT-3 showed no big improvement over the state of the art. So the tasks for which we can expect such a jump are the ones GPT-3 (or previous GPT) made breakthrough at. Daniel actually relies on such tasks, as shown in his reference to this extrapolation post that goes into more detail on this reasoning, and what we can expect from future versions of GPT models. But he fails to make this important matter explicit enough to help us think through the argument and decide whether we're convinced. Instead the only way to find out is either to know already that line of reasoning, or to think very hard about his post and the references in that spec way specifically.

In essence, we find that in this post, almost all the information we would want for thinking about these scenarios exists in the references, but isn't summarized in nearly enough detail in the post itself to make reading self-contained. Of course, we can't ask of Daniel that he explains every little point about his references and assumptions. Yet we still feel like he could probably do a better job, given that he already has all the right pointers (as his references show).

Relevance of the post to the field

The relevance of this work appears to rely mostly on the hypothesis that the +12 OOMs of magnitude of compute and all relevant resources could plausibly be obtained in a short time frame. If not, then the arguments made by Daniel wouldn't have the consequence of making people have shorter timelines.

The first problem we noted was that this hypothesis isn't defended anywhere in the post. Arguments for it are not even summarized. This in turns means that if we read this post by itself, without being fully up to date with its main reference, there is no reason to update towards shorter timelines.

Of course, not having a defense of this position is hardly strong evidence against the hypothesis. Yet we all agreed that it was counterintuitive enough that the burden of proving at least plausibility laid on people defending it.

Another issue with this hypothesis is that it assumes, under the hood, exactly the kind of breakthrough that Daniel is trying so hard to remove from the software side. Our cursory look at Ajeya's report (focused on the speed-up instead of the cost reduction) showed that almost all the hardware improvement forecasted came from breakthrough into currently not working (or not scalable) hardware. Even without mentioning the issue that none of these technologies look like they can provide anywhere near the improvement expected, there is still the fact that getting these orders of magnitude of compute requires many hardware breakthroughs, which contradicts Daniel's stance on not needing new technology or ideas, just scaling.

(Very important note: we haven't studied Ajeya's report in full. It is completely possible that our issues are actually addressed somewhere in it, and that the full-fledged argument for why this increase in compute will be possible looks convincing. Also, she herself writes that at least the hardware forecasting part looks under-informed to her. We're mostly highlighting the same problem as in the previous section -- Daniel not summarizing enough the references that are crucial to his point -- with the difference that this time, when looking quickly at the reference, we failed to find convincing enough arguments).

Lastly, Daniel edited his post to add that the +12 OOMs increase applied to every relevant resource, like memory and bandwidth. But bandwidth for example is known to increase far slower than compute. We understand that this edit was a quick one made to respond to critics that some of his scenarios would require a lot of other resources, but it considerably weakens his claim by making his hypothesis almost impossible to satisfy. That is, even if we could see an argument for that much short term increase in compute, a similar argument for bandwidth looks much less probable.

One counterargument is that the scenarios don't need that much bandwidth, which sounds reasonable. But then what's missing is a ballpark estimate of how much each type of resource is needed, and an argument for why that increase might be done in a short timeline scale.

To summarize, how we interpret this work depends on an hypothesis that is neither obvious nor defended in an easy-to-find argument. As such, we are unable to really judge what should be done following the argument in this post. If the hypothesis is indeed plausible and defended, then we feel that Daniel is making a good point for updating timelines towards shorter ones. If the hypothesis cannot be plausibly defended, then this post might even have the opposite effect: if the most convincing scenarios we have for TAI using modern ML looks like they require an amount of compute we won't get anytime soon, some might update towards longer timelines (at least compared to Daniel's very short ones).

Follow-up work we would be excited about

Most of our issues with this post come from its main premise. As such, we would be particularly excited by any further research arguing for it, be it by extracting the relevant part from sources like Ajeya's report, or by making a whole new argument.

If the argument can only be made for a smaller increase in compute, then looking for scenarios using this much would be the obvious next step.

Less important but still valuable, fleshing out the scenarios and operationalizing them as much as possible (for example with the requirements in the various other resources, or the plausible bottlenecks) would be a good follow-up.

Fitness with framing on AI Alignment research

Finally, how does this work fit in the framing of AI Alignment research proposed by one of us (Adam) here? To refresh memories, this framing splits AI Alignment research into categories around 3 aspects of the field: studying which AIs we're most likely to build, and thus which one we should try to align; studying what well-behaved means for AIs, that is, what we want; and based on at least preliminary answers from the previous two, studying how to solve the problem of making that kind of AIs well-behaved in that way.

Adam finds that Daniel's post fits perfectly in the first category: it argues for the fact that scaled up current AI (à la prosaic AGI) is the kind of AI we should worry about and make well-behaved. And similarly, this post makes no contribution to the other two categories.

On the other hand, the other two reviewers are less convinced that this fully captures Daniel's intentions. For example, they argue that it's not that intuitive that an argument about timelines (when we're going to build AI) fits into the "what kind of AI we're likely to build" part. Or that the point of the post looks more like a warning for people expecting a need for algorithmic improvement than a defense of a specific kind of AI we're likely to build.

What do you think?

1%

2%

3%

4%

5%

6%

7%

8%

9%

10%

11%

12%

13%

14%

15%

16%

17%

18%

19%

20%

21%

22%

23%

24%

25%

26%

27%

28%

29%

30%

31%

32%

33%

34%

35%

36%

37%

38%

39%

40%

41%

42%

43%

44%

45%

46%

47%

48%

49%

50%

51%

52%

53%

54%

55%

56%

57%

58%

59%

60%

61%

62%

63%

64%

65%

66%

67%

68%

69%

70%

71%

72%

73%

74%

75%

76%

77%

78%

79%

80%

81%

82%

83%

84%

85%

86%

87%

88%

89%

90%

91%

92%

93%

94%

95%

96%

97%

98%

99%

Daniel Kokotajlo (90%),adam.shimi (90%)
1%
(Agreement) This work fits neatly within the first category of this framing
(1 = total disagreement, 99 = total agreement)
99%

Conclusion

We find that this post is a well-written argument for short timelines, painting a vivid picture of possible transformative AIs with current ML technology, and operationalizing a crux for the timeline debate. That being said, the post also suffers from never defending his main premise, and not pointing to a source from which it can be extracted without a tremendous investment of work.

In that condition, we can’t be confident about how this work will be relevant to the field. But additional research and argument about that premise would definitely help convince us that this is crucial work for AI Alignment.

Review of "Learning Normativity: A Research Agenda"

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Introduction

We (Adam Shimi, Joe Collman & myself) are trying to emulate peer review feedback for Alignment Forum posts. This is the second review in the series. The first's introduction sums up our motivation and approach rather well, we will not duplicate it here.

Instead, let's dive into today's reviewed work: *Learning Normativity: A Research Agenda* by Abram Demski. We'll follow the same structure as before: summarize the work, locate its hypotheses, and examine its relevance to the field.

This post was written by Jérémy; as such, his perspective will likely bias its content, even if both Adam and Joe approve of it.

Summary

The post describes a conceptual target for AI alignment, *normativity*, that differs in significant ways from other approaches such as value learning and imitation learning.

Pointing at *norms* instead of a specific set of values has several interesting features, especially how it handles the uncertainty of feedback from humans. Norms are reflected imperfectly in human behavior, approval from humans regarding norm-following is sparse and imperfect, yet norms roughly convey what a machine *should* do.

Abram points at language learning as a major motivating example. We don't learn English by knowing all the rules, we don't even know all of them, those we can articulate don't fit the data perfectly. Nevertheless, we can *succeed*, by human standards, at language acquisition.

Going beyond those standards, Abram mentions *superhuman performance in language* based on various properties of texts, with the implicit statement (if we interpret the post correctly) that this performance stems from a better adherence to the underlying norms of language.

This example shows that learning should be possible even in the absence of a *gold standard*, an ideal reference to which the output of an agent may be compared for performance. No feedback can be fully trusted; a system should eventually learn to distrust entire types of feedback, be robust to ontology shifts in feedback, and more generally be able to *reinterpret* any kind of feedback.

This creates a hierarchy in feedback, which is mirrored in the value specification problem: if one cannot specify values directly, one may try to specify how they are learned, and barring that one may try to learn how to learn, etc. Abram suggests there are ways to learn all relevant levels at once, generalizing over all of them. This would represent a possible approach to outer alignment.

Abram then describes *process-level feedback*, where the methods to obtain results are also subject to norms. This type of feedback also suffers from the hierarchy problem outlined previously (i.e. feedback about how to process feedback, piled up recursively). We would like to collapse the levels again, to generalize over all feedback without having an untouchable, possibly malignant, top level.

The suggestion, then, is a "need to specify what it means to learn all the levels" in a given setting, in order to integrate process-level feedback properly. Abram lists three extra obstacles for this approach towards inner alignment.

After a summary of the main requirements that a "learning normativity" agenda would have to address, Abram evaluates the *recursive quantilizers* approach against those requirements. A technical argument follows, concluding that the approach falls short of the reinterpretable feedback requirement, and that process-level feedback may not actually be achieved there.

Do the examples fit the framework?

All three core desiderata for *learning normativity* involve a notion of a hierarchy of levels: object-level feedback, feedback *about* feedback, etc. and the problem of infinite regress. It also posits the existence of a *correct behavior*, which is where the term *normativity* stems from.

We find two issues with the examples used in the post:

- language, taken as a whole, does not have a single normatively correct usage, which makes it awkward to use as an example of single task where a normative ideal *exists in the abstract*;
- with not enough deconfusion around the terms of *process*, *feedback* and *learning* in this framework, the recursive quantilizer approach appears too technical compared to the rest of the post.

Norms and language

Abram's description of *language learning* focuses on initial acquisition, with the rough goal of being able to convey meaning to other people (the primary task of language). In that context, *normativity* is indeed *the result of a complex negotiation between humans*, and fluent speakers will indeed be able to recognize correct usage, insofar they understand each other.

Yet, while *language use* is a good example of a learnable skill with no gold standard, we find that the post could be clearer on whether linguistic norms can be unified in a single objective norm, which we'll argue against.

Rules for correct usage do not exist, not because they're inaccessible or hard to pinpoint, but because looking for them is irrelevant to the task of language acquisition. Attempting to establish rules for language is linguistic prescription: it usually has a social or political aim; rule-following can be seen as a social signal, or as a coordination mechanism.

Each context, each rough task taken in isolation, will lack a commonly accepted ideal, with requirements often at odds with each other. Various journalistic standards require different treatment of facts; fiction norms about suspension of disbelief vary by genre, but also by reader.

It may make sense to talk about *superhuman performance* at syntax and grammar, but this does not extend to all tasks involving language. Creativity, compellingness in writing involve norms that are not seen as extensions of rules of grammar.

If a system is able to follow a variety of human norms in a wide range of contexts (and if we don't move the goalposts endlessly), we might have an objective notion of *superhuman performance*. The point being, it would not correspond to getting closer to any single ideal use of language, but to rank higher than humans at many tasks, each involving distinct sets of norms.

As another framing for this example, this recent DeepMind paper on language agents comes to mind for alignment-relevant uses of language where superhuman performance may be more crisply defined, while preserving the absence of a gold standard, as Abram's framework requires.

Confusion over quantilization

The post does not attempt to give definitions of *feedback*, of *process*, with the phrase "you need to specify what it means to learn in this setting" outlining the operationalization gap to bridge. There is a great deal of deconfusion about learning that needs to happen before piling up layers of feedback, and we find that trying to operationalize them doesn't seem to fit well in the first post.

In Abram's attempt to use recursive quantilizers, he targets a fixed point of a quantilization process, where *learning* is defined as "*an update against a UTAA [which] produces an update against initial distributions which produce that UTAA*". It makes more sense in context, with the little drawings, though the post would benefit from more of them.

This last section on quantilizers has a significantly different tone than the rest of the post. We understand the motive behind displaying a proof-of-concept, but its audience is different, more technical. We expect that splitting the post in two would have been beneficial to readers, with the agenda post simply asserting how the recursive quantilization framework performs on the desiderata, as a commentary on their tractability.

Relevance to the field

How *Learning Normativity* is meant to fit into the field is described right in the introduction. Abram writes that he's pointing at *correct behavior*, in a way that differs from other approaches. Does it? Does this agenda bring a new perspective for alignment research?

First, learning normativity vs. value learning. The concepts being learned are of different natures. A clear distinction between norm-building and value-targeting is explored by Rohin Shah in this post. There, norms are described as expressing what *not* to do, to “read and predict the responses of human normative structure”. The point about norms being the result of complex negotiations, in complex situations, is referenced as well.

Second, learning normativity vs. imitation learning. The post clearly expands the point that direct observation of humans isn’t sufficient to learn norms, no argument on the distinction here.

Third & fourth, learning normativity vs. approval-directed learning vs. rule application. In the same way, this approval (or these rules) can be viewed as a single level of feedback, whereas *normativity* encompasses e.g. feedback about approval and so on.

The post makes a clear argument about how the infinite regresses are unsolved problems. In this review we’ll note that while the post argues that the desiderata point at something useful, it doesn’t argue they’re safety-critical. One could dismiss the desiderata as “nice to have” without further arguments… which Abram has provided in a later post complementing this one.

Where does the post stand?

In Adam’s epistemological framing, Abram’s work fits squarely in the second category: studying what *well-behaved* means for AI, by providing new desiderata for aligned systems. The last quantilizer section might fit in the third category, but it’s not where most of the post’s contribution fits.

All three core desiderata for *learning normativity* involve a notion of multiple levels of feedback, feedback *about* feedback, etc. and the problem of infinite regress.

Several methods are suggested to approach this:

- *to learn a mapping from the feedback humans actually give to what they really mean:* that seems like a restatement of the infinite specification problem (as opposed to a solution), since no feedback can be perfectly trusted;
- *collapse all the levels into one learner:* maybe some insight can be applied to all levels, like Occam’s razor, but then what will give feedback on its application? This also would require that there is a finite amount of useful information that can be extracted from all the levels at once, that going up the meta-ladder doesn’t provide increasingly relevant feedback;
- *one level which is capable of accepting process-level feedback about itself:* that’s one way to stop the infinite regress, which would be highly dependent on its initial prior. There is no guarantee all self-examining processes will converge to the same conclusions about feedback and norms.

None of these methods are satisfying… but they don’t have to be. The document is a research agenda, detailing and motivating a particular array of open problems.

Conclusion

Overall, the post succeeds at outlining the various ways where learning encounters infinite regress, where *normative statements* can be stacked upon themselves, while making a neat distinction between:

- unreliability of feedback at all levels of interpretation;
- unreliability of underlying values at all levels of approximation;
- unreliability of learning processes at all levels of reflection.

We find this post to be a thought-provoking description of a difficult class of open problems related to feedback in learning. Even if “normativity” cannot be reduced to a crisp property of systems, it’s useful to enable a learning process to manipulate the meta-level(s).

The main issues with the post are: the normative sharpness of the main motivating example, which distracts from the otherwise valid concerns about uncertain feedback; the last part about quantilizers that could have belonged in a separate post; conversely, the arguments motivating the importance of the agenda that could have been included here.

We are quite hopeful this avenue of research will lead to interesting results in AI alignment.