

Predicting researcher interest in AI alignment

Vael Gates, 2/1/2022

(This is a copy of the posts on the [EA Forum](#) and [LessWrong](#), but with better formatting.)

TLDR: As someone who talks to researchers about AI alignment, I'm curious if there are ways to predict how well a conversation might go. For example, does demographic information help predict whether a conversation will have lasting effects months later? To answer this, I sifted through two sets of results, focusing on the newly-released quantitative analysis of 97 AI researcher interviews. It was a messy process and I've given up on releasing something that shows my work. However, jump down to the [Overall Takeaways](#) section to read what updates I'm making about predicting research interest in AI alignment.

Introduction

In February-March 2022, I conducted 40-60 min [interviews](#) with 97 AI researchers about their perceptions of AI and the future of the field. The core of these conversations was discussing potential risks from advanced AI systems – I presented some arguments for why we might be concerned, then we discussed what they thought and why.

Maheen Shermohammed and I recently released a [report](#) (interactive graph [version](#)) analyzing these interviews. It's an enormous report, and the main findings are described in [this summary post](#).

However, the summary post doesn't discuss the interaction effects between all of the variables we investigated. By understanding these interactions, we can answer a question I'm very interested in:

If you talk to an AI researcher, and you know something about them – demographics from their website, or information about their knowledge of AI safety, sympathy towards AI safety, or their timelines to AGI – can you predict anything about their future sympathy or whether the conversation will be useful to them?

All the relevant information is in our [report](#). However, it is admirably massive! Ideally there would be a separate document just on the important interaction effects. To that end, I've spent more than 40 hours working through the interaction effect analyses, checking my independent understanding of the graphs against Maheen's careful observations, incorporating some additional results from an [EA Forum post](#) that analyzed [AI Impacts 2022 survey](#) data, lovingly

creating this massive document that lays out all my reasoning as I squint at dozens of graphs and correlations...

...and then I realized how much editing it'd need to post publicly. And you know what, I give up. Message me if you'd like the full, messy, graph-by-graph reasoning doc. Otherwise, here are my takeaways, mostly oriented towards AI safety field-building.

This post is organized as follows:

1. [Variables](#) list, and links to where those variables are defined.
2. [Overall Takeaways](#)
3. [Appendix](#): A simplified version of my full messy doc. For each section below, there's a table showing the data / graphs I was referencing, and then my "interim summary" for that section. If you are not a casual sightseer and are dissatisfied with this light appendix, ask me for the full messy doc where I show my work better.

Acknowledgements: Thanks to Maheen Shermohammed for doing this entire analysis, including writing her interpretations for each graph, which were great to cross-check against. Thanks to Lukas Trötmüller, Michael Keenan, and David Spearman for helping edit the post.

Variables

Variables from [Our Report](#)

Demographic Variables		
h-index	Link to Report	
Age	Link to Report	Based on university graduation date
Field	Link to Report	Field was evaluated in two ways: asking the participant directly in the interview (Field1) and by looking up participants' websites and Google Scholar Interests (Field2). All analyses here use Field2.
Sector	Link to Report	Academia vs. industry
There are additional demographic measures included in this dataset which are included in the " Demographics Correlation Matrix " section. Demographic information was sourced from Google Scholar and personal websites / LinkedIn.		

Measures of “Sympathy to AI Risk Arguments”		
“Alignment + Instrumental”	Link to Report	<p>This was the primary measure of what I’m calling “sympathy”.</p> <p>In these interviews, I described the alignment problem, and asked researchers if this argument seemed valid or invalid.</p> <p>I then described the idea of instrumental incentives, and asked researchers if that argument seemed valid or invalid.</p> <p>“Alignment + Instrumental” is a combined measure:</p> <ul style="list-style-type: none"> - 1 if researchers thought both these arguments seemed valid - 0 otherwise <p>This combined measure is meant to be more robust than agreement to either “Alignment Problem” or “Instrumental Incentives” alone.</p>
Alignment Problem	Link to Report	<p>The core question was:</p> <p>What do you think of the argument: “highly intelligent systems will fail to optimize exactly what their designers intended them to, and this is dangerous?”</p>
Instrumental Incentives	Link to Report	<p>The core question was:</p> <p>What do you think about the argument: “highly intelligent systems will have an incentive to behave in ways to ensure that they are not shut off or</p>

		limited in pursuing their goals, and this is dangerous?”
Work on this [alignment research]		<p>This question was asked in a bunch of different ways, but the core was:</p> <p>Would you work on AI alignment research?</p> <p>It could be considered a sympathy measure, and I included it as such for some sections below. However, I don't think it tracks “sympathy” very well: see the “What is the ‘work on this’ variable tracking?” section for details.</p>
Main Variables		
When will we get AGI?	Link to Report	<p>This was asked in different ways, and advanced AI systems / AGI was imprecisely defined, but roughly:</p> <p>When do you think we'll have very general capable systems, perhaps with the cognitive capabilities to replace all current human jobs (so you could have a CEO AI or a scientist AI), if we do?</p>
Heard of AI safety?	Link to Report	
Heard of AI alignment?	Link to Report	
Work on this [alignment research]	Link to Report	<p>This question was asked in a bunch of different ways, but the core was:</p> <p>Would you work on AI alignment research?</p>

Did you change your mind?	Link to Report	“Have you changed your mind on anything during this interview and how was this interview for you?”
Follow-up Questions: Lasting Effects	Link to Report	I emailed researchers 5-6 months later, and asked this binary question (Y/N): “Did the interview have a lasting effect on your beliefs?”
Follow-up Questions: New Actions	Link to Report	I emailed researchers 5-6 months later, and asked this binary question (Y/N): “Did the interview cause you to take any new actions in your work?”

Variables from [AI Impacts 2022 Survey Analysis](#)

I also analyzed graphs from this [EA Forum post](#) that looked at interaction effects in the [AI Impacts 2022 Expert Survey on Progress in AI](#) data. I did not independently verify the results from that post. Those variables are included in the Appendix.

Overall Takeaways

Neat findings

More sympathy for AI risk arguments, and sooner AGI timelines, among top researchers

- Researchers who thought the AI risk arguments were invalid had [lower h-indices on average](#). Researchers with shorter AGI timelines tended to have [higher h-indices on average](#), though this effect was not strong.
- Being sympathetic to the AI risk arguments is correlated with being in a higher ranked university (Spearman’s rho=-0.37, n=67, p=0.002, no correction for multiple comparisons). Believing AGI will happen is correlated with being in a higher-ranked university (Spearman’s rho=-0.27, n=71, p=0.02, no correction for multiple comparisons). The combined measure, “believing AGI would happen and also being sympathetic to

both AI risk arguments” was thus also correlated with being in a higher-ranked university (Spearman’s $\rho = -0.33$, $n=67$, $p=0.006$, no correction for multiple comparisons).

Timelines are important

- Thinking AGI will happen seems to be approximately a prerequisite to being concerned about AI risk, with earlier AGI timelines corresponding with being more interested in doing alignment research.

When asked about starting alignment research, AI researchers want to know what research directions exist that are close to their current research interests and skillsets

- When asking people if they would work on / what would cause them to work on AI alignment research, they maybe heard the question: “Do I have enough research flexibility to work on an alignment-related project (based on my personal understanding of what that entails— some interviewees interpret this as ethics) related to my research?”
 - Where the three things in that sentence – research flexibility, personal understanding of alignment, and related to my research – are all important.
- This reinforces the idea that AI risk arguments are presented to people, one needs to present concrete alignment projects / problems rather than general philosophical ideas, and these problems must be related to what they’re already doing, if one wants to enable people to add an alignment project to their lives (which is in contrast to something more like a complete research upheaval).
- Interestingly, the three people who were tagged as already working on alignment research were very similar: timelines ≤ 50 years, valid on both sympathy to AI risk argument measures, all in academia.

If you’ve already heard of AI safety / alignment before, you’re (weakly but fairly robustly) more likely to be sympathetic to the AI risk arguments

- It’s hard to tell, but there’s also maybe an effect where if you’re new to AI safety / alignment, the interview is more likely to have a long-term effect (i.e. take a new action at work, or lasting change in beliefs).

These interviews seemed useful

- I was quite surprised by the number of researchers who replied to my follow-up email or reminder emails (82/86 contacted, 95%). Also, a surprisingly high number of people, 51%, said that the interview had a lasting effect on their beliefs. Additionally, 15% said the interview caused them to take a new action(s) at work, though strangely none of those people had said they’d be interested in working on AI alignment research during the interview.
- We also had a question “Have you changed your mind on anything during this interview, and how was this interview for you”, and 24 people said “Yes”. This is neat, especially considering how this variable correlates with the interview having a lasting effect 5-6 months later. (The percentage is 24/58 (41%), but this is inflated due to a [selection](#) effect for this question.)

Field-building specific

The demographics that matter for prediction are probably AI subfield and h-index

- Field (AI subfield) is probably relevant to predicting beliefs pertinent to AI alignment research. The field analyses were chaotic and should not be trusted, but my personal takeaways:
 - “Inference” and “Reinforcement Learning” – I hope to reach out to more researchers in these subfields. Researchers tended to be more sympathetic to AI risk arguments.
 - “Math and Theory”, “NLP”, and “Near-term Safety and Related” – these seem like interesting subfields to reach out to. These subfields had relatively complicated profiles, but seem highly relevant.
 - “Optimization” and “Computer Vision” – I will probably be less interested in these subfields. Researchers tended to be less interested in AI alignment.
- Age and sector (academia vs. industry) don’t matter in my data, though the AI Impacts data suggests [early-career](#) and [industry](#) researchers are more concerned about AI risks.
- As mentioned above, highly-ranked researchers / researchers at highly-ranked universities are more likely to be sympathetic to AI risk arguments.

You can ask people if they changed their mind during the conversation as a litmus test

- If you ask people if they changed their mind during the interview, and they say yes, this means something – it’s highly correlated with them saying the interview had a lasting effect on their beliefs 5-6 months later.
 - This is the second-highest correlation of the main variables (Spearman’s $\rho=0.49$, $n=50$, $p=0.0004$, no correction for multiple comparisons). Moreover, it’s the first interesting correlation, since the highest correlation of the main variables was predetermined (“Heard of AI alignment” \leftrightarrow “Heard of AI safety”).
- Also, a much weaker effect ($\rho=0.24$, $n=57$, $p=0.07$): If people don’t believe AGI will happen, they’re less likely to report changing their mind during the interview. (Though there was a set of interviewees who were tagged with both “AGI won’t happen” and “AGI will happen”, so people’s opinions can vary.)

Besides the “changed mind” question, it’s hard to know who will be affected by the interview

- Knowing someone’s AGI timelines or their sympathy towards AI risk arguments doesn’t predict whether the interview will result in meaningful effects months later. (“Meaningful effect”: a lasting effect on beliefs or taking new action(s) at work, 5-6 months later.)
- You can’t really predict who will change their minds or have a lasting belief change 5-6 months later. (Excepting the correlation mentioned above between “AGI won’t happen” and saying no to “Did you change your mind?”)

Appendix

The remaining sections are excerpts from the full messy doc, which is the source for the takeaways above. I first analyze the demographics variables, then all of the other variables we hypothesized could be predictive. For each section, there's a table showing what graphs I analyzed (follow "Link to Report" to see them), and then an interim conclusion.

Demographics

Field

This report	
Alignment + Instrumental Combined, split by Field2	Link to Report
Work on this, split by Field2	Link to Report
When will we get AGI?, split by Field2	Link to Report
Have you heard of AI safety?, split by Field2	Link to Report
Have you heard of AI alignment?, split by Field2	Link to Report
AI Impacts 2022 , split by "By Specific AI Field"	
Society should invest more / much more in AI safety research	
>=5% chance that HLMI would be extremely bad	
>=5% chance AI leading to bad outcomes	
>=5% chance humans can't control AI leading to bad outcomes	

The data in this section was way more chaotic than any other section, and I don't trust most of it. That said, here are my conclusions about what subfields of AI I'm planning to pay more attention to, which should be taken with a grain of salt. Note that I'm focusing on the subfields where the skillsets are overlapping with existing research directions in alignment.

- "Inference" and "Reinforcement Learning" – I hope to reach out to more researchers in these subfields. Researchers tended to be more sympathetic to AI risk arguments.
- "Math and Theory", "NLP", and "Near-term Safety and Related" – these seem like interesting subfields to reach out to. These subfields had relatively complicated profiles, but seem highly relevant.
- "Optimization" and "Computer Vision" – I will probably be less interested in these subfields. Researchers tended to be less interested in AI alignment.

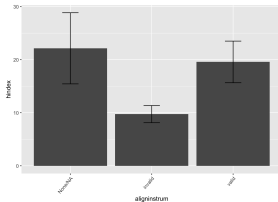
Age

This report	
Alignment + Instrumental Combined, split by Age	Link to Report
Work on this, split by Age	Link to Report
When will we get AGI?, split by Age	Link to Report
AI Impacts 2022 , split by “By Time in Career”	
Society should invest more / much more in AI safety research	
>=5% chance AI leading to bad outcomes	
>=5% chance humans can't control AI leading to bad outcomes	

In my data, age doesn't particularly seem to affect AGI timelines, or sympathy to arguments, though people who think the AI risk arguments are invalid (compared to valid) are maybe slightly younger (a couple of years). In the AI Impacts data, early-career people seem more concerned about risks from AI (directionally this is true across three questions but the strength of that effect differs).

H-Index

(This section includes the two extra columns from the full messy doc. Most of the extra columns aren't as neat as this one.)

Alignment + Instrumental Combined, split by h-index	Link to Report	Invalid tends to have lower h-index	
---	--------------------------------	-------------------------------------	---

Work on this, split by h-index	Link to Report	No differences	
When will we get AGI?, split by h-index	Link to Report	Shorter timelines maybe tend to have higher h-indices, but overlapping error bars	

The mean h-index of researchers who thought the AI risk arguments were invalid was lower than the mean h-index of researchers who thought the AI risk arguments (alignment problem and instrumental incentives arguments) were valid.

Researchers with shorter AGI timelines tended to have higher h-indices on average than researchers with longer AGI timelines, though this effect was not strong.

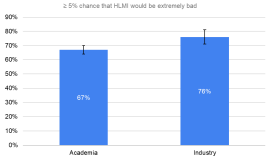

Sector

(This section includes the two extra columns from the full messy doc. Most of the extra columns aren't as neat as this one.)

Note we're ignoring the data from researchers at research institutes (n=3/97), and only comparing academic versus industry researchers.

This report

Alignment + Instrumental Combined, split by Sector	Link to Report	No differences	<p>Note: participants could be tagged in multiple sectors</p> <p>Note: participants could be tagged in multiple categories</p>
Work on this, split by Sector	Link to Report	Academics somewhat more interested than industry	<p>Note: participants could be tagged in multiple sectors</p> <p>Note: participants could be tagged in multiple categories</p>
When will we get AGI?, split by Sector	Link to Report	Relatively similar proportions of academic / industry researchers across all timelines, but industry researchers tend towards shorter timelines (that's an observation made by looking at the count graph rather than the proportion graph, though those counts are very low). "Won't happen" does something separate.	<p>Note: participants could be tagged in multiple sectors</p> <p>Note: participants could be tagged in multiple categories</p>
AI Impacts 2022 , split by "By Industry"			

>=5% chance that HLMI would be extremely bad		Industry is more worried than academia	 <table><caption>≥ 5% chance that HLMI would be extremely bad</caption><tr><th>Group</th><th>Percentage</th></tr><tr><td>Academia</td><td>67%</td></tr><tr><td>Industry</td><td>76%</td></tr></table>	Group	Percentage	Academia	67%	Industry	76%
Group	Percentage								
Academia	67%								
Industry	76%								
>=5% humans can't control AI leading to bad outcomes		Industry is more worried than academia	 <table><caption>≥ 5% chance humans can't control AI leading to bad outcomes</caption><tr><th>Group</th><th>Percentage</th></tr><tr><td>Academia</td><td>65%</td></tr><tr><td>Industry</td><td>81%</td></tr></table>	Group	Percentage	Academia	65%	Industry	81%
Group	Percentage								
Academia	65%								
Industry	81%								

In our report, academics are slightly more interested in working on AI alignment research than industry researchers.

In our report, there are no big differences in sympathy towards AI risk arguments between people in industry and academia. In the AI Impacts data, people in industry are more worried about AI risk than people in academia.

In our report, industry researchers tend towards shorter timelines, but this is a weak effect. (That industry researchers have short timelines and academic researchers have long timelines is the usual claim, which this data doesn't particularly support.)

Demographics Correlation Matrix

Anything we missed that has high correlations in "Demographics x Main Questions"? In particular, we're interested in correlations that we haven't seen yet in the graphs above, because the variables didn't have associated "Split-by" graphs.

Demographics X Main Questions, Using Field2 Labels	Link to Report
--	--------------------------------

My actual summary of the new information we haven't seen above is (ranked by correlation strength, where the first is by far the strongest):

- Being sympathetic to the AI risk arguments is correlated with being in a higher ranked university ($\rho = -0.3736683$, $n=67$, $p=0.0018413$).
- You're more likely to have heard of AI safety and alignment (heardofsafetyandalignment) if you're more senior ($\rho = -0.2523710$, $n=96$, $p=0.0131153$). (Not included in these three points but relevant / supportive / probably driving the effect: heardofAIalignment x professionalrank_ord, $\rho = -0.2697567$, $n=96$, $p=0.0078628$.)
- Believing AGI will happen is correlated with being in a higher ranked university ($\rho = -0.2679489$, $n=71$, $p=0.0238719$). (Note also the not included align_instrum_AGI_allValid x university_ranking_overall: $\rho = -0.3329199$, $n=67$, $p=0.0059091$.)

And there's some other correlations relevant here but they're lower p-values so not including.

Other Predictive Measures

What's the relationship between "AGI timeline estimates" and "sympathy to AI risk arguments"?

Main Questions X Main Questions	Link to Report
Alignment + Instrumental Combined, split by "When will we get AGI?"	Link to Report
Work on this, split by "When will we get AGI?"	Link to Report

Thinking AGI will happen seems to be approximately a prerequisite to being concerned about AI risk, with earlier AGI timelines corresponding with being more interested in doing alignment research.

Let's say we know someone's AGI timelines, or how sympathetic they are towards AI risk arguments. Can we predict whether the interview will have meaningful effects (i.e. lasting effect on beliefs or new actions at work)?

Follow-up Questions: Lasting Effects, split by "When will we get AGI?"	Link to Report
Follow-up Questions: New Actions, split by "When will we get AGI?"	Link to Report
Follow-up Questions: Lasting Effects, split by alignment+instrumental	Link to Report , also Link to Report and Link to Report
Follow-up Questions: New Actions, split by alignment+instrumental	Link to Report , also Link to Report and Link to Report

Short answer: No.

We're asking whether AGI timelines and sympathy towards AI risk arguments are predictive for either of two potential effects:

- Whether the interviewee reports that the interview had a lasting effect on their beliefs 5-6 months later

- Whether the interviewee reports that the interview caused them to take a new action(s) at work 5-6 months later

I don't think the data really makes sense in any direction, and this is a wash – the answer to the question is no.

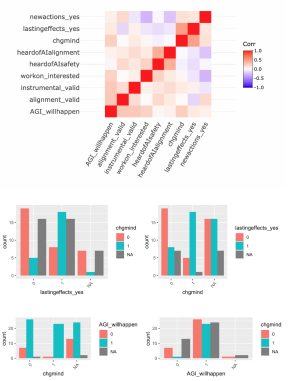
Let's say we know that someone has thought about AI safety or alignment before. Does that predict how sympathetic they'll be towards AI risk arguments? Does that predict whether the interview will have meaningful effects (i.e. lasting effect on beliefs or new actions at work)?

This report	
Main Questions x Main Questions, "Meaningful effects" (The relevant split for "meaningful effects" would be "Lasting Effects" and "New Actions" split by "Heard of AI safety" and "Heard of AI alignment". We don't have those splits, but do have Main Questions x Main Questions correlations.)	Link to Report
Main Questions x Main Questions, "Sympathy"	Link to Report
Alignment + Instrumental Combined, split by "Heard of AI safety"	Link to Report , also Link to Report and Link to Report
Alignment + Instrumental Combined, split by "Heard of AI alignment"	Link to Report , also Link to Report and Link to Report
Work on this, split by "Heard of AI safety"	Link to Report
Work on this, split by "Heard of AI alignment"	Link to Report
AI Impacts 2022	
Split by "By How Much Thought on HLMI" <ul style="list-style-type: none"> • Society should invest more / much more in AI safety research • >=5% chance that HLMI would be extremely bad 	Link to Post
Split by "By How Much Thought on Social Impacts" <ul style="list-style-type: none"> • Society should invest more / much more in AI safety research • >=5% chance that HLMI would be extremely bad 	Link to Post

What I'm actually taking away from this: The general idea that if you've already heard of AI safety / alignment before, you're (weakly but fairly robustly) more likely to be sympathetic to the AI risk arguments (but there's nothing particularly actionable on that). There's also maybe an effect where, if you're new to AI safety / alignment, the interview is more likely to have a long-term effect on you, but it's hard to tell.

If someone reports changing their mind during the interview, does that last?

(This section includes the two extra columns from the full messy doc. Most of the extra columns aren't as neat as this one.)

<p>Main Questions x Main Questions</p>	<p>Link to Report</p> <p>Also for reference:</p> <p>“Did you change your mind?” Link to Report</p> <p>Lasting effects Link to Report</p>	<p>The second most significant correlation in the Main Questions x Main Questions correlation was “Did you change your mind?” and “Lasting effects” ($\rho=0.4851420$, $n=50$, $p=0.0003559$), which is also plotted in a graph. (Note: the highest ranking correlation is uninteresting, since that “Heard of AI alignment” and “Heard of AI safety” would correlate was predetermined.)</p> <p>This means: People who said that they changed their minds during the interview were more likely to report later that the interview had a lasting effect on their beliefs. Similarly, if they said they didn't change their mind, they were less likely to report a lasting effect.</p> <p>“Did you change your mind?” also appears in the 8th-highest rank correlation (higher p-value, not to be taken particularly seriously): Saying one changed one's mind during the interview is correlated with thinking AGI will happen ($\rho=0.2422859$, $n=57$, $p=0.0693929$) – in particular, people who didn't believe AGI would happen basically did not report changing their minds.</p> <p>“Did you change your mind?” also generally follows an expected series of correlations, directionally: positively correlated with “New actions” ($\rho=0.15$) and “Lasting effects” ($\rho=0.49$), very mildly negatively correlated with “Heard of AI safety” ($\rho=-0.06$) and “Heard of AI alignment” ($\rho=-0.02$), positively</p>	<p>Main ?s X Main ?s</p> 
--	--	--	---

		correlated with thinking AGI will happen ($\rho=0.24$), very mildly or mildly positively correlated with thinking the alignment problem ($\rho=0.01$) and instrumental incentive problems ($\rho=0.10$) were valid, and very mildly positively correlated with being interested in working on AI alignment ($\rho=0.04$).	
Follow-up Questions: Lasting effects, split by “Did you change your mind?”	Link to Report	People who said “yes” to whether they changed their minds during the interview were more likely to say that the interview had a lasting change on their beliefs than people who said “no”. People whose responses were tagged as “ambiguous” as to whether they’d changed their minds basically never said the interview had a lasting change on their beliefs. People with “None/NA” responses to whether they changed their minds were split for reporting if the interview had a lasting change on their beliefs or not. (<-- I’m going to ignore most of this except for the main Yes vs No effect though, since that’s the clearer one.)	
Follow-up Questions: New actions, split by “Did you change your mind?”	Link to Report	Interviewees who said they had not changed their mind during the interview never reported that the interview caused them to take a new action(s) in their work.	

This finding is super interesting! It turns out that people saying they changed their mind during the interview (which isn’t even that rare: Yes: 24/58 (41%), Ambiguous: 12/58 (21%), No: 22/58 (38%), though noting there’s [high selection + temporal bias](#) for this question, so the true “Yes” probability is lower) is really very indicative of them saying the interview had a lasting effect on their beliefs 5-6 months later! That’s great, means you can ask that during a conversation and it means something.

- I also like that “Did you change your mind?” just has very reasonable correlations across the board on the Main Questions x Main Questions, which makes me trust it more as a measure.

- More minor evidence for the point that “Did you change your mind” means something: if someone said “no” to changing their mind, they didn’t take any new action(s) at work, though these variables didn’t have a strong correlation: $\rho=0.15$.

There’s also the weaker finding that people who don’t believe AGI will happen tend to not report changing their minds during the interview. This is perhaps expected.

- However, an interesting related finding: For a sense of how often people revised their estimates, search for “an alternative solution” [here](#), which shows how often participants were tagged with multiple timelines estimates during the conversation. Of the people who said at any point that AGI wouldn’t happen, 9/30 (30%) were also tagged with different timeline estimates (4 with “200+y”, 3 with “50-200y”, and 2 with “<50y”) during the conversation. (Checking four of those, the earlier-in-conversation tag was “won’t happen”.) That’s some measure of changing one’s mind about timeline estimates, even if it’s not responding “Yes” to the question “Have you changed your mind on anything during this interview and how was this interview for you?” (Note: this is the same [finding](#) as “there are people who are tagged with both ‘AGI will happen’ and ‘AGI won’t happen’”.)

Can we predict how flexible someone’s beliefs will be?

My original question here was: I’d like to know whether there’s a way to tell how flexible individuals are in their beliefs and if that correlates with anything.

There are two variables directly related to “flexibility of beliefs”: “Did you change your mind?” and “Did the interview have a lasting effect on your beliefs?”

I’m most interested in the correlations here (“Demographics x Main Questions”, and “Main Questions x Main Questions”) rather than any split-by breakdowns (almost all of which we’ve looked at already, if they’re available.)

In particular, I’m looking for any “Demographics x Main Questions” correlations under $p < .05$ that involve “chgmind” or “lastingeffects_yes”. After that, even more liberally, I’m looking for any correlations with $\rho \geq 0.20$. And then I’m doing a similar search within “Main Questions x Main Questions”.

Demographics X Main Questions, Using Field2 Labels, “Did you change your mind?”	Link to Report
Main Questions x Main Questions, “Did you change your mind?”	Link to Report
Demographics X Main Questions, Using Field2 Labels, Lasting effects	Link to Report

Main Questions x Main Questions, Lasting effects	Link to Report
--	--------------------------------

Overall, I'd say the answer is "no, we can't predict individual flexibility in beliefs".

(Outside of what's already been mentioned earlier, about the strong correlation between "changed mind" <> "lasting effects" (which is a bit circular and thus not really relevant here), and weaker correlation ($\rho=.24$) between "changed mind" <> "thinking AGI would happen". Note "lasting effects" <> "thinking AGI would happen" is only $\rho=.10$.)

We've got some weak effects where women are less likely to report having changed their minds during the interview compared to men (there were only 8/97 (8%) women in this series), and some correlations with specific fields (the strongest ones are: NLP is associated with reporting not changing one's mind, and Computing with reporting a lasting effect) that are maybe loosely tied to "previous exposure to AI safety / alignment makes one less likely to report changing one's mind" but that could be spurious.

What is the "work on this" variable tracking?

Something that's come up when I've been plotting graphs for the "work on this" variable, and looking at "workon_interestedOrYes" correlations, is that I'd previously thought this variable was tracking something like "sympathy to the AI risk arguments", and now I don't think it's tracking that. What is it tracking, though?

Let's do some correlations with "work on this". I'm looking for correlations under $p < .05$ that involve workon_interestedOrYes, in "Demographics x Main Questions", or "Main Questions x Main Questions". After that, even more liberally, I'm looking for any correlations with $\rho \geq 0.20$. And there's also some extra split-by graphs for inspection.

Demographics X Main Questions, Using Field2 Labels, workon_interestedOrYes	Link to Report
Qualitative response to Follow-up Question: New Actions	Not available elsewhere – this is everyone who sent me an optional qualitative note about this question
Main Questions x Main Questions, workon_interestedOrYes	Link to Report
Follow-up Questions: Lasting Effects, split by "Work on this"	Link to Report

Follow-up Questions: New Actions, split by “Work on this”	Link to Report
Work on this, split by Alignment Problem	Link to Report
Work on this, split by Instrumental Incentives	Link to Report

- Interestingly, “work on this” doesn’t really seem to be tracking sympathy for AI risk arguments. (Though counterfactually “work on this” probably would have tracked the sympathy to AI risk measures a little better if I’d [asked this question without selection effects](#), and gotten more “no” / “invalid” pairs, for “workonthis” / “alignment”+“instrumental”).)
- When asking people if they would work on AI alignment research / what would cause them to work on alignment research, they maybe heard the question “do I have enough research flexibility to work on an “alignment-related, based on my personal understanding of what that entails (some interviewees interpret this as ethics)” project related to my research?” Where the three things in that sentence – research flexibility, personal understanding of alignment, and related to my research – are all important.
 - But contradiction of this hypothesis is that the people who said they were “interested in long-term safety” (n=13) or “Yes [I’m already working on alignment]” (n=3) did not take new actions at work later, in contrast to people who said “no” to work on this. (Though I guess a new action could be getting a colleague involved, which doesn’t involve one’s research?) So that discounts the variable as a whole to some extent. (But the idea that “interested people maybe don’t take actions” is a little softened by the inclusion of the 3 people who are already working on alignment, who would not be expected to take new actions due to familiarity with the arguments, and none of them in fact do (2 “No”, 1 “None/NA” to New Actions.))
- These replies reinforce the idea that when the AI risk arguments are presented to people, one needs to present concrete alignment projects / problems rather than general philosophical ideas, and these problems should be related to what they’re already doing, if one wants to enable people to add an alignment project to their lives (which is in contrast to something more like a complete research upheaval).
- Interesting that the three people tagged as “Yes” for “work on this” (meaning they’re already working on alignment research) are quite similar: timelines $\leq 50y$, valid on both sympathy to AI risk measures, all in academia.

Aside: How useful were these interviews?

This question is not an interaction effect, but I’m pretty curious anyway. We’ve got a few ways to measure this: whether people changed their mind during the interview, two follow-up questions: lasting effects and new actions, and “work on this”. Just a quick skim through those results:

Did you change your mind?	Link to Report	<p>No: 22/58 (38%). Ambiguous: 12/58 (21%). Yes: 24/58 (41%).</p> <p>Note there's bias for this question.</p> <ul style="list-style-type: none"> - Response Bias: The interviewer tended to avoid asking this question to people who seemed very unlikely to have changed their minds, especially those who seemed frustrated with the interview. - Order effects: The interviewer asked this question explicitly only in later interviews.
Follow-up Questions: Lasting Effects	Link to Report	<p>Responses present for 82/86 (95%) emailed participants.</p> <p>Of the participants, 42 (51%) said yes.</p>
Follow-up Questions: Lasting Effects	Link to Report	<p>Responses present for 82/86 (95%) emailed participants.</p> <p>Of the participants, 12 (15%) said yes.</p>
Work on this (noting that this isn't exactly a sympathy measure)	Link to Report	<p>Yes: 3/97, "Interested in long-term safety but": 13/97, No: 35/97, None/NA: 46/97, but only 55 people were asked.</p> <p>Note there's bias for this question.</p> <ul style="list-style-type: none"> - Response bias: The interviewer tended not to ask this question to people who believed AGI would never happen and/or the alignment/instrumental arguments were invalid, to reduce interviewee frustration. - Order bias: This question tended to be asked in later interviews rather than earlier interviews.

I was quite surprised by the number of researchers who replied to my follow-up email or reminder emails (82/86 contacted). 51% is also a high number for people saying that the interview had a lasting effect on their beliefs. 15% saying the interview caused them to take a new action(s) at work seems interesting (though note none of those people had said they'd be interested in working on AI alignment research during the interview). It's hard to know what "a new action" was, or what the lasting effect on their beliefs was. Reassuringly, the qualitative commentary that some interviewees left with respect to "a lasting effect on beliefs" suggested this was tracking something meaningful to many people, and the people who left "no" comments with respect to "a new action(s)" were mentioning things like projects or decision-making (search "Qualitative response to Follow-up Question: New Actions" to find that data).

It's also hard to know what a "Yes" to "Have you changed your mind on anything in this interview and how was this interview for you" means, and 24/58 (41%) is an inflated percentage because of the [selection](#) effect for this question, but 24 people saying they changed their minds during the interview is pretty cool. Especially neat especially considering how this variable correlates with the interview having a lasting effect 5-6 months later.

"Work on this" doesn't exactly seem to correlate with sympathy to AI risk measures, and I wasn't super convinced anyone in the "Interested in long-term safety but" were going to take actions, plus there's a [selection bias](#) in how the question was asked, so I'm uncertain how to interpret this measure.