

SURVEY REPORT

Survey of AI safety leaders on x-risk, AGI timelines, and resource allocation (Feb 2026)

Ollie Rodriguez and Jemima Jones

I Summary

- We summarise the results of a survey carried out before the 2026 Summit on Existential Security. Respondents were Summit attendees: leaders and key thinkers in the x-risk and AI safety communities.
- The survey asked attendees about their estimates of existential risk, AGI timelines, and resource allocation priorities.
- Survey data comes from the 59 respondents who consented to their answers being shared publicly.¹ Data was collected in February 2026.

X-RISK AND TIMELINES

KEY RESULTS

Probability of human extinction or permanent human disempowerment before 2100	Median: 25% Mean: 34%
50% chance of AGI	Median: 2033 Mean: 2034
25% chance of AGI (n=48)	Median: 2030 Mean: 2030
Assigned $\geq 50\%$ chance of AGI by 2030	22% of respondents
Assigned $\geq 50\%$ chance of AGI by 2035	73% of respondents

We defined AGI as:

An AI system (or collection of systems) that can fully automate the vast majority (>90%) of roles in the 2025 economy. A job is fully automatable when machines could be built to carry out the job better and more cheaply than human workers. Think feasibility, not adoption.

RESOURCE ALLOCATION

- The strongest consensus in resource allocation is that the x-risk community should direct more effort towards **AI-enabled human takeover** scenarios (mean score +0.78 on a -2 to +2 scale, with 43 of 59 saying 'More' or 'Much more') and towards **better futures** work (mean +0.51).
- Respondents lean slightly towards fewer resources on **misaligned AI takeover** (mean -0.14), though this finding is driven almost entirely by those with longer AGI timelines.
- The sub-fields that respondents think deserve significant additional investment are advocacy, policy and governance, and corporate advocacy.

AREAS OF DEBATE AND CONSENSUS AT THE EVENT

- Summit attendees broadly agree that talent is the binding constraint on AI safety, and that risks from aligned AI (such as authoritarian lock-in) deserve far more attention than they currently receive.
- Key debates remain over how well alignment is actually going and whether automated alignment research is a genuine strategy or merely a hope.

Existential risk estimates

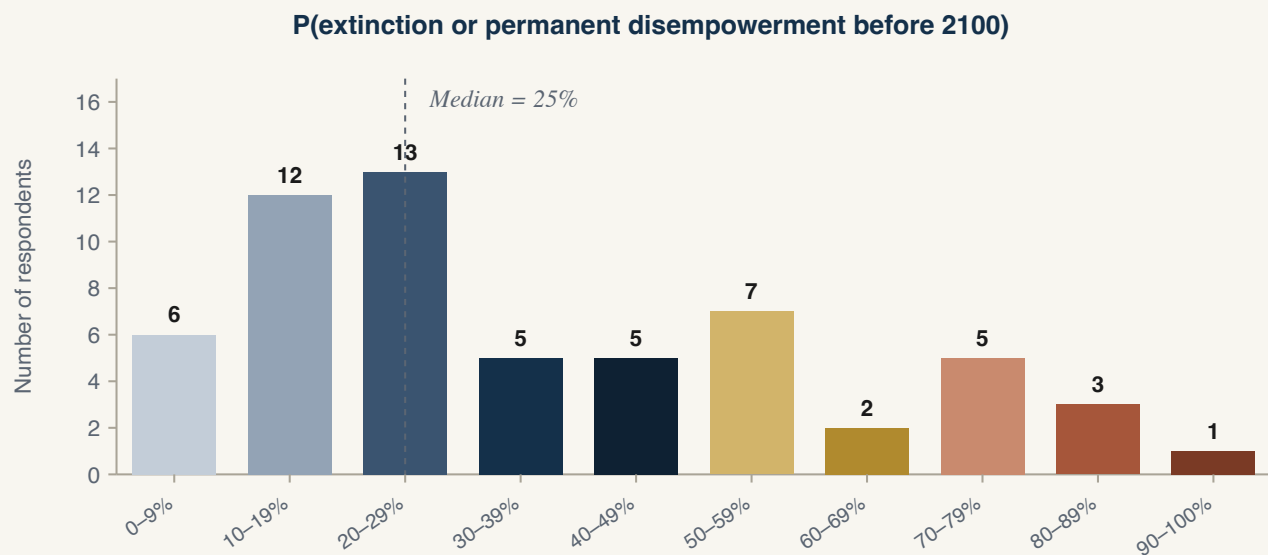
Respondents were asked:

“What is the probability of human extinction or permanent human disempowerment before 2100?”

SUMMARY STATISTICS

MEAN	MEDIAN	STD DEV	IQR	RANGE
34.1%	25.0%	24.2%	14–50%	5–95%

DISTRIBUTION



The distribution is right-skewed with a long tail of high estimates. The modal range is 20–29%, but there is a secondary cluster at 50%+, with 9 respondents (15%) placing their estimate at 70% or above. Several respondents offered important caveats: some distinguished between extinction and disempowerment, noting that permanent disempowerment is not necessarily an existential catastrophe if human values are preserved.

A note on interpretation: these figures reflect the views of a self-selected group of practitioners working on existential risk. The sample therefore likely skews towards people who have already concluded that these risks warrant dedicated attention.

NOTABLE COMMENTS

“I’m not counting it to be human disempowerment if reflected human preferences maintain a major influence on the future, even if humans as they exist today aren’t around.”

“I’d put the chance of human extinction itself below 10%, but I think there’s a significant chance of permanent human disempowerment (which may not in itself mean that the future will be bad/devoid of value).”

“I’m reading human disempowerment as a future where humanity has lost control and few to none of our values are preserved. This would be distinct from a future where we deliberately empower a successor species that retains many of our values, even if AIs and not humans are the ones calling the shots.”

“Competition between companies and countries means we build superintelligence. And I just don’t see how humans retain control of something way more capable and intelligent than us.”

“I think full extinction is fairly hard and most of my probability comes from effects downstream of technological advancement (i.e. nuclear war from civilisational collapse). ‘Permanent human disempowerment’ is hard to define — to what extent are most of us in control of what we spend our time doing even today?”

“If you see progress as fundamentally disempowering, I think extinction is also quite likely. Because ceding control, with whatever alignment methods, will likely not go well. I think we can avoid disempowerment.”

AGI timelines

Respondents were asked:

“In what year do you estimate there's a 50% chance we will have developed AGI?”

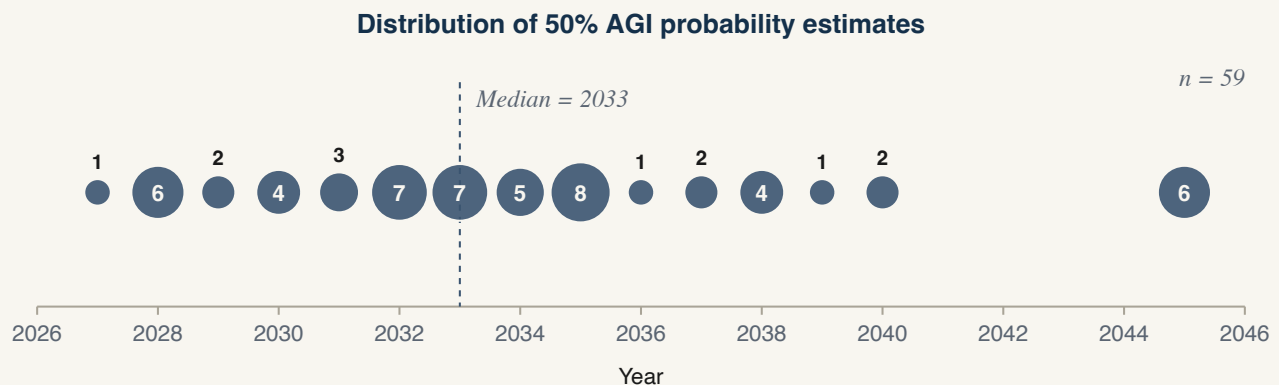
An optional question asked the same at 25% probability. Our definition of AGI was:

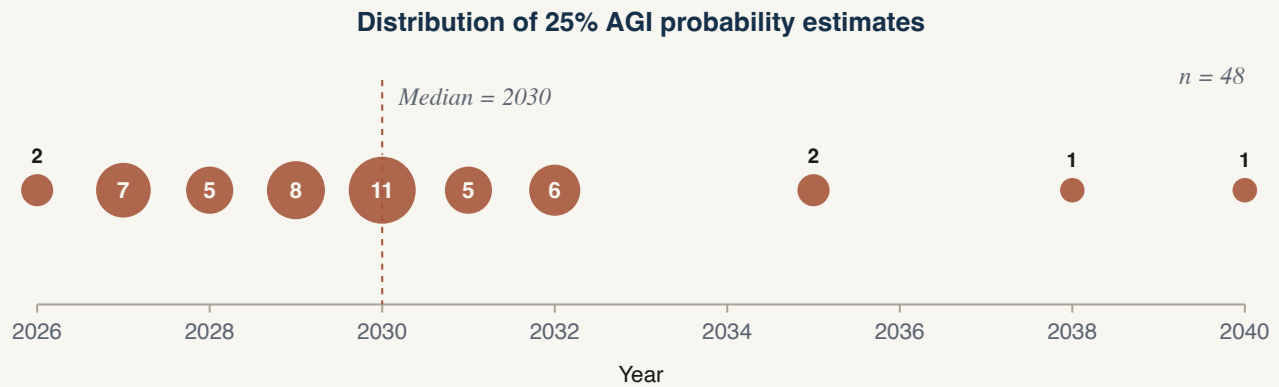
“An AI system (or collection of systems) that can fully automate the vast majority (>90%) of roles in the 2025 economy. A job is fully automatable when machines could be built to carry out the job better and more cheaply than human workers. Think feasibility, not adoption.”

SUMMARY STATISTICS

	50% AGI (N=59)	25% AGI (N=48)
Mean	2034.3	2030.0
Median	2033	2030
IQR	2031–2036	—
Range	2027–2045	2026–2040

DISTRIBUTION OF AGI ESTIMATES





The bulk of 50% AGI estimates cluster in the 2030–2035 range, with 73% of respondents placing their 50% estimate before the end of 2035. The 25% estimates are more tightly concentrated, with an IQR of just 2028–2031, though a handful of outliers extend as far as 2040. Several respondents noted that the definition’s inclusion of physical labour roles pushed their estimates later than they would be for cognitive-only AGI.

NOTABLE COMMENTS

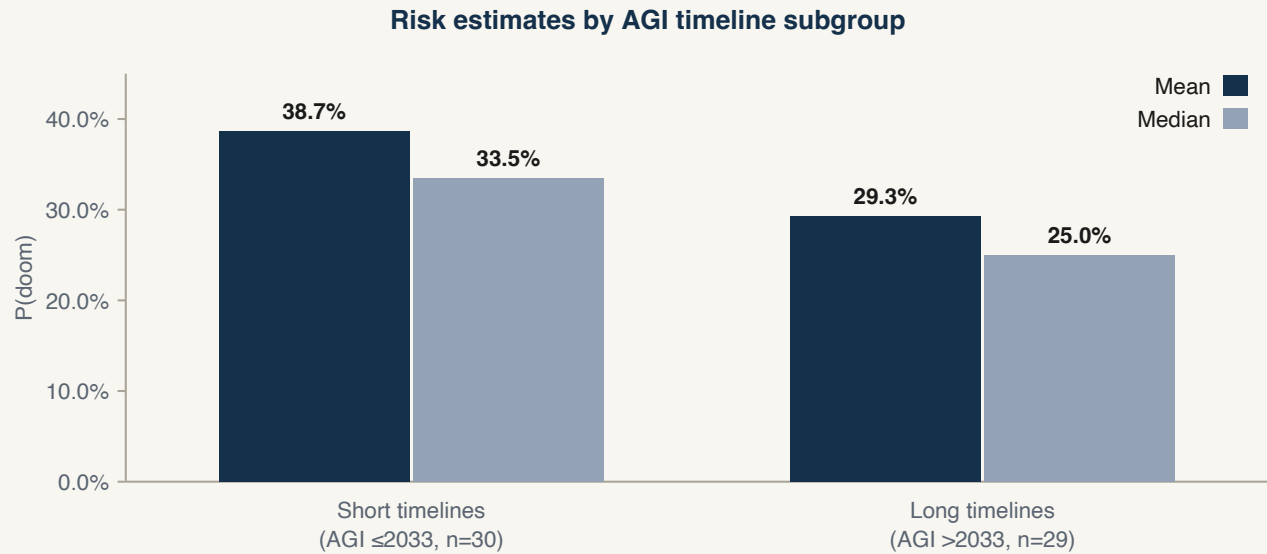
“I think >10% of roles in the 2025 economy are either manual or otherwise require human-like bodies: construction, barbers, restaurant server, etc. If we restrict to knowledge workers (roughly, jobs that can be done on a laptop), these dates move even closer.”

“My numbers would be lower by 1–3 years if the question were just about computer jobs. I expect the development of robots to be the limiting factor for reaching >90%.”

“I understand the question to state that a system exists capable of automating the vast majority of roles, and I feel that many of the arguments for slower timelines (around takeoff, rather than AGI) revolve around a slow pace of diffusion and unhobbling required to actually distribute those capabilities throughout the economy.”

RELATIONSHIP BETWEEN TIMELINES AND RISK

There is a weak negative correlation ($r = -0.25$) between AGI timeline estimates and existential risk estimates: those with shorter timelines tend to assign higher existential risk, though the relationship is noisy.



This subgroup difference is more pronounced when examining resource allocation preferences for misaligned AI, as discussed in the next section.

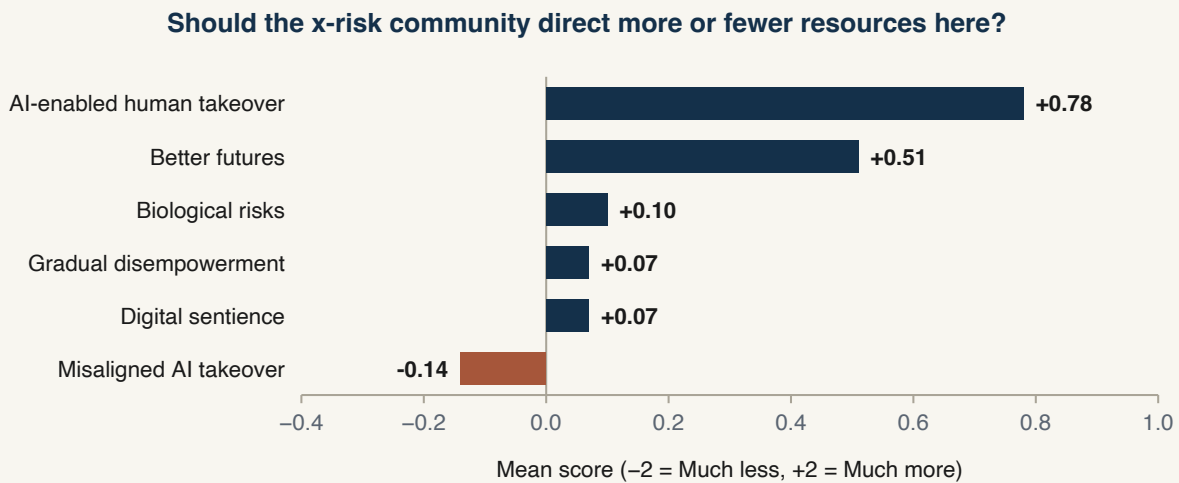
Resource allocation priorities

Respondents were asked:

On the current margin, should the community tackling x-risks direct more or less resources towards addressing the following risk categories and fields?

Below, we represent answers on a five-point scale from 'Much less' (-2) to 'Much more' (+2).

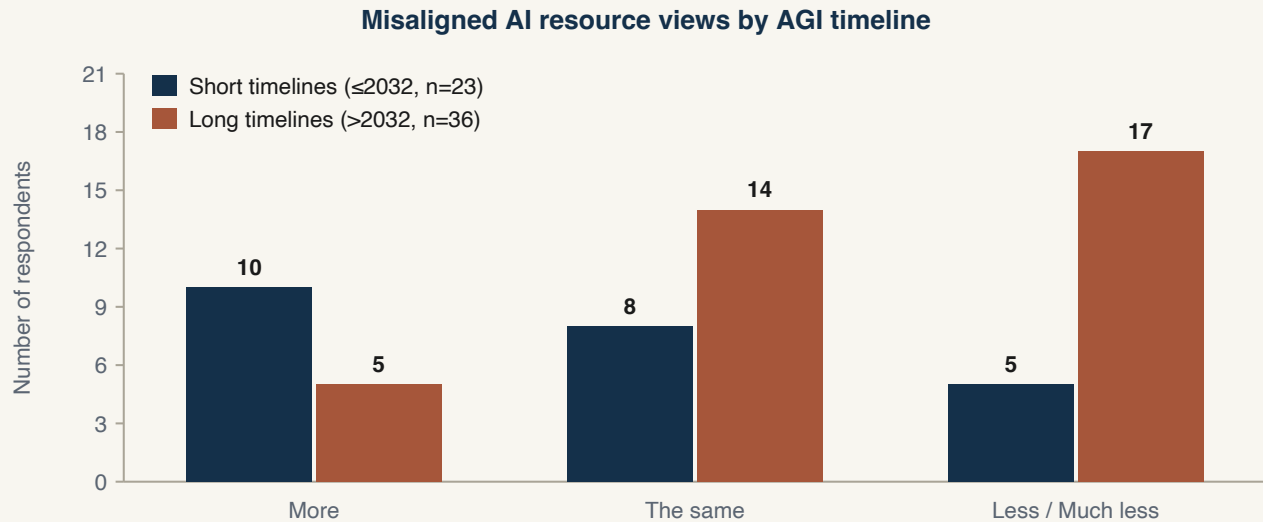
OVERALL PREFERENCES



The clearest signal is the strong preference for more resources on AI-enabled human takeover (e.g. permanent authoritarian state scenarios), with 73% of respondents saying 'More' or 'Much more'. Better futures work also received net-positive support. The slight lean against misaligned AI takeover resources is perhaps the most surprising result for this audience, and merits closer examination.

MISALIGNED AI VIEWS BY TIMELINE SUBGROUP

The aggregate result on misaligned AI masks a significant divergence between subgroups. Among those with shorter AGI timelines (≤ 2032), opinion leans towards more resources; among those with longer timelines (> 2032), it leans substantially towards fewer.



NOTABLE COMMENTS

“I would carve out an AI slowdown/pause/moratorium as a category of its own, and assign it ‘much more’. I completely agree that it’s a heavy lift, but that’s offset by how fragile all the full-speed-ahead alternatives look.”

“AI-enabled bioterrorism is the one most likely to be handled in the near term by natsec, and so less of our resources and attention should be spent on that, relative to the authoritarian state risks which are more [sic] likely to come from natsec.”

“I rated AI-enabled human takeover in the ‘More’ category because I think this is an important threat and potentially more legible to politicians and the public, and digital sentience as ‘More’ because this seems to be the most confused outstanding space.”

“I think that all of these issues are extremely important and deserve much more support than they currently receive, so my answers here focus primarily on which ones seem most neglected.”

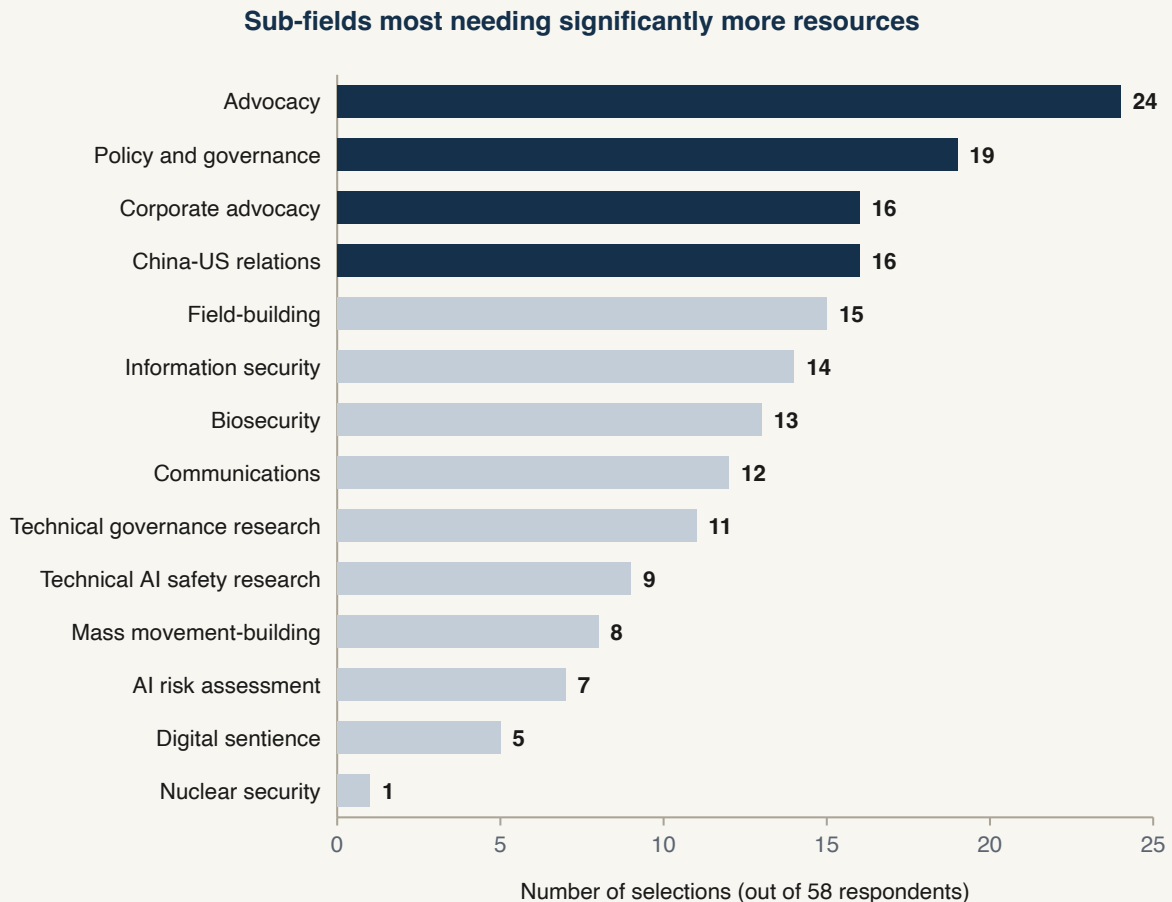
“I see digital sentience / model welfare as a subset of better futures.”

Sub-field priorities

Respondents were asked:

Which of these sub-fields do you think is most important to direct significantly more resources to?

They could select up to three sub-fields. Out of 58 respondents:



The top tier of sub-fields (advocacy, policy and governance, corporate advocacy, and China-US relations) are all oriented towards institutional engagement rather than technical research. Technical AI safety research is selected by only 15% of respondents, ranking 10th of 14 options.

NOTABLE COMMENTS

“My chosen sub-fields fall from my thesis that one of the most important things we can do is lay the groundwork such that a non-existential crisis / ‘shot across the bow’, if it occurs, leads to maximal positive action and societal response. I suspect the following course of events will be highly sensitive to initial conditions, and thus this is the type of space where concerted action from those in this community could do a lot of good.”

“This question is highly contingent on who’s giving the money and how it’s being dispersed! I think China-US relations are highly important. I think field-building targeted at more experienced people who can go on to start and lead competent organisations (and who have networks to convert a load of other experienced people) is much more important than more junior field-building. I think most existing AI safety fellowships act as internship programmes for the labs, which is fine, but we should be doing more than that.”

“Information security, biosecurity, and nuclear security seem like the paths to avoiding takeover risk.”

■ Areas of consensus and debate at the event

This section summarises common themes from attendee-written memos shared at the Summit.

AREAS OF BROAD CONSENSUS

Talent, not funding, is currently the binding constraint on the AI safety field.

Current safety frameworks and evaluation practices are necessary but insufficient. There was widespread agreement that voluntary safety commitments need stronger governance structures, independent oversight, and systematic transparency.

Risks from aligned AI (power concentration, lock-in, authoritarian entrenchment) could be equally important to alignment and deserve much more attention from the AI safety community than they're currently receiving.

Attendee memos proposed crisis action plans, post-AGI scenario exercises, and resilience infrastructure, reflecting a broad consensus that being ready for multiple futures matters more than predicting which one arrives.

AI character and identity are being shaped now, mostly unreflectively. Multiple memos argue this is an under-attended parameter with large downstream consequences.

AREAS OF ACTIVE DEBATE

Timelines and risk levels. Disagreements here are largely captured in the above survey results.

How well is alignment going? Some attendees think that alignment could be solvable and that concerning signals in RL-trained models have been substantially mitigated. Others hold that current models are meaningfully misaligned and that this would be existentially dangerous at greater capability levels.

Is automated alignment research a plan or a prediction? Some memos treat it as a viable strategy; others question whether it contains anything differential that distinguishes it from “build a superintelligence and hope for the best.”

¹ We recognize this represents a form of sampling bias above and beyond those who attended the summit.