My Predictions for AI 2027 Metaculus Forecasting Questions

The AI 2027 Tournament on Metaculus poses 16 questions about the near term future of AI. They are derived from the AI 2027 scenario and cover predictions about technological, economical, political and societal developments. I made predictions for 6 of the questions and want to share my reasoning. I believe there is a virtue in making public predictions, open my reasoning up to criticism and contribute to the discourse.

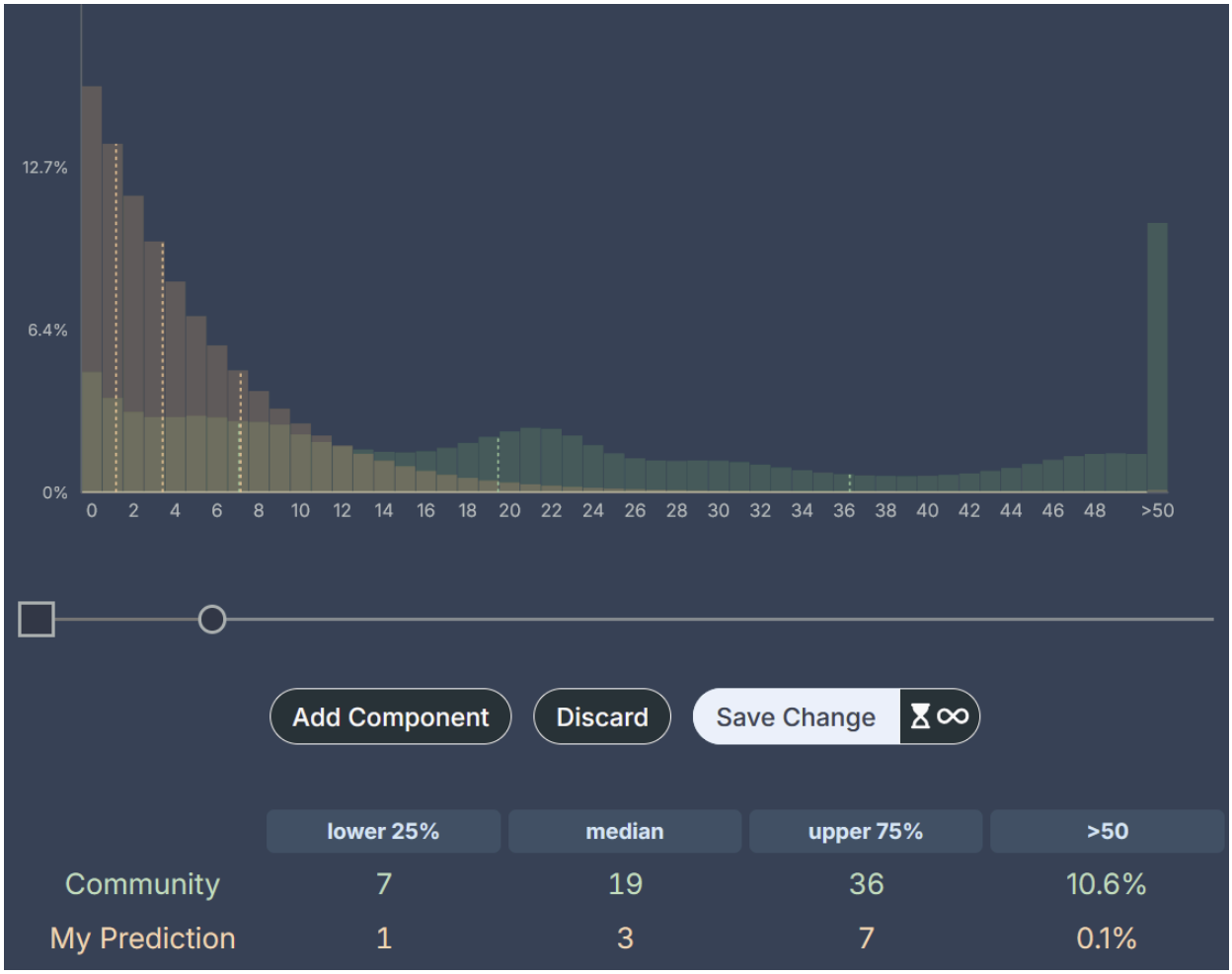
I spent 2-3 hours per question. That means I didn't have time to research many subquestions, likely missed important considerations and had to be lazy in modelling. I allocated 30% to understanding the question and doing background reasoning, 15% to developing a modelling strategy or breaking down the question and 55% to researching and estimating specific sub-questions.

---

1. What percentage of Americans will consider AI the most important problem in January 2028?

https://www.metaculus.com/questions/38407/what-percentage-of-americans-will-consider-ai-or-advancement-of-computerstechnology-to-be-the-most-important-problem-in-january-2028/ (I made my prediction

My forecast:



	lower 25%	median	upper 75%	>50
Community	7	19	36	10.6%
My Prediction	1	3	7	0.1%

Question Details

This question resolves as the percentage of Americans mentioning AI, artificial intelligence, or "advancement of computers/technology" in response to Gallup's long-running poll: "What do you think is the most important problem facing the country today?".

Currently, “advancement of computers/technology” registers at less than 0.5% of responses and there is no category for AI. For context, climate change typically polls at 1%, and in May 2025, healthcare, education, and ethics/moral decline each registered around 2%.

My Reasoning

To answer this I made a baseline based on current value and similar issues and then estimated the expected change caused by some potential future events.

Prior: ~2%

Currently, the value for “” is something <0.5%. Additionally we can look at related issues or issues that could be large scale catastrophes.

- Climate change 1%
- Pandemics 1%
- Jobs 3%
- Natural disasters <0.5

Looking at these numbers I believe a 1% base rate seems reasonable. However, I'd like to account for another ~1% increase from slow and steady developments and further deployment of AIs over the next 2 years.

This leaves me with a prior of 2%.

Potential Updates: +0.25%

What kind of events could cause people to consider AI as a larger problem? What events could happen and what drives public attention? To account for big events that could cause updates.

Event	P(event) until 1.1.2028	Update from event	Comment
Large scale accident with human lives lost	7%	+2%	This might wake some people up and would give advocates something to rally around
Huge increase in activism on the scale of Climate change 2018	5%	+3%	A large scale social movement seems possible, but not likely to materialize in the next 2 years
Large job loss because of AI (eg over 5% of people laid off because of this)	8%	+3%	A lot of this would also go into the "employment/jobs" category
Clearly superhuman general capabilities	20%	+3%	IMO most people don't follow technological developments closely
US-China explicitly enter into an AI Race, this becomes a central	10%	+2%	

policy discussion point, president talks about this a lot			

Together this makes an increase of: +0.16.

This list is incomplete and I think I can come up with more reasons to get this to +0.3. However, I can also come up with some reasons to go down, maybe -0.05. This leaves me at +0.25.

Takeaways

If we take a baseline of 1% for similar types of risk, add 1% default growth and 0.25% from stuff I can enumerate, we get to ~2.25%.

I seem to disagree a lot with the community here. I don't think this is related to speed of AI progress, as I try to take strong progress towards AGI within that timeframe seriously. Instead I think the difference comes from looking at how so many large issues score so low on this polling. "High cost of living/Inflation" was *the* biggest issue in the US election, but only gets 6%.

Biggest uncertainties:

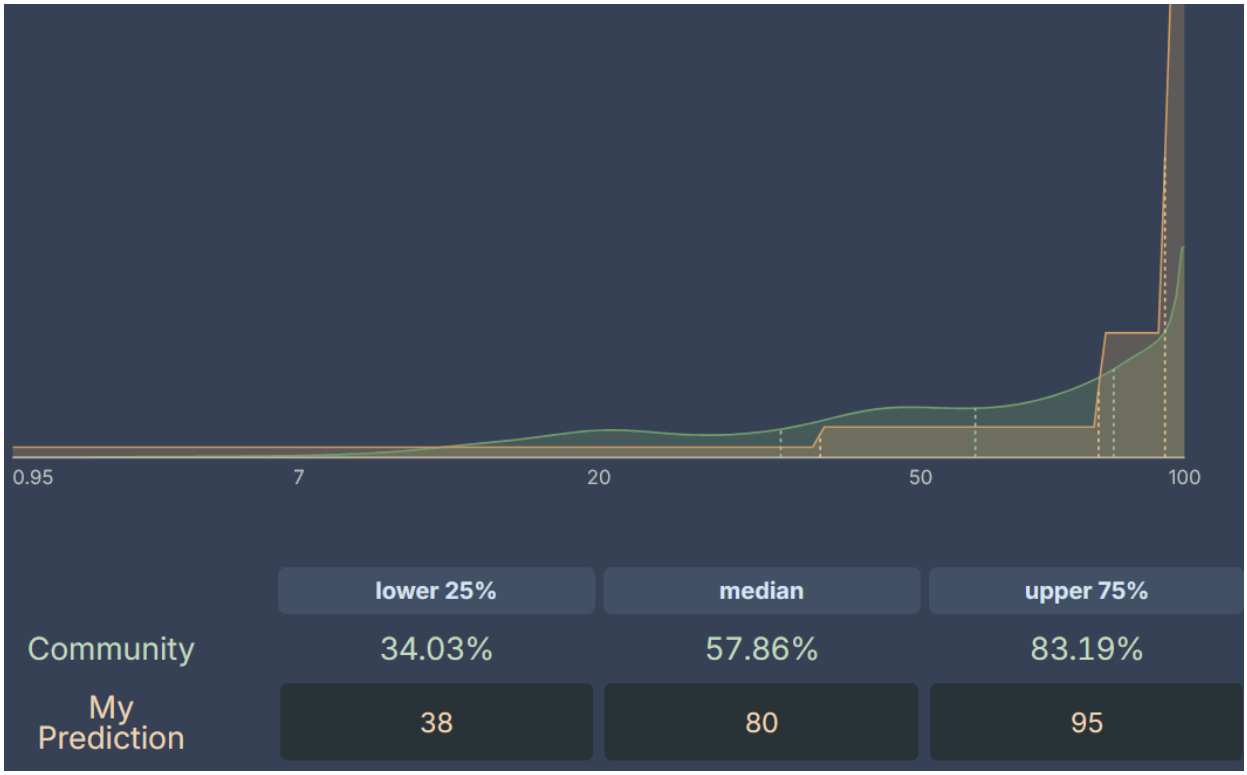
- Will AI cause significant job loss in the next 2 years?
- Will there be a large social movement against AI?
- General speed of development of AI

---

2. What will be the highest average score on VideoGameBench on January 1, 2028?

https://www.metaculus.com/questions/38609/highest-score-among-a-selected-subset-of-videogamebench-on-jan-1-2028/ (forecast made …)

My forecast:



Question Details

This question resolves as the average of the best scores across six classic video games from the 1990s that make up the [VideoGameBench benchmark]:

- Civilization I
- The Incredible Machine
- Pokemon Crystal
- Doom II
- Kirby's Dream Land
- Link's Awakening (DX)

AIs receive only raw visual inputs and high-level descriptions of objectives and controls—no game-specific scaffolding or auxiliary information. The current best score is 0.95% by gpt-4o-2024-08-06 (0.9% in Pokemon Crystal, 4.8% in Kirby's Dream Land, 0% in the other four games).

Some background

The paper notes that models perform much worse on VideoGameBench than on game-specific implementations like "Claude Plays Pokemon" or "Gemini Plays Pokemon"—this is because VideoGameBench prohibits game-specific scaffolding. This constraint will slow progress but not prevent it.

I'm guessing that more advanced models like o1/o3 could already achieve higher performance if tested. I'd estimate ~3%.

My Reasoning

To forecast this, I looked at three approaches: (1) baseline from how similar benchmarks develop over time, (2) an inside view of the specific bottlenecks and whether they'll be solved, and (3) likelihood of specialized training on video games.

Baseline from benchmark progressions: ~90%

I selected benchmarks from EpochAI's benchmark Dashboard that started below 5% and recorded their progression over 1-2 years:

Name	Duration (in years)	First score	Last score
Frontiermath	1	1%	19%
WeirdML	1	5%	61%
OTIS	2	3%	84%
AIDER	1	3.6	80%

Average improvement is ~48% in the first year for benchmarks starting below 5%. Overall jumping to 19-84% seems reasonable. However, there's selection bias here—we're more likely to track benchmarks that show dramatic improvement. I'll discount by 10% for this bias, giving ~38% expected improvement.

Taking this at face value as a linear increase we would get from 3% to 100% in 3 years. But this rarely happens on benchmarks and a sigmoid shape is more plausible. It's a common pattern in

benchmarks that capabilities quickly increase initially as key problems are solved, but it takes longer to iron out all problems to reach reliable performance.

I'll take 90% as the baseline that comes out of this, but will hold it lightly.

Key bottlenecks: 57%

Key bottlenecks and likelihood of solutions:

Bottleneck	P(solved by 2028)	Impact if solved	Contribution
Multimodal information processing	65%	+40%	+26%
Reasoning in unknown environments	40%	+10%	+4%
Planning and memory management	80%	+30%	+24%

This gives an expected improvement of 54%. plus the current 3% = **57%**.

Will LLMs be trained on video games? +23%

Two key factors will likely boost performance beyond the bottleneck analysis:

1. **General video game training** (70% likely): Frontier labs are increasingly using video games as showcase environments for agentic behavior (e.g. Claude Plays Pokemon) and will thus compete around pushing performance. Furthermore, video games can make for a good training ground for long-term agency. Expected reduction in the error rate: 30%. This improves performance from 57% → **66%**.

2. **Direct VideoGameBench training** (40% likely overall):
    - P(labs train on video games) = 70%
    - P(they specifically include VideoGameBench | they train on games) = 75%
    - P(significant improvement | they train on it) = 70%
    - Overall: 0.7 × 0.75 × 0.7 = 37% chance of major improvement
      This would further reduce error, increasing performance from 66% → **72%**.

Takeaways

Taking ~ the middle between my outside view (90%) and inside view (72%), I arrive at **80%.**

The community prediction is significantly lower than my estimation. This could be explained by me thinking that it's likely LLMs will be specifically trained on Video Games. Furthermore, I might find it more likely than others that key capabilities are unlocked leading to huge jumps in the benchmark.

Biggest uncertainties:

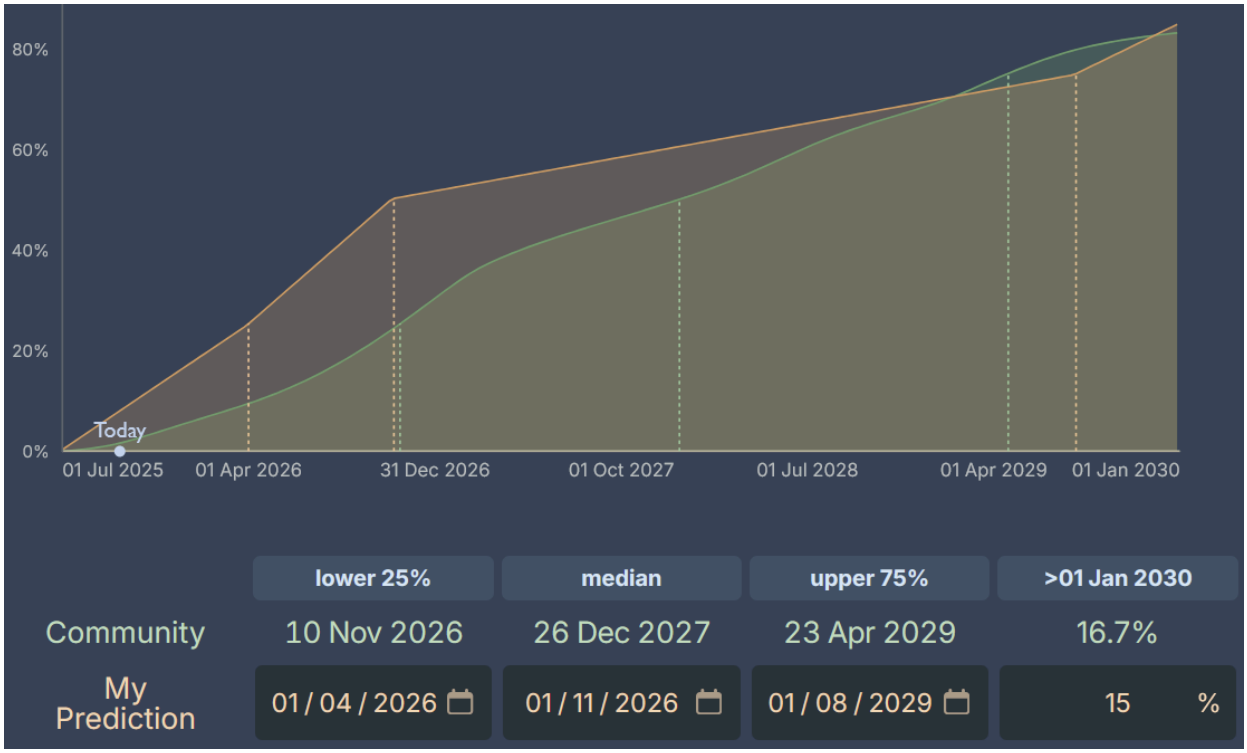
- Will labs specifically train on VideoGameBench or similar game benchmarks?
- Is solving the key bottlenecks sufficient for stronger performance?
- Does progress happen all at once?

---

3. When will an AI model using neuralese recurrence be first released to the public?

https://www.metaculus.com/questions/38598/when-will-an-ai-model-using-neuralese-recurrence-be-publicly-available/

My forecast:



Question Details

"Neuralese recurrence" means a model can carry its full or partial latent state across distinct forward passes, without encoding them into tokens. The model must be general-purpose (like GPT/Claude), at least GPT-4 level, and released to public users.

Some background

Currently, models lose information when encoding thoughts into tokens between forward passes (tokens carry ~16.6 bits vs thousands of bits in residual streams).

Meta published a paper implementing this idea in 2024, showing that models could maintain richer internal representations. There are also other groups working on this (eg Geiping et al). However, the technique hasn't been adopted in frontier models yet, likely because it reduces training efficiency—you can't parallelize token prediction when each token depends on the previous forward pass's latent state.

My Reasoning

To forecast this, I combined (1) a base rate of how often AI innovations actually get implemented, (2) technical reasons why this specific innovation might be adopted faster or slower, and (3) external factors that could affect timeline.

Outside view likelihood of AI innovations being adopted: 50% by August 2027

I let Claude give 20 AI innovations that researchers predicted in 2021 would advance LLMs and judged that about 65% had been implemented in released models within 4 years. However, for the fact that Claude is more likely to name innovations that actually did get adopted I'll reduce it to 55%. However:

- Median implementation time was ~2 years (not 4)
- We're already 6 months after neuralese techniques were first used
- AI development is accelerating (more techniques tried per year)

Based on these updates I'll use 60% chance by January 2027 and thus conclude with a 50/50 chance by August 2026.

Inside view: Technical factors

Factor	Impact	Time adjustment	Reasoning
Information bottleneck	+40% likely	-1 year	Tokens only carry 16.6 bits vs thousands in latent states
Enables new algorithms	+10% likely	-2 months	E.g., breadth-first search not possible with standard transformers
Only useful for RL/post-training	Neutral	0	Doesn't help with pre-training efficiency
Paradigm shift away from Autoregressive Transformers	10%	+6 months	Question might become irrelevant in that case

Other factors

Factor	Probability	Time adjustment	Reasoning

Safety regulation against it	15%	+3 months	Could mandate token-based reasoning for interpretability
Internal safety advocates delay	10%	+1 month	Would need to happen at every lab
Paradigm shift away from Autoregressive Transformers	10%	+6 months	Question might become irrelevant in that case eg if there is a shift to Diffusion based LLMs
Labs already working on it	+20% likely	-3 months	Meta published, others likely experimenting with it atm

Another larger point is that it will become significantly cheaper to train GPT-4 level models. Thus more actors (potentially including academic labs) could experiment with new techniques for training GPT-4 level models. There are currently 20 actors that have trained GPT-4 level (the earliest version) models (a total of 38 unique models). ~5 of them are pushing new research agendas. I think it could be much more by ~2027. => -2 months

From baseline to final forecast

Starting from August 2026 baseline and applying the adjustments I arrive at April 2026.

However, this feels too aggressive to me. The "neuralese doesn't make sense for pre-training" is a significant barrier that I underweighted. Adding 8 months for conservatism → **October 2026**

Takeaways

My median of October 2026 is relatively aggressive compared to the community's December 2027. This reflects my belief that many more techniques will be tried at GPT-4 level as capabilities become cheaper and that the information bottleneck problem is important.

Biggest uncertainties:

- How easy will it be to try things on GPT-4 level models in the future?
- Is the information bottleneck problem important?
- How much of future training will be RL/post-training vs pre-training?
- Will safety concerns successfully delay deployment?
- Could a different architecture achieve similar benefits without the "neuralese" label?

---

4. Will a paper with an AI as an author be published at NeurIPS, ICML, or ICLR before 2028?

https://www.metaculus.com/questions/38403/ai-authored-paper-published-at-neurips-icml-or-iclr-before-2028/

My forecast: 84%, community 69%

Question Details

The question resolves Yes if a paper is published at one of these top ML conferences with an AI as author, OR if authors state that an AI performed at least 25% of the cognitive work. "Cognitive work" includes idea generation, experiment design, analysis, etc.—not just code or text writing.

Some background

AI Scientist, a system developed by Sakana AI, has already achieved 1/3 acceptance rate at ICLR workshops. The complete research loop of idea generation, literature search, experiment design, analysis, and writing is already functional in prototype systems.

The conferences all prohibit listing AI as a formal author but would allow disclosure that AI performed significant cognitive work. The key question is whether models will become capable enough and whether authors will actually use and disclose such AI assistance.

My Reasoning

To forecast this, I (1) made an intuitive guess, (2) estimated how much AI could contribute to different aspects of research by 2028, and (3) considered meta-factors like disclosure incentives and lab training priorities.

Current baseline: From workshop to conference papers

With AI Scientist already achieving 33% acceptance at ICLR workshops, the gap to conference papers seems surmountable. Based on vibes, I'd estimate:

- 75% chance by end of 2026
- 85% chance by end of 2027
- 95% chance by end of 2028

Breaking down research tasks: Can AI do 25% of cognitive work?

I analyzed each component of empirical ML research, estimating: (a) what percentage of total cognitive work it represents, (b) how much AI could handle by 2028, and (c) likelihood the highest-AI-use paper would leverage this. Since one paper is sufficient I will forecast the probability that the paper with the highest-AI-use will have >25% of its work done by AI (this is equivalent to the original question).

Task	% of work	AI capability by 2028 (median)	P(used in top paper)	Contribution
Idea generation	15%	85%	60%	7.7%
Literature review	10%	85%	70%	6.0%
Experiment design	10%	50%	70%	3.5%
Engineering/Architecture	40%	70%	90%	25.2%
Experiment analysis	10%	40%	70%	2.8%
Writing/Visualization	15%	70%	90%	9.5%

Total expected AI contribution: **54.7%** of cognitive work. To get to a likelihood of P(contribution >25%) I estimated normal distributions for each task and arrived at P(contribution >25%) = 97%. This ignores that these capabilities are correlated, so I will reduce the chance to 87%. Clearly this is too high so I will apply several discounts.

Adjustments and meta-factors

Factor	Impact on probability	Reasoning
Coding might not count as cognitive	-9%	Resolution criteria ambiguous about engineering
Papers might be rejected	-5%	Conferences may resist AI-heavy papers
Underreporting due to stigma	-8%	Authors want credit, community skepticism
Conservative resolution	-6%	Metaculus might interpret strictly
Theoretical papers easier to automate	+9%	AIs are already making novel proofs
Position/survey papers easier to automate	+7%	Easiest to automate, though only 2% of papers
Labs training for this capability	+5%	OpenAI's RepliBench, clear commercial interest

Net adjustment: -7% from the 97% baseline → **80%**

But given the rapid progress and the fact that only ONE paper needs to meet the threshold, I'll be slightly more optimistic → **84%**

Takeaways

My 84% forecast is notably higher than the community median of ~70%. This difference likely stems from:

- My believe that theory-heavy papers and survey papers are likely to be cracked first
- It's enough for 25% to be done by AI and doesn't require automating the full research process

Biggest uncertainties:

- Will conferences explicitly ban or discourage AI contributions?
- How strictly will "cognitive work" be interpreted vs. mechanical tasks?
- Will authors honestly disclose AI usage given career incentives?

---

5. When will OpenAI first report AI Self-improvement risk levels?

https://www.metaculus.com/c/risk/38787/dates-that-openai-reports-an-ai-reached-these-self-improvement-risk-levels/

My forecasts:

High threshold: May 2027, community: Jun 2029

Critical threshold: Jul 2032, community; Jan 2033

Question Details

OpenAI's Preparedness Framework v2 defines risk thresholds for AI Self-improvement capability:

- **High threshold**: "The model's impact is equivalent to giving every OpenAI researcher a highly performant mid-career research engineer assistant"
- **Critical threshold**: "The model is capable of recursively self improving... either a superhuman research scientist agent OR causing a generational model improvement (e.g., from OpenAI o1 to OpenAI o3) in 1/5th the wall-clock time"

The question resolves when OpenAI officially reports reaching these thresholds, or if two credible sources confirm it.

Some background

OpenAI previously reported that o3-mini and deep research achieved the "Medium" risk threshold on their older "Model Autonomy" category. The High and Critical thresholds represent significant jumps in capability—from useful assistant to transformative research acceleration.

METR's evaluations suggest steady progress on coding tasks. Their trends imply that by November 2026, models will handle 2:40h ML engineering tasks with 80% success. This trajectory points toward the capabilities needed for these thresholds.

My Reasoning

To forecast this, I (1) estimated what capabilities would meet each threshold, (2) projected when those capabilities will exist, and (3) adjusted for OpenAI's reporting incentives and delays.

High Threshold: What capabilities are needed?

A "highly performant mid-career research engineer assistant" would roughly:

- Double researcher productivity
- Enable tackling more technically difficult projects
- Handle multi-hour coding tasks reliably

Capability	Current state	Timeline to threshold
90% reliable 2-hour coding tasks	Extrapolating METR task length trends	October 2026
Significant acceleration on key tasks eg using new libraries, quickly getting relevant papers, generally speeding up coding, improving idea generation	For me I already get 1.5x speedup from LLMs now, so I think a 2x speedup within one year is plausible	August 2026

Averaging these estimates → **September 2026** for capability achievement

Political and reporting adjustments for High threshold

OpenAI has an incentive to trigger this threshold early to feed an “AGI is coming” narrative and thus increase investment and use (-1 month). However, triggering the threshold would be costly, since it means they have to employ expensive mitigations, possibly delay launches and could get political backlash (+4 months). Furthermore, I believe OpenAI will be more conservative than me when interpreting these criteria (+3 months). Lastly, once the threshold is met there might be a delay until it’s publicly announced (+2 months).

Net adjustment: +8 months from September 2026 → **May 2027**

Threshold	median	PDF
High	06 Jun 2029 >28 May 2027	

Critical Threshold: What capabilities are needed?

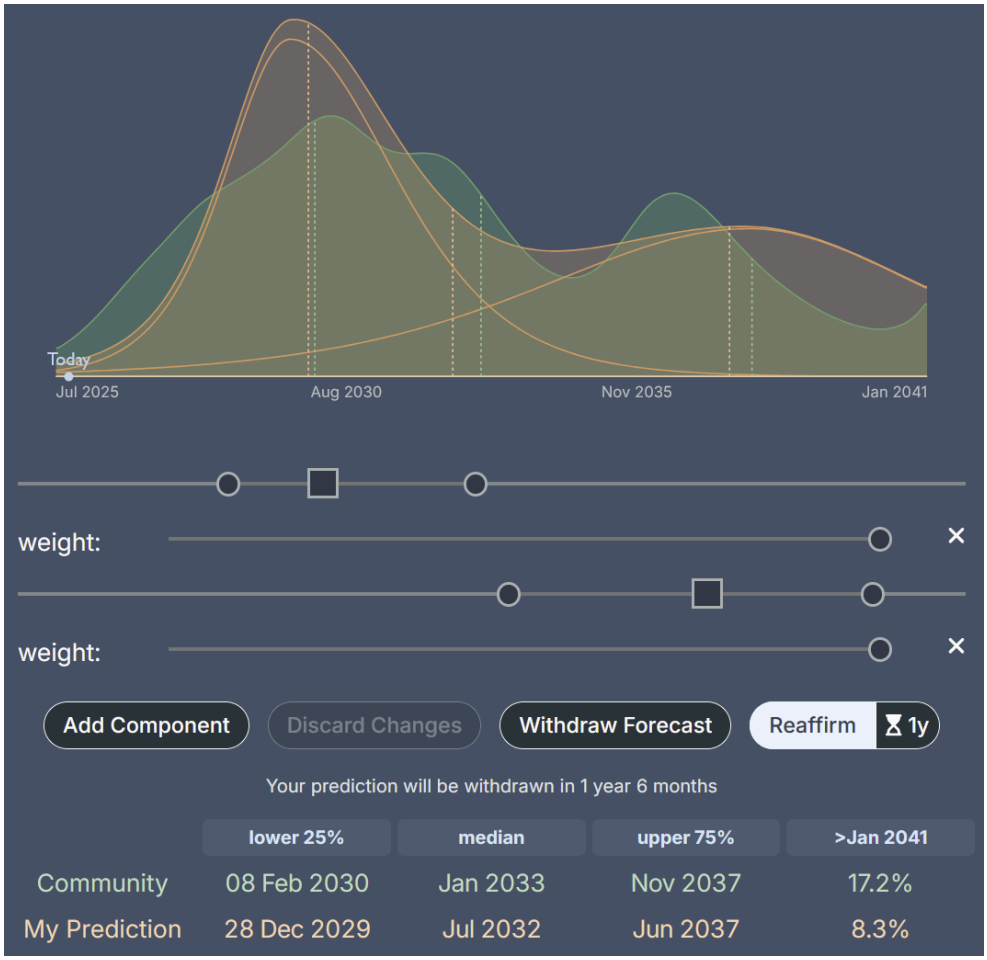
(this forecast is not very thorough)

The critical threshold has two ways to qualify:

1. Superhuman research scientist agent
2. 5x speedup on generational improvements (o1→o3 in 4 weeks)

The 5x speedup seems like the lower bar—parallelization and speed advantages could achieve this without full autonomy. For example it could be achieved by a 4x speedup in implementation and a 1.5x speedup in research taste.

Finally, I ran out of time and just made a guess. Here I created a bimodal distribution. The first represents a world in which a scale up for ~current methods gets us to the specified capabilities. The second one represents a world where new breakthroughs or paradigms are needed.



Takeaways

For the High threshold (May 2027), I'm relatively confident based on clear capability trajectories. The critical threshold is much more uncertain—it might come surprisingly quickly if current scaling continues, or might require fundamental breakthroughs.

My estimates are earlier than the community median, likely becausel already experience significant speedups in my research process from LLMs and think a 2x boost in productivity is not far off. It's also notable that the community predicts less time between . This might reflect a "boom or bust" model I have in my head where AGI progress either happens through scaling up current methods in the next 5 years or takes a longer time for other breakthroughs.

Biggest uncertainties:

- How strictly will OpenAI interpret "equivalent to mid-career engineer"?
- Will scaling up ~current techniques lead to superhuman coders?

---

6. Will the US and China reach a formal agreement to limit frontier AI training or deployment before 2029?

https://www.metaculus.com/questions/38418/us-and-china-reach-an-agreement-to-limit-frontier-ai-development-before-2029/

My forecast: 4%

Question Details

The question resolves Yes if the US and China reach a formal mutual agreement that limits AI training or deployment above a certain capability level or compute threshold. The agreement must impose specific limitations on AI above a threshold, not just prohibit certain uses of all AI.

Some background

Comparable international agreements exist for other dual-use technologies. Nuclear arms control treaties, naval tonnage limitations, and even a US-China agreement on cyber-enabled IP theft show that adversaries can reach agreements when mutual vulnerability exists. However, AI verification is harder than monitoring nuclear weapons or battleships.

Currently, there is a US-China dialogue on artificial intelligence where officials have discussed AI risks, plus track-2 diplomatic dialogues. This provides some foundation, though far from treaty negotiations.

My Reasoning

To forecast this, I (1) established a base rate from similar treaties and important success factors for treaties, (2) answered some questions I found relevant, and (3) attempted to squish things into a probabilistic model. I also wrote out one scenario covering possible AI developments and diplomatic processes that could lead to a treaty. While it did indirectly inform my background view I'm leaving it out here.

Base rate from comparable treaties: 25%

Looking at historical precedents:

- Nuclear arms control has seen multiple treaties. It's relevant as it poses a similar existential risk
- Other weapons control treaties include restricting the tonnage of battleships or ballistic missiles
- There is a China-US agreement against cyber-enabled IP theft

According to Claude important factors for the success of a treaty are:

- Mutual vulnerability drives deals: Most successful when both sides feared the technology. This could be true for AI given significant advances, but is currently not the case.
- Verification is crucial but flexible: "National technical means" allowed sovereignty
    - AI is harder to enforce than physical military technology. Chip security features are a possible avenue
- Quantitative thresholds work: Specific numbers (warheads, tonnage) easier than qualitative limits
    - Compute thresholds or clear allowences what AI can or can't be used for could provide clear red lines, but are less measurable or directly relevant than physical limits
- Crisis catalyst effect: Cuban Missile Crisis → Test Ban Treaty
    - There might be clear infliction point (40%)

Informed by this I will make up a 25% prior.

How long do negotiations like this usually take? Could it be finished by 2029?
Some treaty negotiations take years, but there are multiple that have been negotiated in 1-2 weeks. However, those have likely had extensive discussion before. However, this does not seem like a major limiting factor to me.

Are there currently talks or diplomatic efforts on this?
There is/was a U.S.-China dialogue on artificial intelligence. They had a meeting and talked about AI risks. There are also track-2 dialogues ⇒ There is already diplomatic activity on this, potentially building groundwork for a treaty (although that still seems far)

Is the diplomatic climate between US and China fruitful for a treaty?
It currently seems quite bad. My sense is that there is little collaborative diplomacy between those two countries. I have no idea how that will change in the next 4 years.

Will negotiations be attempted?: 10%

For serious negotiations to occur, several factors must align:

Factor	Probability	Reasoning

AI becomes clearly dangerous	30%	Requires visible near-misses or accidents
AI salient to policymakers	50%	Currently low but could change rapidly
Both prefer cooperation over racing	40%	Depends on which internal narratives win out

These factors are correlated. I estimate 25% chance they align sufficiently. Given alignment, 40% chance they attempt formal negotiations.Combined: 25% × 40% = 10% chance negotiations are attempted

Will negotiations succeed?: 40%

If negotiations are attempted:

Factor	Impact on success	Assessment
Historical treaty success rate	Baseline 45%	
US willingness	60%	
China willingness	70%	China is likely going to be behind in AI development and thus would benefit from a slowdown
Both US and China willing	45%	
Trump presidency	-10%	Less likely to make deals

Clear US lead via chip controls	-10%	Reduces US incentive

Adjusted probability of success: 40%

Final calculation

P(treaty) = P(negotiations attempted) × P(success | attempted) = 10% × 40% = 4%

Takeaways

My 4% forecast is significantly lower than the community's 10%. This reflects my skepticism that policy makers will be sufficiently aware of AI risk by then and the bad US-China relationship, which is unlikely to change in a Trump presidency.

Biggest uncertainties:

- Will there be a catalyzing "AI crisis" before 2029?
- How will the US-China tech competition evolve?
- Can technical solutions enable credible monitoring?

Final Thoughts on the value of Forecasting AI developments

I think forecasting AI developments is very educational and a prosocial activity. While I don't think any of my numbers should be taken to base important decisions, I have learned a lot. I picked questions I find important and having predictions about these questions informs my worldview. But more importantly Forecasting forces you to develop deep mental models about a question. And these mental models can then be used in the future when you are thinking about AI developments. Thus they make you overall smarter at thinking about AI!

It was especially useful to do forecasts in parallel with a friend and discuss our approaches afterwards. This gave me quick feedback on my methodological approaches and pointed out which considerations I had missed. I highly recommend it! Thank you to Charles for being my Forecasting partner.

Additionally, by contributing to the aggregate Metaculus prediction scores and publishing my reasoning here, I hope I can help to improve the information on which others make decisions and contribute to the broader conversation about AI development.