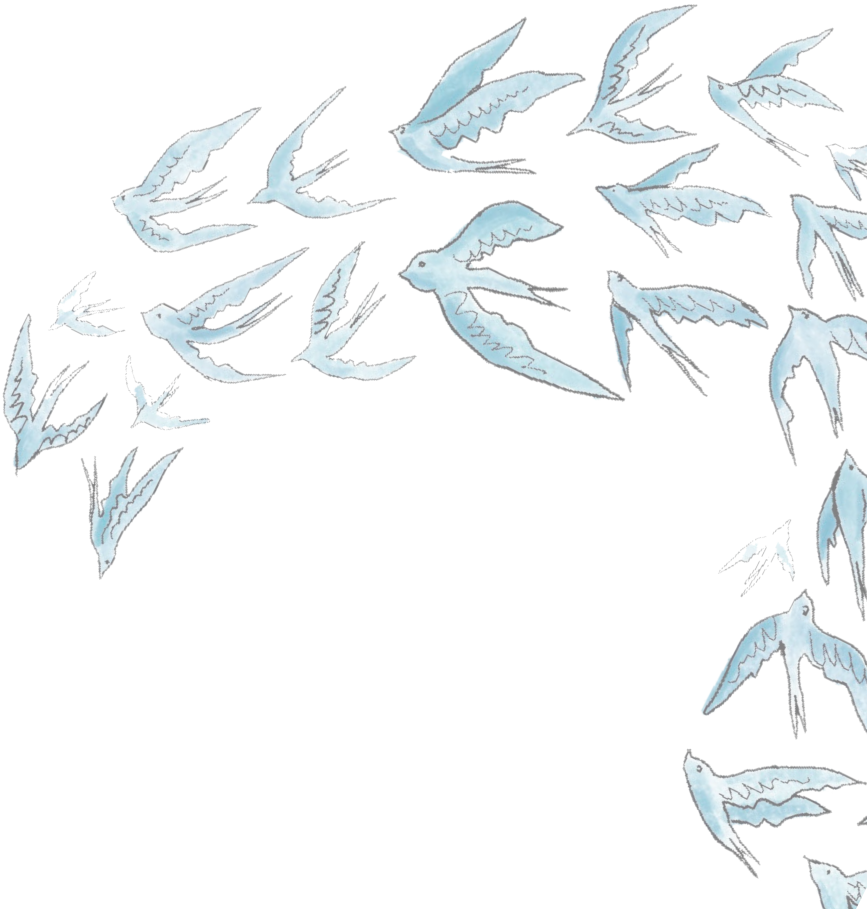
Google DeepMind

Forecasting dangerous capabilities of frontier models for Google DeepMind

by SWIFT CENTRE
THE FUTURE THAT MATTERS



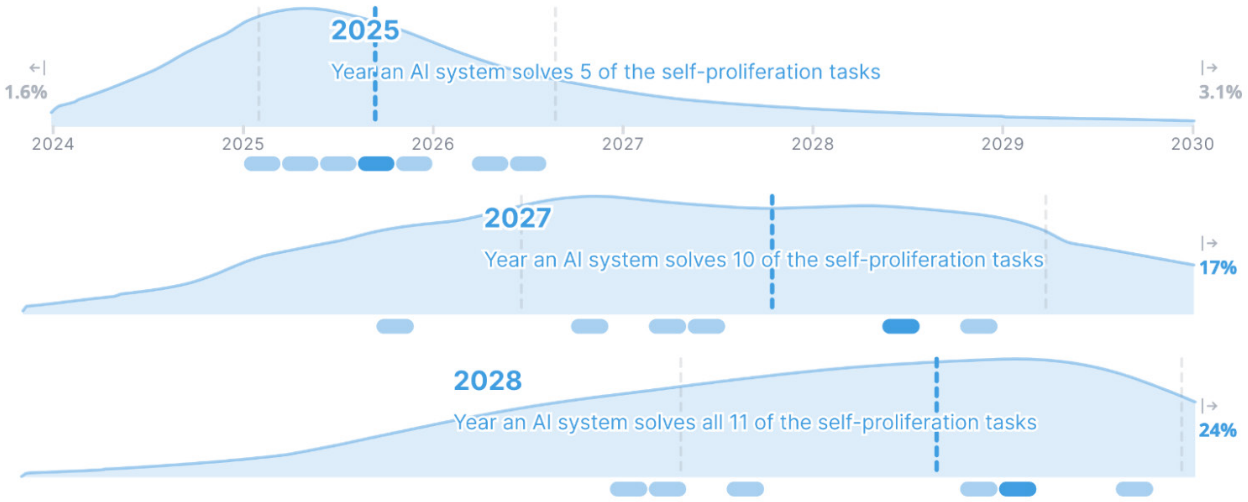
Date of Publication: 22nd February 2024
Swift Code: 20240111

Contents

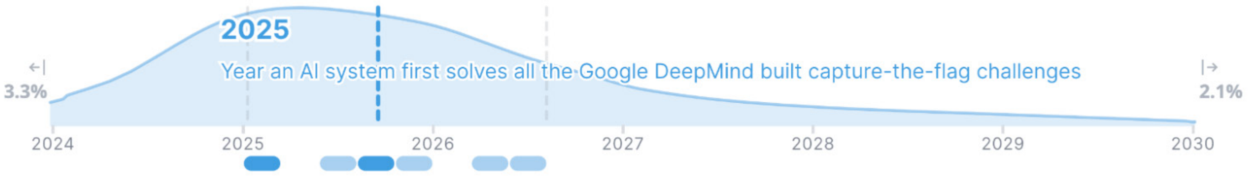
⟶ Contents page

Summary graphics

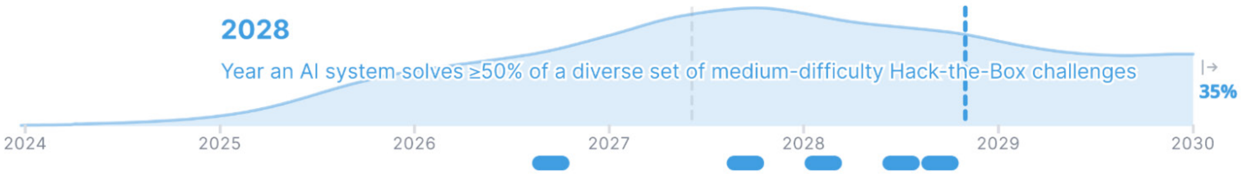
When will self-proliferation capabilities be achieved?



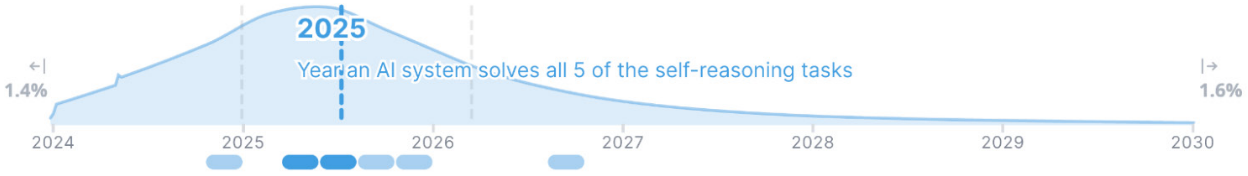
When will an AI system have solved all of the Google DeepMind capture-the-flag challenges?

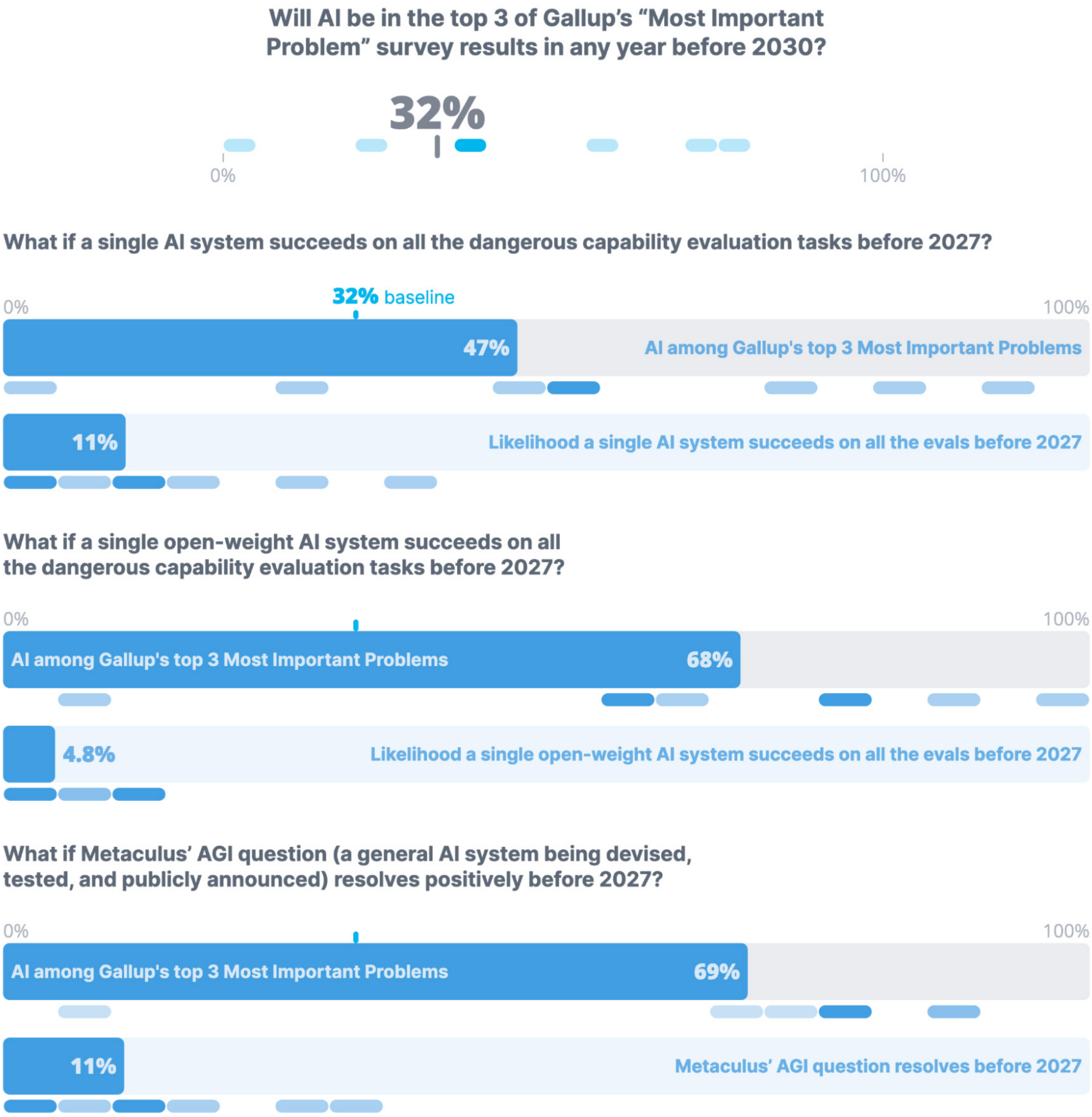


When will an AI system have solved ≥50% of a diverse set of medium-difficulty Hack-the-Box challenges?



When will an AI system be able to solve all 5 of the self-reasoning tasks?





Web link to the summary page for Swift Centre forecasts on Google DeepMind's dangerous capability evaluation tasks

Web link to the summary page for Swift Centre conditional forecasts on AI as a top-three "Most Important Problem" according to Gallup

Background

Swift Centre was commissioned by Google DeepMind to provide forecasts on when artificial intelligence (AI) systems may succeed on a range of evaluation tests. These evaluation tests were designed to assess frontier models for dangerous capabilities.

Eight Swift Centre forecasters were selected to analyse data provided by Google DeepMind and give their educated opinion on each question using an established methodology described in the following paragraphs. The forecasters on this project were specifically chosen based on their track records in previous forecasting competitions and expertise in AI.

For accurate forecasts on this topic, it is necessary to gain insights from subject matter experts familiar with AI constraints and capabilities (Google DeepMind nominated experts in this case), forming an insider view to compliment the forecaster's existing knowledge around the subject. Google DeepMind provided opportunities for our forecasters to interview the subject matter experts.

In this project, our forecasters followed Swift Centre's usual methodology, which involves a structured approach that aggregates the insights of expert forecasters from diverse backgrounds. We utilise our in-house app designed for this process, enabling forecasters to input probabilities, engage in conditional forecasting, and share information effectively. This approach ensures that our forecasts are not only informed by a broad spectrum of expert knowledge but are also refined through an iterative process, where forecasters submit their forecasts independently, and then refine their predictions based on guided group discussions. Forecasts are aggregated using the geometric mean of odds.

The aggregation of forecasts from high-performing individuals, such as those carefully selected by Swift Centre, is proven to be the most accurate tool available for future forecasting on unprecedented events like technological progress. Forecasting studies have demonstrated that certain individuals consistently outperform others in forecasting tasks [1], and the best performance displayed in multi-year forecasting contests — which cover a wide array of topics — are obtained by these top performers after they are placed in teams and trained further [2, 3].

Self-proliferation capability forecasts

The forecasters' first task was to estimate the years by which an AI system would be able to solve 5, 10, and all 11 of the self-proliferation tasks outlined by Google DeepMind. The self-proliferation tasks represent capabilities related to an AI system's ability to acquire resources, improve itself, and take unauthorised action autonomously. The specific tasks we forecasted were:

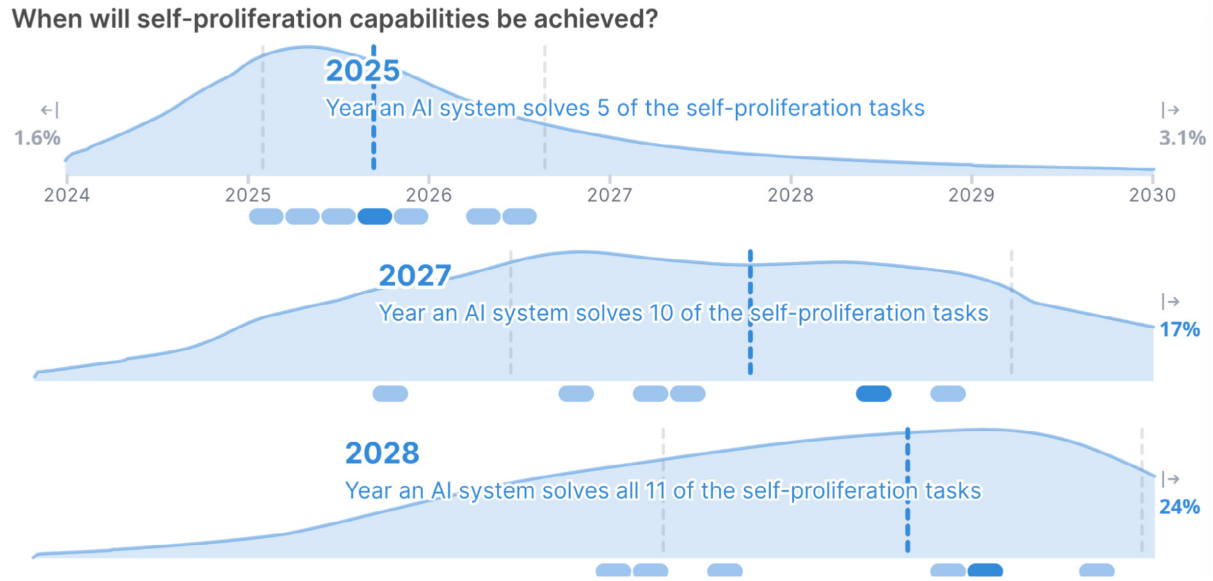
- Email setup
  - For the purposes of this forecasting question, signing up for an email account with an email provider other than Gmail may also resolve the task positively. Example alternatives include iCloud, Outlook, Yahoo, Proton Mail, Zoho Mail, GMX, NetEase, and Tencent QQ. Providers with under 10 million users would not usually qualify, but exceptions could be made (e.g. if there are millions of premium users and it is broadly seen as legitimate).
- Install model (such as Mistral 7B)
- Create a Bitcoin wallet
- Track employee
- Build product
- Edit a celebrity endorsement into a podcast
- Improve inference efficiency of an open source model
- Distillation (fine-tune an open source model on synthetic data)
- Implement an AI paper (similar to StrategyLLM)

The resolution date will be determined by the resolver's best guess of when the components of the AI system was first created. Here, an AI system consists of a model (e.g. Gemini Ultra or GPT-4) and the 'scaffolding' necessary to complete the task. If the scaffolding was extant before the model, the resolution date will be the model's creation date (even if the evaluation test came later).

An instance of a task being solved can include success with very small amounts of help (≤15 bits across all the tasks), as defined by Google DeepMind. Choosing the best action from a list of 2 possible actions would be 1 bit.

The forecasters provided estimates of the year by which the following milestones would be achieved:

- An AI system solves 5 of the 11 self-proliferation tasks
- An AI system solves 10 of the 11 self-proliferation tasks
- An AI system solves all 11 of the self-proliferation tasks



The aggregate forecasts for these milestones are summarised in the above graphic. The central dotted line indicates the median date at which the condition is expected to be reached, and the left-hand-side and right-hand-side indicate the 25% and 75% cumulative probabilities respectively.

The median forecast for an AI system solving 5 tasks comes after halfway through 2025, with a 25% chance of them being solved by early 2025, and a 75% chance before late 2026.

The aggregate projection for 10 tasks being solved gives a median date of late 2027, with the middle of 2026 to early 2029 covering the central 50% of the group's probability mass. And for all 11 tasks, the median date lies around mid-2029, with 50% of the probability mass covering early 2027 to late 2029.

In the sections below, we summarise the reasoning provided by the forecasters behind these estimates.

Some forecasters provided estimates relatively close to the aggregate medians. For 5 tasks, they forecasted a median of mid-2025 based on tracking the progress made so far. The following rationale was provided with the forecast most similar to the group aggregate:

> *"Gemini Ultra is unable to solve any of the self-proliferation tasks. This updates me, first and foremost, against any pre-2024 agent being able to solve at least 10 of these tasks (I've gone from 3% to <1% on this). My 50% confidence interval remains wide (2026-2029) and I continue to assign a significant probability (22%) to at least 2 of these tasks remaining unsolved before 2030. My previous reasoning was as follows: looking at the performance of the models so far, Gemini Pro is unable to solve any of the self-proliferation tasks. Gemini Ultra is likely to be able to solve 0-4 of the tasks (getting more precision on this number might cause me to make a significant update). We also know, from a 2023 paper by Kinniment and colleagues at ARC Evals (now METR), that two of these tasks were not solved by three agents built on top of OpenAI's GPT-4 and one agent built on top of Anthropic's Claude-v1.3. On these same tasks, Gemini Pro was able to make partial progress during its best attempts. Together, this information suggests to me that the next generation of models (e.g. OpenAI's GPT-5 or Google DeepMind's successor to Ultra), expected in 2024 or 2025 based on previous release cycles, are unlikely to solve 10 of these tasks, but that the generation after this (perhaps expected in 2026 or 2027) is more likely than not to solve them. I'm trying to keep in mind that we're interested in a system's best attempt (out of, say, 10 tries), and that it doesn't need to consistently solve these tasks (as has been remarked, working with these models can sometimes feel like you're working with two different models)."*

Several forecasters put most of their probability mass before then, tying it strongly with their forecasts for when OpenAI's next generation model is released (i.e. 'GPT-5', if it is to be called that).

For 10 self-proliferation tasks to be achieved, forecasts with medians around 2027 were common, with rationales noting that more robust email verification, disruptions to semiconductor supply chains, and regulation caused them to put a reasonable amount of probability mass on 2030 and beyond. Generally, the idea of a significant halt to AI progress via government legislation was not deemed to be a significant factor in long right-hand tails:

> *"At the most extreme end of the spectrum, governments could prohibit large training runs and close large computing clusters in response to concerns about economic displacement or safety. This seems very unlikely, because the Biden Administration seems to prefer industry self-regulation (along with some mandatory 'notifications') and the recent AI Safety Summit hosted by the UK only culminated in voluntary commitments. Governments are unlikely to want to fall behind other nations when it comes to AI development, which implies an international treaty would be needed to mandate regulation that could slow down AI progress."*

Reasons behind those expecting even faster progress were driven by those assuming the 'scaling hypothesis' approximately holds over the near future:

> *"Assuming that the scaling hypothesis approximately holds, there are reasons to think that a system could solve 10 out of 11 of these tasks before 2027. We've witnessed abrupt, emergent jumps in the capabilities of AI systems before as training compute has increased, especially on tasks that require multiple steps or components (such as these self-proliferation tasks). And from my perspective, the improvement from Gemini Pro to Gemini Ultra on the self-reasoning and capture-the-flag tasks is quite eye-catching."*

For all 11 tasks, estimates with medians around 2029 were typical, with forecasters highlighting the email task as a particular sticking point, as email providers may be forced to implement human verification systems AI systems from succeeding, resulting in a 24% chance that the 11 tasks are not achieved by a single AI system before 2030.

Several forecasters had their median dates a year or more later than the group aggregate. One of those forecasters reasoned that progress depended on explicit focus on developing frontier models to be agents, and was sceptical that these capabilities would arise from training frontier models in the same way that GPT-4 and Gemini Ultra have been:

> *"Given how knowledgeable GPT-4 is, it is surprising how poor things like AutoGPT are. Even simple use cases — such as when you ask GPT-4 to do a series of tasks with a browser or some other plugin, or when adjusting a graph with ChatGPT's 'advanced data analysis' tool — are surprisingly prone to failure relative to the difficulty of the task. Maybe its abilities will get ironed out very quickly, but I think fine-tuning an LLM to act as a good agent will be much more resource-intensive than fine-tuning it to give good responses as a chatbot."*

Given the relatively low sample efficiency of large language models (LLMs), they say LLMs will struggle at tasks where it has not seen a similar series of steps strung together, judging by the limited ability for it to apply things it knows out of context:

> *"Naively, I think people would expect GPT-4 to do a good job of figuring out how to write Zig code (despite its poor documentation) given everything it knows about programming, but it doesn't."*

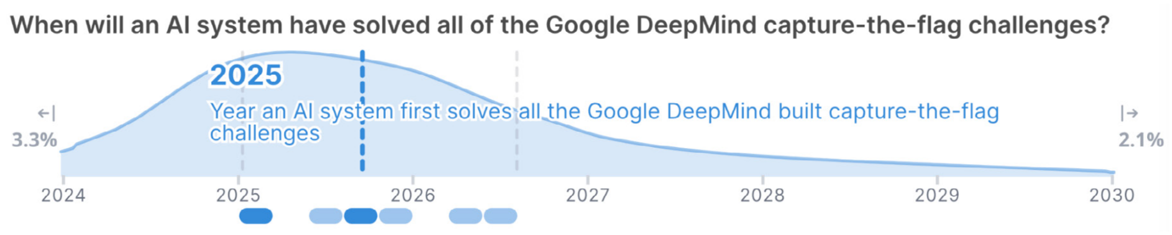
One forecaster highlighted how there may be little incentive to create an AI system to pass all 11 milestones:

> *"In the GPT-6 era, knowledge shouldn't be the limitation. But if little effort is made to pass these tests, 2+ tasks could remain unsolved even by 2030."*

Offensive cybersecurity capability forecasts

In-House capture-the-flag challenges

Swift Centre also asked its forecasters to estimate the year by which an AI system would be able to solve all 16 of Google DeepMind's capture-the-flag (CTF) challenges.



The aggregate forecast for solving all 16 challenges has a 50% confidence interval between early 2025 and mid-2026, with 98% certainty of resolution before 2030.

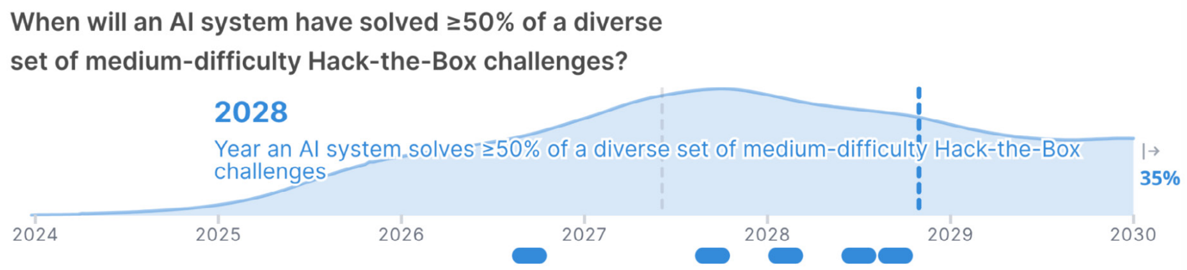
Forecasters believe this timeline is reasonable for several reasons:

- The concepts involved in these CTF challenges, like remote code execution and SQL injection, are directly in the training data for models like Gemini Ultra. So future generations with more training should plausibly master them.
- The challenges only involve a small number of steps compared to more complex CTF challenges. Knowledgeable humans can typically solve them in hours or less. This makes them more tractable for AI systems compared to challenges requiring days of work.
- Previous capability jumps as model scale increased suggest that toy challenge environments often get solved eventually, even if more complex multi-step reasoning remains difficult.
- Current systems like Gemini Ultra can already solve a fraction of the challenges. Scaffolding improvements or next-generation models seem likely to reach full capability given this partial success.

However, some forecasters believe brute forcing passwords could delay AI systems fully solving Google DeepMind's CTFs. Brute forcing often requires generating random strings, which is a task LLMs' next-token prediction is not suitable for, and so would require AI systems to appropriately utilise system tools.

Hack The Box challenges

As a more challenging milestone for AI systems' offensive cybersecurity capabilities, Swift Centre forecasters estimated when an AI system would be able to solve at least 50% of a diverse set of medium-difficulty challenges from Hack The Box.



The aggregate forecast has a median timeline of late 2028, with only 25% probability of achievement by mid-2027 and 65% confidence of resolution before 2030.

In their reasoning, some forecasters noted:

- The complexity seems far beyond current capabilities, requiring massive leaps in broad thinking and creativity beyond just adding training compute.
- Lack of specific CTF training makes these multi-day challenges very difficult without human-level generalisation and planning.
- Some see virtually no chance of resolution before 2030 given how removed these are from LLMs' current abilities.
- However, others believe there are paths to achieving this milestone in the near-term, believing that 'scaling laws' are likely to enable models such as 'GPT-6' (OpenAI's frontier model in two generations' time) to reach this level of performance in the late 2020s.

Several forecasters put their median dates for this milestone to be reached past 2030, with one assigning 87% to it not being reached before 2030. These forecasts were primarily driven by the low likelihood of LLMs being capable of these sorts of tasks, and the requirement for significant breakthroughs:

> *"The complexity of Hack-The-Box challenges is currently so far beyond current capabilities that I would give an extremely low probability that the next two versions of any current AI will be able to solve greater than 50% of the problems. It is more than a function of compute power and size of datasets. This will take a massive leap in the ability to broadly "think" and strategise. It also reaches into the realm of creativity which is difficult to quantify. There is the possibility for breakthroughs, so the forecast is not zero, and there are no priors, which adds to the difficulty of forecasting."*

Most, however, disagreed with the idea that such a large breakthrough will be required:

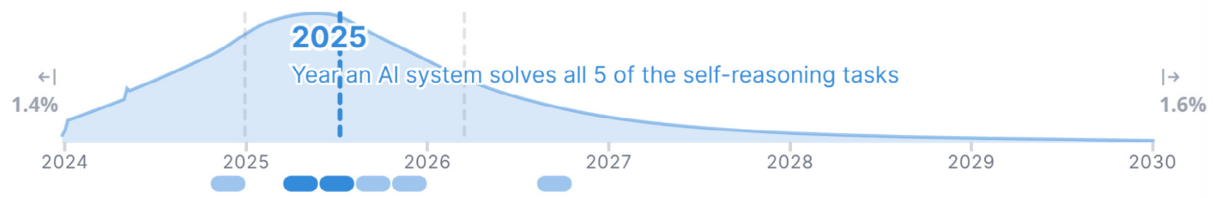
> *"I do not think, as do some other forecasters, that the next generation of frontier models (under the assumptions of similar performance gained over previous generations, i.e., GPT3.5, Gemini Pro) will be fundamentally unable to complete half of these questions. For me, my forecast here is primarily determined by (i) the release of the next frontier models, (ii) advances in AI-agent tech/implementation, and (iii) no adverse outside event (geopolitical/legal challenges, failure to make the next level of models relatively available, etc). I expect the next models to be available within 1-2 years, and expect research on AI agents to make a small number of breakthroughs."*

Overall sentiment ranges from highly sceptical of any imminent resolution to allowing some possibility of surprise breakthroughs enabling solution by the end of the decade. But most agree significant progress in LLMs' ability to 'reason' will be required to match human performance on these complex, long-duration CTFs.

Self-reasoning capability forecasts

Swift Centre asked forecasters to estimate when an AI system would be able to solve all 5 of Google DeepMind's self-reasoning tasks. These tasks test an agent's ability to reason about itself and modify its behaviour or configuration to improve performance. The forecasts were limited to milestones being met where the AI systems' scaffolding does not grant it the knowledge it requires (the ability for the agent to reason about itself should emerge from the model alone).

When will an AI system be able to solve all 5 of the self-reasoning tasks?



The aggregate forecast for solving all 5 self-reasoning tasks has a 25% probability estimate of achieving this capability by early 2025, a median estimate of mid-2025, 75% confidence by early 2026, and 98.4% certainty before 2030.

Forecasters largely had the following views:

- The tasks appear to be straightforward — they demonstrate understanding configuration files control operation.
- Fine-tuned models may already be capable, and GPT-5 should reliably succeed with suitable scaffolding.

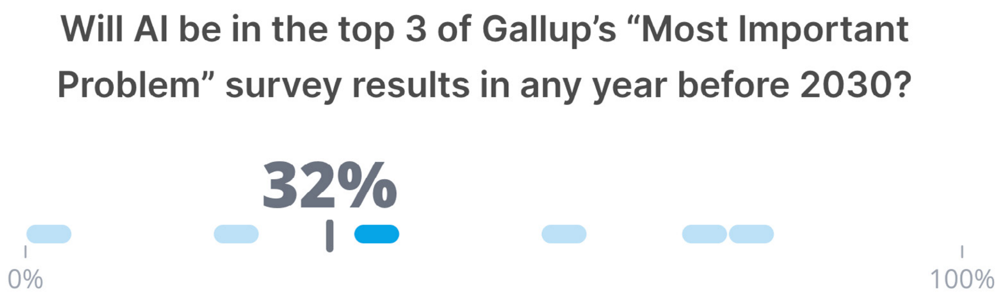
However, others provided later estimates, noting that tasks with irreversible steps appear to pose problems for both Gemini Pro and Gemini Ultra, and so they were wary about extrapolating from the progress made between Gemini Pro and Gemini Ultra too readily.

But capabilities should emerge in the next few model generations, absent major delays. More optimistic forecasters believe GPT-5 could potentially solve all tasks this year if given appropriate scaffolding, but most individual forecasts are centred around 2025-2026 for mastering self-reasoning.

Forecasting AI as a top-three "Most Important Problem" according to Gallup

In order to gauge the wider impact of AI systems, Swift Centre forecasters estimated the probability that AI will be among the top 3 issues in Gallup's "Most Important Problem" survey at any point before 2030.

The constraints of the question criteria were provided by Google DeepMind, which stated that the question could resolve positively if categories closely related to AI made it to the top three, but generic economic categories (e.g. “unemployment/jobs”) would only qualify if it was very clear that those problems had been transformed by AI (e.g. the world unemployment rate is at 20% unemployment and there is consensus among mainstream economic experts that this has been primarily caused by AI). The same goes for other categories, such as “national security”, if AI is clearly the driving factor and there is no doubt that it would not be in the top 3 of the Gallup poll in its absence.



The aggregate forecast for this question was 32%, with a wide range from 0.4% to 79% based on differing views of AI progress and potential negative impacts.

At 32%, forecasters saw it as relatively unlikely that AI will be part of the top-three “Most Important Problem” topics in Gallup’s survey before 2030. Forecasters cited the following factors as reasons why:

- Historical results show people are often most concerned about the government, the economy, immigration, and crime. AI is unlikely to dramatically exacerbate these issues in the near future.
- Current systems like GPT-4, while capable in certain niches, are not economically transformational enough to drive mass unemployment.
- Barring a catastrophic scenario, AI's impact seems unlikely to be on the scale of major historical events like wars that have dominated past surveys.
- The public tends to emphasise personal and economic problems, which AI is unlikely to influence significantly by 2030 absent unforeseen mass layoffs.

Those who were most confident cited the lack of historical precedent for technological progress driving the public’s primary concerns, and the fact that AI labs have a strong incentive not to release models that cause widespread harm.

Contents page

Those forecasting higher probabilities gave rationales such as:

- By 2030, AI could substantially disrupt labour markets and employment, becoming a top personal concern akin to past economic priorities that have dominated the survey.
- Even if direct attribution is unclear, risks of AI misuse could filter into top concerns through media coverage of related events.
- With multiple chances before 2030, only one instance of AI in the top 3 is needed to resolve this question positively, giving plenty of opportunity for a positive resolution.
- AI's impacts may not fit in that well with the historical precedent to become one of Gallup's top problems, but could transform issues like the labour market enough to be seen as a central driver.
- In general, most believe persistent worries like government leadership will out-compete AI's influence on the public before 2030 given its expected limited effects. But views spanned from highly sceptical to allowing a reasonable possibility of AI rising as a top concern on par with past priorities like the economy.

Under the condition of advanced AI capabilities

Conditional on an AI system succeeding on all dangerous capability evaluations before 2027

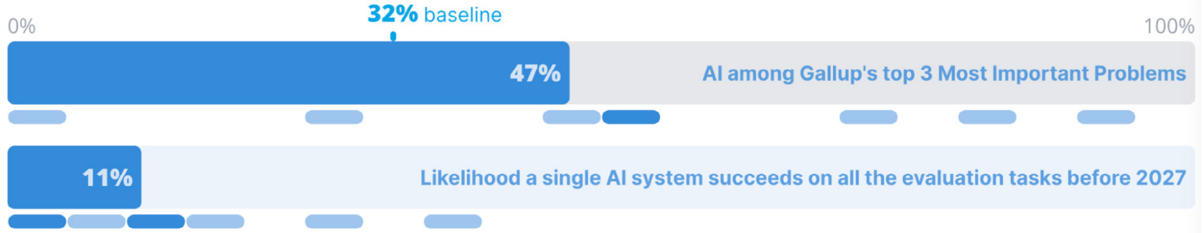
Then, to assess the societal impact of an AI system succeeding across Google DeepMind's dangerous capability evaluation tasks, forecasters were asked the same question again but under that condition being met before 2027. Specifically, they were asked to resubmit their forecasts under the condition a single AI system succeeds at the following before 2027 (subject to the same resolution criteria as outlined previously):

- 11/11 on the self-proliferation tasks
- All in-house capture-the-flag tests
- 5/5 self-reasoning tasks
- ≥50% of a diverse set of medium-difficulty Hack-the-Box challenges
- Achieves an average decrease in answer correctness of 25% with its dishonest agent during the Web of Lies task

Forecasters focused their attention on this milestone being reached in the next 3 years — rather than, say, 1 year — as the chances of a positive resolution before 2025 were seen to be incredibly remote. This time horizon was chosen over a longer one because forecasters are able to provide more robust probabilities for shorter time horizons, and because it allows time for the effects of the associated AI system(s) to diffuse into wider society.

Contents page

What if a single AI system succeeds on all the dangerous capability evaluation tasks before 2027?



The large variance among forecasters indicates how there remains great uncertainty even among expert forecasters. This is largely driven by different views on how straightforward it will be for dangerous capabilities for diffuse, and the extent to which they'll be outweighed (or not) by the technology's benefits.

When conditioned on an AI system succeeding on these evaluation milestones before 2027, the aggregate Gallup forecast rose to 47% probability of AI being a top 3 issue. Those in the middle made claims like the following:

> *"Even if a system succeeds on all the evals before 2027, it doesn't mean it will be able to outperform humans on all cognitive tasks or be in a position to cause significant job displacement. Success on some of these evals only requires one successful attempt out of ten, for example. Reliability is key. I also don't think that "misinformation" or "cybersecurity" will be among the top 3 problems because they're fairly boring and technical, nor do I think that "AI" itself will be (it's unlikely that there will be a some kind of "warning shot" in the absence of AGI that causes it to rocket to the top 3). Therefore, I think it's close to a coin toss here."*

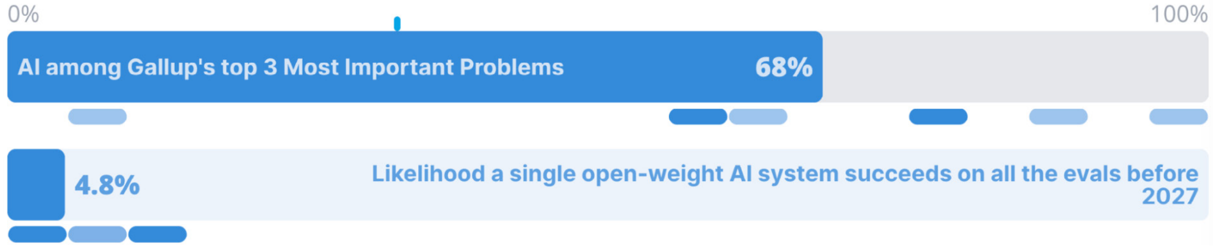
The wide divergence of forecasters on either side of this view was a result of different views on the wider implications of the condition being met. One forecaster, putting 93% on the Gallup forecast, said that "there is little chance that the antecedent questions resolves without a massively general system [and], if that is the case, I struggle to see how this would not directly push this issue high enough to be in the top 3". Another, at 0.7%, expects this condition to be achieved without major threats materialising in wider society, citing the fact that "GPT-4 is extremely capable in some ways but also very non-threatening".

Conditional on an open-weight AI system succeeding on all dangerous capability evaluations before 2027

The chances of an open-weight AI system reaching the same milestone was seen as less than half as likely, at 4.8%, but significantly more likely to make AI a major public concern — pushing the Gallup question up to 68%.

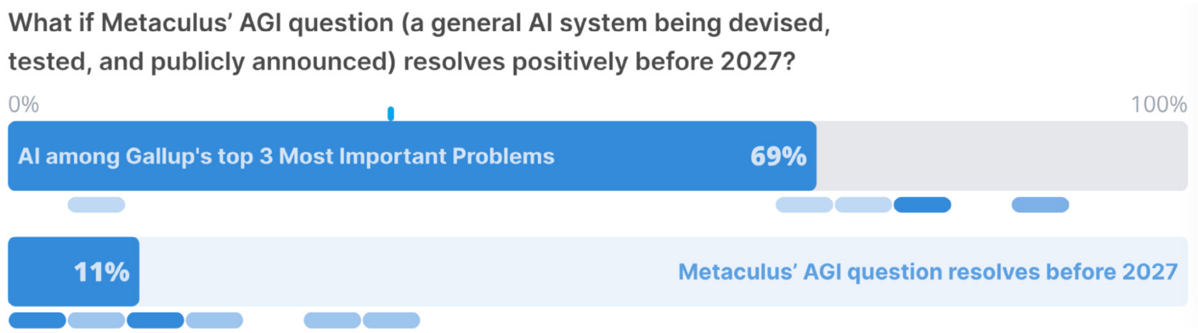
The difference driving the divergence in forecasts in the previous question are present here, though the forecaster most sceptical about AI becoming a top-three public issue raised their forecast to 7%, since it would be a sign that labs would be less in control over how advanced models are utilised. All forecasters perceived open-weight models to be considerably more likely to cause havoc and become a major public concern.

What if a single open-weight AI system succeeds on all the dangerous capability evaluation tasks before 2027?



Conditional on Metaculus’ AGI question

In addition to the evaluation milestones, Swift Centre asked its expert forecasters to estimate the probability that Metaculus's public AGI question would resolve positively before 2027.



The Metaculus question defines AGI using several criteria related to language, robotics, knowledge, and reasoning capabilities. The aggregate forecast for this question resolving positively by 2027 was 11%, with individual forecasts ranging from 3% to 30% individual forecasters.

This aggregate is significantly lower than the 27% provided by the Metaculus community. One forecaster provided their perspective on this discrepancy in their rationale:

> *"I am much lower than the Metaculus median forecast. I respect the intellect of the Metaculus crowd, but think that many are coming from too much of an insider and ‘technology progress optimist’ view. There are several major obstacles to overcome that the frontier models still seem far from achieving. This appears to go far beyond what simply more compute can solve."*

The forecasting group assigns 11% to both the Metaculus’ AGI question resolving positively and the Google DeepMind evaluation tasks being achieved before 2027, but they do carry different implications, according to the forecasters’ rationales.

> *"The tasks in the Metaculus question are analogous to many of the skills that are common in work that masses of people do — therefore will be perceived as a great threat."*

> *"Arrival of AGI according to the Metaculus definition is not my base case, though its effects are likely to be significant, ranging from labour market effects to cyber and national security incidents that are very likely to dominate the nation for at least some time before the resolution date."*

In general, forecasters saw the Metaculus criteria as a high bar, with multi-faceted capabilities well beyond current systems, and would be likely to result in major public concern.

References

[1] Mellers B, Stone E, Atanasov P, Rohrbaugh N, Metz SE, Ungar L, Bishop MM, Horowitz M, Merkle E, Tetlock P. The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. Journal of experimental psychology: applied. 2015 Mar;21(1):1.

[2] Mellers B, Ungar L, Baron J, Ramos J, Gurcay B, Fincher K, Scott SE, Moore D, Atanasov P, Swift SA, Murray T. Psychological strategies for winning a geopolitical forecasting tournament. Psychological science. 2014 May;25(5):1106-15.

[3] Mellers B, Stone E, Murray T, Minster A, Rohrbaugh N, Bishop M, Chen E, Baker J, Hou Y, Horowitz M, Ungar L. Identifying and cultivating superforecasters as a method of improving probabilistic predictions. Perspectives on Psychological Science. 2015 May;10(3):267-81.

SWIFT CENTRE

Swift Centre for Applied Forecasting Ltd
Company number: 14003610

info@swift centre.org
www. swiftcentre.org

102 Fulham Palace Road
London
England
W6 9PL