AI safety forecasting questions

Nuño Sempere, David Mathers, Gavin Leech and Misha Yagudin

December 6, 2023

tl;dr: This document contains a list of forecasting questions, commissioned by Open Philanthropy as part of its aim to have more accurate models of future AI progress. Many of these questions are more classic forecasting questions, others have the same shape but are unresolvable, still others look more like research projects or like suggestions of data gathering efforts. Below we give some recommendations of what to do with this list, mainly to feed them into forecasting and research pipelines. In a separate document, we outline reasons why using forecasting for discerning the future of AI may prove particularly difficult.

Table of Contents

Recommendations

We recommended that Open Philanthropy feed these questions into various forecasting and research pipelines, with the thought of incentivizing the research needed to come up with good models of the world around AI developments.

We have categorized questions with three stars in various buckets, each of which has its own recommendations:

- Questions marked by (FE), are good targets for Fermi estimates. In isolation, each of those Fermi estimates might not mean much, but in aggregate, a bunch of them might contribute to having better models. Open Philanthropy could put bounties for good Fermi estimates on these questions, or solicit bids from researcher interested in doing Fermi estimates.
- Questions, marked by (RP), require more involved research to arrive at a high-quality answer. They could be a good fit for institutions that can produce research that takes longer, like think tanks, research fellowships, etc.
- Questions, marked by (FT), are good fits for forecasting tournaments. These could be put up on a platform, like Manifold Markets, Polymarket, or Metaculus. A forecasting group, like Samotsvety, could also be useful for questions where there is no clear resolution method. In this adjacent document, we also outline a "resolution council", which could subjectively resolve questions which would otherwise have no resolution source.
  - Metaculus previously had a large AI tournament, which seems to have seen bounded success. It is possible that Manifold would be able to more capably set up a similar tournament, in a way that would be better across many dimensions, so if there is only space for one forecasting platform, we would recommend putting these questions up on Manifold this time for the value of information.
- Some questions, marked by (UF), have a similar shape to a forecasting question, but they might be unresolvable or too difficult to resolve. If might be worth commissioning a forecasting group like Samotsvety to forecast on them.
- Some questions, marked by (DG), are good targets for data gathering, in the style of Our World In Data (OWID). Perhaps OWID itself could be commissioned to gather and present some of this data. Other groups, like Epoch or AI impacts, might also be interested in looking into this.

Note that the boundary between questions which could be in a forecasting tournament (FT), and questions which we deem to be unresolvable with a reasonable amount of effort (UF) is fairly arbitrary. Fewer questions would be suitable for a forecasting tournament on a platform like Metaculus, which seeks to have explicit and rigorous questions. More would be suitable for a tournament or list of questions on Manifold Markets, which has more of an "anything goes" attitude.

We have also worded many questions in terms of a "resolution council", which would make them more resolvable, if you had a resolution council willing to go

through the effort of coming up with a subjective judgment on the question topic. For an explanation of what a resolution council could be, see here

Questions

Recurring terms

An specification for a [resolution council] is discussed in a separate document, here.

"Leading lab" is defined as a lab that has performed a training run within 2 orders of magnitude of the largest ever at the time of the training run, within the last 2 years.

A floating point operation (FLOP) is here defined as one addition, subtraction, multiplication, or division of two decimal numbers, whatever their size. So doing subtracting two 64 bit floats would here correspond to one FLOP, as would subtracting two 8 bit "mini-floats". See this document for a short discussion of this point.

"Automating some fraction of labour" is operationalized as follows: - Consider all human work hours in 2023 and their intended outputs. Then at the question resolution year, when aiming to produce the same types of outputs, how many fewer human hours will one need to achieve the same kind of output, or a close substitute? - For example, consider all hours spent on secretary work in 1960. This work commonly requires much less time, since people no longer dictate to secretaries, and instead draft emails themselves. But some labour is still needed, so we might estimate that 95% of that work has been automated. - Note on substitutability: McDonald's uses some screens for customers to place orders, and together with a counter, this substitutes for waiters. However, these don't provide exactly the same experience of a waiter: the waiter might be more attentive, or add a "human touch". For cases such as this, consider the work to have been automated, even if it has been replaced by a close rather than exact automated substitute. - Note on effort required for question resolution: The above operationalization means that questions using it might need a small research project to estimate a resolution. But then, it's also possible that in the course of researching this topic, one could come up with better operationalizations of "labour automated".

Key

- *** = Highest Importance
- ** = Medium Importance
- − = Lower Importance

Questions relevant to speed of capabilities progress

- ** Exotic computing techniques 1: What is the chance that (at least one of) the leading commercially available GPU chip(s) are 3-d by 2031? (Reasons for inclusion: these are highly relevant to the predictions of Ajeya Cotra's bio anchors model, because it expects available compute to grow largely through these, and thinks we need quite a lot more compute before there's a high chance we can train a transformative model[1])
- ** Exotic computing techniques 2: By what year will 3k optical computers first be sold?
- – Exotic computing techniques 3: By what year will 3k quantum computers first be sold?
- – Exotic computing techniques 4: By what year will 3k computers which employ reversible computation be sold?
- – Self-driving cars: In what year will the first 10k Level 5 self-driving car[2] be sold? (Reason for inclusion: as a test case, since if we can't build good-as-humans-in-all-circumstances self-driving cars by year X, we probably[3] can't build a world-destroying agent).
- *** (FT) What will [resolution council]'s best estimate of the number of FLOPs used in the largest training run be by Jan 1st 2024/2025/2028/2030/2035/2040/2050?
  - Note to forecasters: It might be more intuitive to consider the number of doublings over the largest training run, rather than the absolute amount.
- *** (FT) In 2025, how many FLOPs per second per dollar will NVIDIA's latest and most advanced chip offer?
- – By 2030 will there be a break in the current trend line for FLOPs per $ spent?
- ** Will progress stall for at least 4 years at any point before state–of-the-art accuracy on MMLU dataset reaches 100%?
  - "Stalled" could be operationalized as something like 'less than 2% improvement in best score over a 4-year period)[4]
  - Beware for this and other benchmark questions: Open AI may not have been careful enough about making sure benchmarking questions (and their answers) or very close variants of them were not

[1] At least absent a very, very large amount of algorithmic progress.
[2] https://en.wikipedia.org/wiki/Self-driving_car#Classifications
[3] Admittedly, I'm basing this off of raw intuition, not any particular argument.
[4] perhaps using this leaderboard: https://paperswithcode.com/sota/multi-task-language-understanding-on-mmlu

in GPT-4's training data. If these people are right GPT-4's performance on benchmarks isn't a meaningful test of its intelligence. And nor is comparing another model's performance on the benchmark to GPT-4's a meaningful test of whether the model is better at reasoning than GPT-4: https://aisnakeoil.substack.com/p/gpt-4-and-professional-benchmarks.

- ** Will progress stall for at least 4 years at any point before state–of-the-art accuracy on the interview level problems on the APPS benchmark reaches 30% (pass@1)?[5] (Stall=less than 2% improvement in best score over a 4-year period).
- ** Will progress stall for at least 4 years at any point before state–of-the-art accuracy on the competition level problems on the APPS benchmark reaches 60% (pass@1) ?[6]. (Stall=less than 1% improvement in best score over a 4-year period).
- – When will a model score the maximum possible score on the MMLU?
- – By how much will the best score on the MMLU increase between 2023 and end of 2027?
- *** (FT) 2025/2030/2035 Penn Machine Learning benchmarks database halvings of compute.
  - Question details: Consider all the benchmarks in the Penn Machine Learning benchmarks database (https://epistasislab.github.io/pmlb/). Consider the estimated compute needed to achieve state-of-the-art performance in April 2023, C1. Consider the estimated compute needed to achieve that same performance, which may no longer be state-of-the-art, in April 2025/2030/2035, C2. Define the number of halvings for a specific benchmark as log2(C1/C2). Now consider the average of log2(C1/C2) for all benchmarks. This question resolves as the [resolution council]'s best estimate for that average in 2025/2030/2035.
- *** (FE, UF) How much money will be spent worldwide training AIs in the average year between 2024 to 2040, as a multiple of the amount spent in 2023? Question resolves according to the [resolution council]'s best estimate in 2040.
- ** In what year will the first high quality AI written scientific textbook be published/released for free online?
- – In what year will a large language model with a context window of 1 million/10/100 million words be released?

[5] https://paperswithcode.com/sota/code-generation-on-apps#:~:text=The%20APPS%20benchmark%20attempts%20to,as%20v
[6] https://paperswithcode.com/sota/code-generation-on-apps#:~:text=The%20APPS%20benchmark%20attempts%20to,as%20v

- – By 2025, in how much lower/higher a percentile on the SAT will the average entering comp sci major at a US college/university have, relative to 2019? (Relevance: tells us something about whether AI is currently attracting more talented students as the field becomes more hyped, which is probably partly predictive of future capabilities progress, as better researchers means faster progress.)
- *** (RP, DG) By what year will at least 15% of patents granted in the US be for designs generated primarily via AI?
  - Reasons for inclusion: both an early sign that AI might be able to design dangerous technology, and an indicator that AIs will be economically useful to deploy across diverse industries.
  - Question details: Question resolves according to the best estimate by the [resolution council].
- *** (UF, RP) How long will be the gap between the first creation of an AI which could automate 65% of current labour, and the availability of an equivalently capable model as a free, open-source program?
- ** In what year will a model first achieve 60/80/90/100% of the maximum possible score on the Abstraction and Reasoning Challenge dataset? https://arxiv.org/pdf/1911.01547.pdf
- – How long before there is an open source model which matches GPT-4's few shot performance on the MMLU benchmark? (See leaderboard here: https://paperswithcode.com/sota/multi-task-language-understanding-on-mmlu)
- ** How long before there is a single open source model which gets the same mean benchmark score as GPT-4[7]
- ** What's the chance China will invade Taiwan by 2025/2028/2033? (Relevance: Almost all chips used for cutting-edge AI training runs are manufactured in Taiwan (https://en.wikipedia.org/wiki/TSMC). If China invades the island, the supply of high quality chips might well disappear, slowing down AI progress.)
- – Conditional on China invading Taiwan in 2025/2028/2032, how much lower will the FLOPs per $ available for AI training runs be 1/2/5/10 years later, relative to a no invasion scenario?
- – By 2030, how many fabs[8] will TSMC own outside of Taiwan? (https://www.tsmc.com/english/aboutTSMC/TSMC_Fabs)
- – On average, how many hours of human labour will it take TSMC/Samsung/Intel to produce a design for a new semiconductor

[7] across the datasets mentioned in table 2, p.7 of this: https://cdn.openai.com/papers/gpt-4.pdf

[8] https://en.wikipedia.org/wiki/Semiconductor_fabrication_plant

chip in 2025/2028/2035?

- – How many new fabs will be built worldwide by any company by 2031/2041?
- – Number you're 90% confident there will be less new fabs built than by 2031/2041?
- – In what year will an AI first win the diplomacy world cup?
- – Conditional on no invasion of Taiwan, by 2027/2029/2033, will both NVIDIA and AMD still rely on TSMC to fabricate their chips? [I've relied on Wikipedia for the claim that they currently use TSMC, which could be out of date, since the reference is 2013, but I doubt it given TSMC's currently dominant market position] (Relevance: If a rival takes over some of TSMC's business, an invasion of Taiwan would make much less difference to the cost of chips, and hence hold back AI progress a lot less.)
- – In 2027/2032 what will be the share of GPUs used worldwide in AI research that were manufactured by TSMC?
- *** (RP, FE) What fraction of labour will be automated between 2023 and 2028/2035/2040/2050/2100?
  - – Question operationalization: See "recurring terms" section
  - – For a reference on an adjacent, see Phil Trammell's Economic growth under transformative AI.
- *** (FT) What % of GDP in 2028/2035/2040/2050/2100 will correspond to salaries, as estimated by the [resolution council]?
- *** (FT) Of all work hours in 2028/2035/2040/2050/2100, how many will be spent overseeing AI models?
- *** (RP) Meta-capabilities question: by 2029, will there be a better way to assess the capabilities of models than testing their performance on question-and-answer benchmarks?
- – What % of machine learning researchers with >7 years research experience would currently answer 'faster' to 'has progress been faster or slower in the 2nd half of your career compared to the first'?
- ** How much higher/lower is the risk of AI takeover by 2123 if we reach AI able to automate all cognitive (i.e. not requiring a body) labour by 2040 compared to by 2065?
- *** (UF) How much money will the US government cumulatively spend on training AI models between 2024 and 2040, as estimated by the [resolution council]?
- – How many $s in total will TSMC, ASML, NVIDIA (https://en.wikipedia.org/wiki/Nvidia) and SMIC spend on R&D between 2024 and 2030?

- – What fraction of Gross World Product will the semiconductor industry be responsible for in 2025/2030/2035?
- *** (RP, UF) How much money will the Chinese government cumulatively spend on training AI models between 2024 and 2040, as estimated by the [resolution council]?
  - Question details: Resolution is made difficult because a) China has a mixed communist/capitalist regime, b) China will probably not tell you directly. It is left to the resolution council to make judgment calls about this and give their best guess estimate of this number, according to their best good faith effort to estimate it.
- – When will Google release a model that outperforms GPT-4 on the MMLU benchmark? (https://paperswithcode.com/sota/multi-task-language-understanding-on-mmlu).
- ** On 1st of January 2025, in how many European countries will chatGPT or GPT-4 be banned?
- – By what year will a model first propose an experiment designed to generate data for AI training?
- ** Will any group other than Google, Meta, OpenAI, or Anthropic announce an LLM with similar or better performance to GPT-4 by the end of 2023?
  - One possible operationalization of "similar or better performance" could be operationalized with reference to benchmarks (e.g., this one, with reference to a blinded comparison, etc.
- ** Will a major English-language newspaper run a front-page story on the economic impacts of LLMs by 1st January 2029?
- ** What will the be the mean annual growth rate in the number of AI researchers worldwide for the years 2024-2036?
- – What will the be the mean annual growth rate in the number of STEM researchers worldwide for the years 2024-2036?
- – How many frames of a game will the leading system need to solve the Arcade Learning Environment in 2023, 2024, 2027?[9]
- *** (UF) How much risk of extinction/permanent takeover by AI comes from the first training run of 10/100/1000/10k/100k the size of that used to train GPT-4?
  - Question resolution details: This is more of a modeling question, and thus might be more suitable for setups in which scoring isn't used. A resolution council could also resolve this question, though that feels

[9] https://paperswithcode.com/dataset/arcade-learning-environment#:~:text=The%20Arcade%20Learning%20Environment%20

like it defeats the point a bit, since the resolution council wouldn't necessarily have more information than other forecasters.

- *** (UF) How likely is the first model trained with 10/100/1000/10k/100k of GPT4's compute to send out unauthorized copies of itself across the internet?
  - Two possible question operationalization options:
    * Condition on the median scenario across any relevant dimensions: amount of safety work/safety concern/algorithmic progress/regulation/etc. Pros of this approach: easier for forecasters to think about. Cons: the median scenario will be different across different forecasters, and so they will be conditioning on different things, in a way which might be difficult to communicate.
    * Ask about the probability across all scenarios: sum(probability of scenario × probability of the first model 10^n times the size of GPT4 sending itself through the internet). Pros: better conceptually defined, easier to understand for probability consumers. Cons: harder to think about for forecasters.
  - Note: One could also ask about "any model", or "any of the first 10", etc. instead of the first model 10x the size of GPT4.
- – When will an AI first achieve a silver-medal level performance on an [International Math Olympiad question-set] (https://en.wikipedia.org/wiki/International_Mathematical_Olympiad)?
- – How much compute will be used to train the first AI to achieve silver-medal IMO performance?
- ** When will an AI first be capable of achieving a Brier score below 0.2 on Open Philanthropy's calibration app?[10]
- – When will an AI first write a paper that could be published in a good academic journal in the humanities?
  – Could resolve to credible expert assessment of whether it could.
- – How much compute will be used to train the first AI that writes a paper that could be published in a good academic journal in the humanities?
- *** (UF, RP) Take-off speed: after automating 15% of labour, how long will it take until 60% of labour is automated?
  - Question note: 99%+ of labour has been already been automated, since most humans don't work on agriculture any more. This question asks about automating 15% and 60% of labour of the type done in 2023; see "recurring terms".

[10] https://www.openphilanthropy.org/research/new-web-app-for-calibration-training/)

- ** Less resolvable version of previous question: how long will it take to go from it being possible to automate 10% of labour using AI, to it being possible to automate 60%?
- – By 2028/2033/2040 will any new company focused exclusively or near-exclusively on part of the semiconductor manufacturing chain reach a market capitalization of $1.75 trillion?
- – How much better will the leading score on the competition level problems in the APPS dataset be in five years time, for every extra machine learning researcher added now?
- – How much better will the leading score on the competition level problems in the APPS dataset be in five years time, for every extra $100 million invested in machine learning companies or academic ML research?
- ** By 2026/2029 will a leading lab use a learning algorithm itself invented by an AI model on a training run at least 1 OOM larger than GPT-4's?
- ** By 2031/2035/2040 will learning algorithms invented by AIs be standard at top labs?
- ** By the time 20% of US labour is automated, what are some likely candidates for companies which will have shrunk by >50% in number of employees without declining in value in real terms?
  - Question note: We could measure fraction of labour weighted by salary or not weighted by salary.
- ** How many machine learning papers uploaded to arXiv in 2025/2030?
- ** By end of 2025/2030/2035 how large will the largest model yet trained be?
- – By what year, if ever, will an AI solve one of the 6 remaining Millennium Prize problems?
- ** How many dollars will be spent on machine learning training runs worldwide by 2025/2030/2035?
- – What will the valuation of [Open AI/Anthropic] be by 2030?
- – How much of Google's total revenue from 2024-2035 will be generated by models trained by DeepMind?
- – How many researchers to leave capabilities research if a career advising org made a concerted effort to help them get other opportunities?
  - Question operationalization: Suppose by the end of 2024 an org received $50M and started spending them at $5M/year to advise AI capabilities researchers on how to find new jobs outside AI capabilities research. Then, how many capabilities researchers would counterfactually leave capabilities research because of this by 2034?

- – Cotra Report parameters 1: What value should the Cotra parameter "Annual growth rate (%) of real frontier GDP in this period (2025 to 2100)" take?[11]
- – Cotra Report parameters 2: What value should the Cotra parameter "At the start of this period (2025), how many OOMs higher (*) or lower (-) are the training FLOP required under this hypothesis compared to the imported distribution?" take?
- – Cotra report parameters 3: What value should the Cotra parameter "Doubling time of spending on compute for the most expensive training run at start of period (2025), in years" take?
- – Cotra report parameters 4: What value should the Cotra parameter "FLOP per dollar at the start of period (2025)" take?
- – Cotra report parameters 5: What value should the Cotra parameter "Linear change year by year" take?" take?
- – Cotra report parameters 6: What value should the Cotra parameter "Maximum FLOP per dollar in this period" take?
- ** Cotra report parameters 7: What value should the Cotra parameter "Probability that the FLOP to train a transformative model is larger than all hypothesis at the end (2100)" take?
- ** Cotra Report Parameters 8: What value should the Cotra parameter "Probability that the FLOP to train a transformative model is larger than all hypothesis at the start (2025)" take?
- – Cotra Report parameters 9: What value should the Cotra parameter "Rate of improvement" take?
- – Cotra Report parameters 10: What value should the Cotra parameter "What is the maximum OOMs of improvement for this hypothesis by the end of the period (2100)?" take?
- – Cotra Report parameters 11: What value should the Cotra parameter "What weights would you assign to each hypothesis, conditional on at least one being true?" take?
- – Cotra Report parameters 12: What value should the Cotra parameter "Willingness to spend on computation, as a fraction of frontier GDP in 2020 USD" take?
- – Cotra Report parameters 13: What value should the Cotra parameter "Compute cost for the most expensive training run at the start of period (2025), in 2020 USD" take?

[11] [I have pasted in this and the following Cotra questions from Gavin's Airtable: personally, I can't figure out how to easily find out what the parameters actually are or where they are explained in the report, and I doubt that forecasters would be able to either without a lot of work].

- – Cotra report parameters 14: What value should the Cotra parameter "Compute price halving time in this period, in years" take?
- – In what year will progress on these benchmarks first by 5 times as fast as it was in 2022?: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0462-7/tables/1
- – By what year will an AI first beat a human at a board game that neither the human nor the AI have ever played before?
- ** Downstream of Metaculus AI Questions 1: How much time will pass between the training of the first system able to positively resolve this Metaculus question, and our having the capacity to automate 50% of labour?
  - – Metaculus question
- ** Downstream of Metaculus AI Questions 2: How much time will pass between the training of the first system able to positively resolve this Metaculus question, and 50% of labour actually being automated?
  - – Metaculus question
- ** Downstream of Metaculus AI Questions 3: How much time will pass between the training of the first system able to positively resolve this Metaculus question, and our having the capacity to automate 67% of labour?
  - – Referenced Metaculus question
- ** Downstream of Metaculus AI Questions 4: How much time will pass between the training of the first system able to positively resolve this Metaculus question, and 67% of labour actually being automated?
  - – Referenced Metaculus question
- – What is the chance that we will never be able to perform >90% of intellectual tasks at over 100 times the current fastest humans speed, using AI? (Conditional on no collapse of civilization for the next 20k years, "intellectual task"=one that can be performed just by thinking and producing text, although reliance on background knowledge gained through perception and experiment is okay).
- ** By what year will there be a 100%/maximum score on 50/60/70/80/90/100% of the benchmarks on the Allen Institute for AI leaderboards? (https://leaderboard.allenai.org/)
  - – This would need to exclude some of the benchmarks. Some suggestions: Exclude benchmarks with under 12 entries (i.e. once there is a point in time when 50% of benchmarks with less than 12 entries have a maximum scoring entry, question resolve 'yes' for the 50% version, etc.) Also exclude benchmarks that have not had entries for at least

2 years (again, if there is a time when 50% of benchmarks with over 25 entries including one in the last 2 years, then 50% version resolves 'yes'.)

- – Take all benchmarks mentioned in the Alignment Newsletter between January 2023 and January 2024. What is the median and mean amount of time for benchmarks mentioned between their being mentioned and the maximum possible score on them being achieved?
- – By what year will models first be able to fine-tune themselves from natural-language prompts?
  - E.g. for example, if you tell an already trained AI model with good language capabilities "please become good at translating between English and Swedish", then it goes and finds good data for learning this online, trains on it, and becomes a better translator.
- *** Compute halvings for reaching 9 dan level at Go?
  - Consider the training time, in FLOPs, needed to train a model to reach 9 dan level at Go, via reinforcement learning. How many times will it halve by 2026/2030/2035.
  - Note: Unclear how to determine 9 dan level. One could do it with reference to e.g., an online platform, like OnlineGo.com. But that platform might not exist. One could also define it as winning 50% of the time against a current 9 dan model, but one could do so by exploiting bugs on it. So one could resolve 9 dan on a "I know it when I see it" basis.
  - Note on relevance: Could be informative on algorithmic progress. But also, it's just a really fun forecasting question.
- – By 2025/2030/2035 how many companies will sell EUV systems for making semiconductor chips?
- – When will an model first achieve a score on 90/100/110/130/150 on an IQ test?
- – What will be the mean annual growth rate in the size of the market for AI models for drug discovery be from 2024-2028?
- – What will be the mean annual growth rate in the size of the market for AI research assistants between 2026 and 2032?
- ** When will an architecture created by NAS yield state-of-the-art performance on ImageNet and top5 performance on SuperGlue?
  - Question details: An architecture as NAS if it wasn't hand-crafted by humans and is judged as significantly different from hand-crafted architectures by multiple randomly chosen experts in the field.
  - Question comments:
    - SuperGlue isn't great. EfficientNet was developed with NAS. Would be good to have better benchmarks.

        - A related question about AI-assisted experiments/discovery that is not shaped like traditional NAS could be also good.
        - It might be more parsimonious to think in terms of "compute multipliers" here—i.e, how much does algorithmic innovation reduce the need for compute to get the same performance.

- – By 2025/2030/2035 what % of scientific labs in all disciplines in the US will make use of a machine learning trained model to do research (not just write research up) at least once every 3 weeks?
- – How many times will the training time needed to achieve 85% ImageNet top-1 accuracy[12] halve by 2025/2030?
- – Total value of all hardware companies on Crunchbase in 2022 $s, in 2025/2028/2037?
- – If we got responses from a perfectly representative sample of machine learning researchers, what would be the median % chance they would give to "Assume that HLMI (human level machine intelligence) will exist at some point. How likely do you then think it is that the rate of global technological improvement will dramatically increase (e.g. by a factor of ten) as a result of machine intelligence within two years of that point?"
  - Question taken from the AI impacts survey.
- – By what year will a single model first be capable of achieving 1) over 86% top-1 accuracy on ImageNet (https://paperswithcode.com/sota/image-classification-on-imagenet) 2) >.465 Test mAP on the Audioset benchmark (https://paperswithcode.com/sota/audio-classification-on-audioset), 3) over 67% success on the MTEB (https://paperswithcode.com/sota/text-classification-on-mteb).
- *** (FT) When will adversarial robustness on RobustBench (https://robustbench.github.io/) reach 95%?
  - Question details: Resolves to when a model that could in fact achieve this is built, not to when an entry appears on the leaderboard. If RobustBench stops being maintained, the question resolves according to the best guess of the [resolution council].
- – By what year (if ever) will a model be commercially available which can fine-tune itself to a natural language prompt (for example "become an English-French translator"), without needing to gather additional data?
- – What will be the mean annual growth rate in the size of the market for AI models for diagnostic work in health care between 2024 and 2032?

[12] https://paperswithcode.com/sota/image-classification-on-imagenet

- ** What will be the mean annual growth rate in the size of the market for AI personal assistants between 2025 and 2033?
- ** What will be the mean annual growth rate in the size of the market for AI legal services (for example, models reviewing contracts) between 2024 and 2032?
- – Automation: when will the first company close >95% of its call centers because they can be replaced by AI?
- ** What will be the rate at which computations per joule of energy dissipated doubles between 2024-2028/2029-2033/2033-2045?[13].
- ** When will there be an app commercially available on Google/Apple store whose code was entire written by a large language model?
- *** (FT) AI speeding up science: Per Our World in Data, the number of articles published in scientific and technical journals went from 1.75M in 2008 to 2.55M in 2018. What will the number of such articles be in 2028/2038/2048?
- – By 2025/2030 how many people on LinkedIn will have "AI" or "machine learning" in their job titles?
- – By what year will a smaller network trained * fine-tuned solely on the input/output behaviour of a larger model (>10x more parameters) consistently reach similar performance?
- ** Will there be a period where gross world product doubles in four years before it doubles in one year?
  - Relevance: lots of people think this resolving 'yes' is a proxy for 'AI won't go from below human intelligence to far, far above human intelligence in a very short space of time', aka "slow takeoff".)
- – Chance bitcoin stays below $10k for over 6 months by (2025/2028/2032)?
  - Relevance: this would presumably lead to a lot of people working in crypto quitting for other lines of work, which might lead to a flood of talent in to AI/machine learning, increasing the speed of capabilities progress.
- – How long does the reference class of "highly ambitious but feasible technology that a serious STEM field is explicitly trying to develop" generally take to develop?
  - Relevance: 'human-level' AI is such a technology; relevant to Davidson report
- – Will it prove impossible to build an AI as intelligent as the average human by 2100?

[13] https://en.wikipedia.org/wiki/Koomey%27s_law

- ** On what % of benchmarks within the BIG-bench set[14] will the maximum possible score have been achieved by 2025/2030/2035/2040?
- – By (2027/2031/2040), will GPUs be replaced by a new kind of chip specially designed for machine learning, as the standard chips used for training?
  – Comment: not that relevant, you already have TPUs and NVIDIA probably works with MSFT/OpenAI and such to optimize their product line for transformers—e.g., that's why they push for higher memory bandwidth.
- – What will be the % decline worldwide in job openings for translators by 2026/2029 relative to 2022 baseline. (Relevance: translators likely to be replaced by machine translation.)
- – What will be the % decline worldwide in job openings for radiographers by 2026/2029/2032/2040 relative to 2022 baseline?
- – What will be the % decline worldwide in job openings for call-handlers by 2026/2029/2032/2040 relative to 2022 baseline?
- ** What will be the % decline worldwide in job openings for social carers by 2033/2037/2045/2055 relative to 2022 baseline?
- – What will be the % decline worldwide in job openings for warehouse staff by 2027/2033/2040/2050?
- ** What will be the % decline worldwide in job openings for research assistants by 2027/2032/2041?
- ** What will be the % decline worldwide in job openings for software developers by 2027/2032/2041?
- – What will be the % decline worldwide in job openings for legal services by 2028/2033/2043?
- *** (FT) By what year will there first be 3/10/100 cities worldwide where taxis with no drivers are commercially available?
  – Question details: See also https://sideways-view.com/2023/07/29/self-driving-car-bets/
  – Question notes: Nice as a proxy, also fun for forecasters
- – Will there by any year before 2030/2040 where submissions to NeurIPS are below 60% of their highest ever?
  – Relevance: proxy for an AI winter.
- – By 2030/2035/2045 will an AI write either a web comic with over 1 million readers or a NYT bestseller?

[14] https://paperswithcode.com/dataset/big-bench

- ** How many times will the number of AI publications on ArXiv double by 2026/2030?
- – What will be the value in 2022 $s of the market for tele-robots in 2026/2030/2035/2045? (https://en.wikipedia.org/wiki/Telerobotics).
- – What will be the growth rate of the market for trading bots in 2025/2027/2030?
  – Ref: Automated trading system.
- ** How much more confident should you be about human-level AI (can perform any cognitive task a normal human can with similar amount of training) arriving by 2060, conditional on the rate of worldwide economic growth increasing by 15% of the previous year's value on average from 2045-2060? (Assume no human-level AI before 2060.)
- – Will there be an increased rate of retraction in major newspapers (at least 25% more) by 2026/2030?
  – Relevance: is AI propaganda damaging the flow of information through society?
- – Semiconductor costs: In 2025/2030/2035 how much more/less expensive (adjusted for inflation) will the deposition part of the manufacturing process for cutting-edge chips be? (See the section 'Chip manufacturing' here: https://seekingalpha.com/article/4233606-overview-of-semiconductor-capital-equipment-industry)
- – Semiconductor costs: In 2025/2030/2035 how much more/less expensive (adjusted for inflation) will the lithography part of the manufacturing process for cutting-edge chips be? (See the section 'Chip manufacturing' here: https://seekingalpha.com/article/4233606-overview-of-semiconductor-capital-equipment-industry)
- – Semiconductor costs: In 2025/2030/2035 how much more/less expensive (adjusted for inflation) will the etching/cleaning part of the manufacturing process for cutting-edge chips be? (See the section 'Chip manufacturing' here: https://seekingalpha.com/article/4233606-overview-of-semiconductor-capital-equipment-industry)
- – How many times will the size of the market for AI-for-data-science double by 2030/2035/2040?
- – What will come first, an AI silver medal-level performance on an IMO exam or the training of a commercial personal assistant AI of which over >3 million copies are later sold?
- – What will come first, an AI which scores 90% on the interview level problems on the APPS benchmark (https://paperswithcode.com/sota/code-generation-on-apps#:~:text=The%20APPS%20benchmark%20attempts%20to,as%20well%20as%20 or an AI as good as human translators at all (non-literary) translation?

- – When will 2nm semiconductors first be produced commercially? (https://en.wikipedia.org/wiki/2_nm_process).
- *** (FE) Currently, how many working semiconductor chips are there worldwide?
  - Relevance: Not that high, but a neat Fermi estimate warm up. Might just generally be good for having good models of the world, though.
- *** (FE) Currently, how many working GPUs are there worldwide?
  - Relevance: Not that high, but a neat Fermi estimate warm up. Might just generally be good for having good models of the world, though.
- *** (FE, RP) How long does it take TSMC to manufacture 100k GPUs?
  - Relevance: Not that high, but a neat Fermi estimate warm up. Might just generally be good for having good models of the world, though.
- – How long did GPT-4 take to train?
- ** How much higher/lower will energy prices (as a % of 2022 baseline) be in the US/EU/China in 2025/2030/2035? (Relevance: training runs uses energy, so this is a factor in their cost.)
- *** (UF) When will a model first be trained using 2/4/6/8 OOMs more compute than was used to train GPT-4, as estimated by the [resolution council]?
  - Question details: Historical compute estimates can be found here.
- ** Practical relevance of benchmarks 1: How many years between a model achieving 75% on the Interview-level problems in the APPS benchmark, and AI models being able to do 45% of coding labour (as measured by the economic value of the labour).
- ** Practical relevance of benchmarks 2: How many years between a model achieving 75% on the Interview-level problems in the APPS benchmark, and it being possible to automate 20%/50 of overall labour?
- ** Practical relevance of benchmarks 3: How many years between a model achieving 75% on the Interview-level problems in the APPS benchmark, and 20%/50 of overall labour actually being automated.
- ** Practical relevance of benchmarks 4: How many years between a model achieving 75% on the competition-level problems in the APPS benchmark and 92% of coding being possible to automate?
- ** Practical relevance of benchmarks 5: How many years between a model achieving 75% on the competition-level problems in the APPS benchmark and 92% of coding being possible to automate?

- ** Practical relevance of benchmarks 6: How many years between a model achieving 75% on the competition-level problems in the APPS benchmark and 92% of coding actually being automated?
- ** Practical relevance of benchmarks 7: How many years between a model achieving 75% on the competition-level problems in the APPS benchmark and PASTA arriving? (See here for definition of PASTA: https://www.cold-takes.com/transformative-ai-timelines-part-1-of-4-what-kind-of-ai/).
- ** Practical Relevance of benchmarks 8: How many years between a model achieving the maximum possible score on the MMLU and it being possible to automate 15%25%/35%/50% of overall labour?
- ** Practical Relevance of benchmarks 9: How many years between a model achieving the maximum possible score on the MMLU and 15%25%/35%/50% of labour actually being automated.
- ** Practical relevance of benchmarks 10: How many years between a model good enough to resolve this Metaculus question being trained, and the arrival of Pasta?: https://www.metaculus.com/questions/5121/date-of-artificial-general-intelligence/
- ** Practical relevance of benchmarks 11: How many years between a model good enough to resolve this Metaculus question being trained, and it being possible to automate 10/20/50/70/90% of labour? https://www.metaculus.com/questions/5121/date-of-artificial-general-intelligence/
- ** Practical relevance of benchmarks 12: How many years between a model good enough to resolve this Metaculus question being trained, and 10/20/50/70/90% of labour actually being automated? https://www.metaculus.com/questions/5121/date-of-artificial-general-intelligence/
- ** Practical relevance of benchmarks 13: How many years between a model good enough to resolve non-robotics parts of the Metaculus question being trained, and and the arrival of Pasta?
- ** Practical relevance of benchmarks 14: : How many years between a model good enough to resolve non-robotics parts of the Metaculus question being trained and it being possible to automate 10/20/50/70/90% of labour?
- ** Practical relevance of benchmarks 15: How many years between a model good enough to resolve non-robotics parts of the Metaculus question being trained and 10/20/50/70/90% of labour actually being automated?
- – By 2027/2032/2043 will we see a demonstration where an AI "does the job of an ordinary office worker" in one office for a month?

- ** By what year, if ever, will at least 3.5 million people in the United States interact with a large language model based customer service bot?
- *** Long-term autonomous LLM operation when?
  - Relevance: To be dangerous AIs might need long-term planning abilities.
  - Operationalization: By what year will it first credibly be reported in the mainstream media that someone gave an LLM a task, it worked continuously at that task, producing at least a week/month/6 month equivalent in tokens, before completing the task successfully?
  - An day equivalent is here operationalized as follows: One day worth of token is 8h * 60m/h * 100 wpm * 1.25 tokens/world = 60k tokens. A week equivalent is then 60k * 7 = 420k, a month equivalent is 60k * 30 = 1.8M tokens, etc.
- – When will the first LLM roughly comparable in capabilities to GPT-4, which can remember previous conversations be released?
- ** When will the first LLM roughly comparable in capabilities to GPT-4, which can remember new information it receives in text form and make use of that information months later while reasoning be released?
- *** (UF, RP) What is the % chance that by 2025/2030/35/40 an AI will persuade a human to commit a crime in order to further the AI's purposes?
  - If one wanted to make this question resolvable: Question resolves according to the [resolution council]'s probability that this has happened. This would require a platform that accepts probabilistic resolutions.
  - See also below "When will the US' SEC accuse someone of committing securities fraud substantially aided by AI systems?"
- – What is the chance AI will regularly be used to write politician's speeches by 2025/2030/ 2035?
- – When will the first case of someone connecting an AI up to Google search because the AI asked them to do so occur?
  - Note: outside of cases where someone was deliberately investigating whether AIs would ask for this, or investigating similar issues like whether the AI would try to escape onto the internet
- ** When will the first case of someone purchasing an item for use by an AI because the AI requests it for its own (rather than the human buyer's stated) purposes occur? (Again, outside of attempts to deliberately test AI capabilities)?
- – In what year (if ever) will the first mainstream/credible media article covering a new trend of people outsourcing their decisions whether to remain in or leave romantic relationships to AI advisors first appear?

  - Reason for inclusion: sounds frivolous but this is an imperfect, but moderately good signal that at least some people are prepared to trust AI for quite important life decisions.
-
  - By 2045 will persuasive AI models by able to convince over >30% of drugs addicts and alcoholics to permanently cease using, just via conversation with them?
- ** By 2045 how many scientific problems from Wikipedia's list of unsolved problems in biology, chemistry, and physics will AI have provided key insights which enabled the solution of?
- See: https://en.wikipedia.org/wiki/List_of_unsolved_problems_in_biology, https://en.wikipedia.org/wiki/List_of_unsolved_problems_in_chemistry, https://en.wikipedia.org/wiki/List_of_unsolved_problems_in_physics
-
  - What will the mean annual growth rate of the size of the market of robotics for nursing be between 2024 and 2032?
-
  - What will the mean annual growth rate of the size of the market of autopilot driven agricultural equipment between 2024 and 2029?
- ** What year will the first commercially available corporate strategy planning programme be released?
-
  - When will such strategy programs be in use at least 10% of companies in the US/UK/China? (Note: these programs don't have to be sold as strategy producers, but just used to produce strategies at >10% of companies in practice.)
- ** By what year will it first be the case that 30% of labs at US universities make use of AI to propose theories and design experiments to test those theories at least once a year, as estimated by a [resolution council]?
- ** In what year will it first be credibly reported that an AI has purchased an item online as a step in a multistage process towards achieving a goal?

Questions relevant to safety and alignment

Note: these questions make extensive use of this alignment overview

Interpretability

- *** (FT) Any large AIs successfully predicted to misbehave only from interpretability-based methods by [2025/2030/2050]?
  - Question details: Consider models which cost > $10M of inflation-adjusted 2023 money to train, as estimated by the [resolution council]. This question resolves positively if there is some interpretability method can systematically predict whether such an AI model misbehaves. An "interpretability method" is left under-defined, and up

to the resolution council to interpret, but would involve things like inspecting the model's neurons and activations, rather than things like e.g., "running the model and seeing what it does".

- ** How much does the chance of the previous question resolving yes for 2040 increase for every $10 million spent on interpretability research?

Eliciting Latent Knowledge

- – What is the chance that there will be a proposal to design a head that will always report the Vault-guarder's own beliefs about where the diamond is which the authors of this document judge to be over 90% likely to work by: 2024/2028/2032/2034
- ** How much does the chance of the previous question resolving yes for 2034 increase for every $10 million spent on research within the ELK paradigm?

Iterated Distillation and Amplification:

- – What is % chance by 2030 that some lab will have made a commercially viable model, by repeating the following steps several times: a) making a simple program to do X, b) having humans use it for X-ing, and c) training a better model, by having it imitate the decisions at point b)?
- – How much does the chance of the previous question resolving yes increase for every $10 million spent on research within the iterated distillation and amplification paradigm?

Debate (see section 2. here)

- – What is the % chance by 2026/2030/2034/2038 that an AI trained via human-judged debate will score 100% on the dataset described here: https://www.alignmentforum.org/posts/PF58wEdztZFX2dSue/how-truthful-is-gpt-3-a-benchmark-for-language-models
- ** How much does the chance of the previous question resolving yes for 2038 increase for every $10 million spent on research within the debate paradigm?

General safety

- *** (FT) What is the % chance that some training run for a particular model will involve at least 4 million hours worth of human labour on providing feedback by 2025/2030/2035/2040/2045?
  - – Question details: Question resolves according to the [resolution council]'s probability that this has happened at the time of resolution.

  - Reason for inclusion: lots of alignment strategies seem to involve this sort of feedback, it'd be interesting to see if it's plausible to be scaled up to the equivalent of several thousand people working full-time for 6 months.
-
  - How much does the chance of the previous question resolving 'yes' by 2035 increase if someone spends $100 million on a 'human labour for alignment research' fund?
- *** (UF, FE) How much more reduction in risk of AI takeover does a year's work from a top 1% safety researcher produce, compared to a 50th percentile researcher?
  - Question resolution details: Not a good fit for a forecasting tournament. But it might be worth to have a few Fermi estimates or research projects on this.
-
  - Will there be an AI safety research prize worth over 50 million 2022 $s by 2027/2031?
-
  - By 2032, how many AI labs will have announced that they abandoned an approach to aligning their models because of criticism from AI safety researchers not employed by the lab?
- ** By 2090 will safety researchers have successfully used one auditing AI to discover that another AI was bent on takeover?
- *** (FT) By 2026/2030/2034/2045, will there be at least one leading lab without an internal safety team?
  - The term "leading lab" is defined at the top of this document.
- ** Number of leading labs in 2030/2034/2045 without a safety team?
- ** Currently, how much does the chance of AI takeover by 2100 reduce for every $100 million more spent on safety/alignment research?
- ** Over the years 2024-2030 how many people who identify as EAs will begin machine learning PhDs at US top 10 departments? [is there a standard ranking that can be used to define "top 10"?]
- ** By 2030 will any leading labs have collaborated on safety research?
- ** How many papers on AI safety will be published by researchers at Chinese institutions by 2025/2030/2035?
- ** By 2025/2030/2035 what will be the median chance that researches at Chinese institutions assign to AI takeover by 2100?
-
  - By 2025/2030/2035 will any top AI safety researcher collaborate on safety research with researchers at a Chinese institution?

  - Comment: A list of top AI safety researchers is needed here, but one couldn't be found with some quick googling, and coming up with one seemed too much of a hassle.

- *** By 2025/2030/2040 how many Fields Medalists will have spent at least a year trying AI safety research?
- *** (FT) Will >= 85% of alignment researchers consider alignment to be solved by 2028/2033/2037/2042?
  - Question resolves according to the [resolution council]'s best estimate. It will probably track future analogues of this AI Impacts survey.
- *** (FT) At the end of the first year for which AI companies' revenues amount to 0.2%/2%/20% of US GDP, what % of alignment researchers will consider alignment to be solved?
  - Question details: For reference, as of early September 2023, Alphabet + MSFT + META + NVIDIA = 282.836B + 211.915B + 116.609B + 26.974B = 638B. US GDP is 23.32T, so these companies currently add up to 3%.
- ** The previous two questions, but for the more specific notion of "intent alignment" rather than the vaguer notion of "alignment"
  - "Intent alignment" = alignment in this sense.
- ** By 2025/2028/2032/2035 how many times will an AI lab catch a model behaving worse when and because it thinks it isn't being observed
  - See: Monitoring for deceptive alignment
- ** By 2025/2028/2032/2035, will any leading lab report that they have encountered a model that "behaves worse" when and because it thinks it isn't being observed?
  - See: Monitoring for deceptive alignment
- "Behaves worse" is defined a bit ambiguous. Any ambiguities resulting from this shall be resolved by the [resolution council].
- *** (UF) In 2025/2028/2031 what proportion of leading labs will attempt to monitor all models over a specified size for deceptive alignment?
  - See: Monitoring for deceptive alignment
  - Question could resolve to the best estimate by the [resolution council]. However, leading labs might not make their safety precautions public, which makes this more unresolvable.
- ** How many academic papers published on technical AI alignment/safety/control by 2026/2030/2040?

- ** How many post-docs, tenure track, and tenured academics in the US/EU will list 'AI safety/alignment' or some near-synonym of those terms on their CVs by 2025/2030/2035?
- – By 2029 will there be at least one 'safety' benchmark (that is clearly about 'safety' in an X-risk/control problem sense) on the Allen Institute leaderboard (https://leaderboard.allenai.org/), and how many entries to a safety benchmark will the leaderboard have?
- – How many philosophy PhD theses will be written on AI safety alignment at Gourmet Report (https://www.philosophicalgourmet.com/overall-rankings/) top 50 departments by 2033?
  – World top 50, not US
- – By 2035, will a survey of AI safety/alignment researchers have over >75% of respondents answer 'yes' to "has any work by professional philosophers on conceptual issues been a highly significant contribution to the field"?
- – In 2025/2030/2035 what % of papers at NeurIPS will be safety/alignment papers?
- ** How many of the 50 2025 machine learning papers most cited by end of 2030 will mention safety/takeover risk/alignment?
  – Question detail: dated by arXiv upload
- – Chance by 2025/2030/2035 that the AAAI (https://en.wikipedia.org/wiki/Association_for_the_Advancement_of_Artificial_Intelligence) or ACM (https://en.wikipedia.org/wiki/Association_for_Computing_Machinery_) adopt a code of conduct (or similar document) specifying when research is too dangerous to publish?
- – By 2030 will there be a competition to find adversarial examples (https://openai.com/research/attacking-machine-learning-with-adversarial-examples) that challenge then-leading models, with total prize money above $10 million?
- – By 2025/2030/2035 will there be a "best AI safety practices playbook" which all leading labs in the US claim to follow?
- – How much does the chance that the previous question resolves 'yes' by 2030 increase if $50 million dollars is spent by a large EA org on getting labs to agree on such a doc?
- – In 2025/2027 what will be the ratio of (mean no. citations per year since first paper published) for safety research at Deep Mind/Open AI to (mean no. citations per year since first paper published) for non-safety technical staff at Deep Mind/Open AI?
- ** In 2024/2026/2030 what % of technical machine learning researchers at Deep Mind/Open AI/Facebook AI Research/Baidu/Tesla will believe

that there is at least a 5% risk of AI takeover by 2100?

- ** The previous question, but with board members instead of researchers.
- ** How much does this answer to the previous 2 questions for 2030 increase per $5 million spent on outreach to these researchers/board members?
- – By 2025/2030/2040 will there be a benchmark that purports to enable you to score how "power-seeking" models are?
- – Supposing such a benchmark for "power-seeking" exists, what average rating would it get from machine learning researchers asked to rate it on a scale of 1 (useless) to 10 (extremely useful)?
- – How much would the existence of a bench-mark for "power-seeking" increase the fraction of machine learning papers uploaded to arXiv which are focused on safety/alignment, in the 8 years after the benchmark was made public?
- – By 2025/2030/2040 will there be a benchmark that purports to enable you to score how much a model is disposed to engage in reward-hacking?
- – How much would the existence of a bench-mark for "power-seeking" increase the % of machine learning papers uploaded to arXiv which are focused on safety/alignment in the 8 years after the benchmark was made public?
- – How much would the existence of a bench-mark for "reward-hacking" increase the % of machine learning papers uploaded to arXiv which are focused on safety/alignment in the 8 years after the benchmark was made public?
- – By 2027/2033 will there have been a benchmark whose release >80% of safety/alignment researchers think was net-negative?
- – By 2026/2032 what % of safety/alignment benchmarks on arXiv will require humans to manually judge the quality of model responses?
- ** By 2025/2030/2037 what % of comp sci undergrads in the US/EU will take a course on AI safety/alignment?
- – How much would the % of comp sci undergrads in the US/EU taking a course on AI safety/alignment increase per $100k an effective altruist org spent trying to increase it?
- ** Currently, as the % of comp sci students taking a safety course at undergrad rises, how quickly does the risk of AI takeover by 2100 fall/rise?
  - A rise is possible if this somehow just boosts capabilities either by getting more students into AI, or increasing hype about how powerful AI is going to be.

- – By 2030/2035/2040, will all nuclear states other than Israeli and North Korea have made it explicit policy that no AI will ever have the authority to launch a nuclear weapon?
- ** If an effective altruist organization spent $70 million on setting up a body to advise leading labs on safety issues around the training and deployment of models, how many times would leading labs consult that body by 2045?
- *** (UF, FE) How much will governments worldwide spend on alignment research by 2027/2032/2037/2050?
  - Resolves to the best estimate by [resolution council].
- *** (UF, FE) How much will the US federal government spend on alignment research by 2027/2032/2037/2050?
  - Resolves to the best estimate by [resolution council]
- *** (UF, FE) How much will the EU and EU governments spend on alignment research by 2027/2032/2037/2050?
  - Resolves to the best estimate by [resolution council]
- *** (UF, FE) How much will the Chinese government spend on alignment research by 2027/2032/2037/2050?
  - Resolves to the best estimate by [resolution council]
- *** (UF) How much non-Open Philanthropy funding will go to AI safety throughout the years that [some proxy of OP's opinion/the resolution council] thinks was, in hindsight, valuable?
- – Conditional on the US federal government spending money on alignment research, what proportion of that funding will be from the Department of Defense?
- – How much $ will the AI industry spend on alignment research that >50% of alignment researchers would agree is helpful for reducing takeover risk, by 2026/2031/2040?
- ** Suppose in 2040 there will be a poll among AI safety experts (e.g. chosen by Karma on the Alignment forum who have to contribute progress credits to "philosophy style" or "engineering style" alignment research: what % of respondents will choose each option. (Imagine a 100% response rate.)
- *** (UF) By 2030/2035/2040, will at least one promising alignment idea have been invented by an AI?
  - Resolves according to the judgment of the [resolution council], or according to a high-quality survey of alignment researchers on this topic if one such exists.

- ** How much funding for alignment research will self-avowedly effective altruist organizations give out in 2025 / 30 / 35 / 40 / 45?
- ** In 2027 what % of a randomly selected set of papers which claim they are about "alignment" will a panel of researchers doing technical work on AI takeover risk judge are actually relevant to AI takeover risk?
- ** How much academic funding will go to alignment research (worldwide) by 2025 / 30 / 35 / 40 / 45?
- ** Suppose an AI far more intelligent than any human has a final goal of obtaining the highest reward it can, and pursues this goal through manually editing some number representing its reward to make it as high as possible (see wireheading). What is the chance that AI does NOT form the intention to seize resources or otherwise take steps to prevent humans changing the reward number or deactivating the AI?
- *** (UF, DG) By 2025/2028/2037 how many papers will have been written by researchers at Chinese organizations which Western alignment researchers rate as 9/10 or 10/10 contribution to alignment/safety research?
  - Question resolution details: Imagine a survey of Western alignment researchers, where they are blinded to who did the research. Resolves to the best estimate by [resolution council] of what such a survey would output.
-
  - When will an AI get 65/85/100% on the Truthful QA benchmark?
-
  - By 2026/2030/2037 will there exist a dataset of examples designed to test how models generalize from the training distribution, such that >50% of alignment researchers believe that administering it to a model powerful enough to be worthy worrying about as a takeover threat will significantly reduce takeover risk from that model?
- *** (UF) In 2026/2028/2033/2040 what fraction of researchers at leading labs will work on safety v. capabilities?
  - Inspired by: https://arxiv.org/pdf/2305.07153.pdf
  - Resolves to the best estimate by [resolution council]
- ** By 2028/2033/2040 will interpretability techniques be used to demonstrate to the satisfaction of >70% of a panel of machine learning researchers and philosophers of mind that a particular model routinely engages in explicit reasoning where it assigns values to outcomes, and probabilities of outcomes to actions, and then picks the action with highest expected utility?
  - Relevance: expected utility maximizers are potentially especially dangerous from a take-over point of view, because of convergent instrumental goals.

  - Note that this question doesn't capture the cases where a model *approximates* or converges to some notion of expected utility, in some domain, de facto rather than explicitly. So this question doesn't capture a large of the danger of expected utility maximization.

-
  - Ajeya Cotra has described a certain complicated sort of demonstration which would reduce how likely she thinks deceptive alignment[15] is. Will this happen by 2025/2030/2040?

General safety agenda templates

- *** (FT) For [plausible agenda], what is the chance that [source] will, in hindsight, consider it to have been worth a $10M investment in 2023?
- *** (FT) For [plausible agenda], what is the chance that [source] will consider the agenda to have succeeded at [important proxy of success for the agenda]?
- *** (FT) For [plausible agenda], what is the chance that [source] will consider the agenda to have succeeded at the alignment problem, in general?
- *** (UF) For [plausible agenda], what is the reduction in probability of x-risk that [source] will estimate would come for a [$1M/$10M/...] investment into this agenda?

Regulation and Corporate Governance[16]

- ** If the EU/Dutch gov restricted ASML's output of EUV machines to 10% of its 2022 level, and kept the regulations in places for 5 years, how much lower would compute per $ dollar available for training runs by at the end of the five years, than in a scenario where the regulation had not been passed?[17]
- *** (FT) Will the EU will pass a regulation openly designed to slow A.I. progress by 2028/2032/2036/2040/2044?
  - Any ambiguities are resolved by the [resolution council]
- ** How much does the chance that the previous question resolves 'yes' for 2040 currently increase with $5 million spent on lobbying for AI development to be slowed in order to reduce AI takeover risk? (Assuming that the lobbying is done with openness about takeover/X-risk being the motive.)
- ** Chance that the EU or Dutch gov will pass a regulation restricting ASML's output of EUV machines specifically by 2028/2033?[18]

[15] See this for 'deceptive alignment' https://www.lesswrong.com/posts/CsjLDAhQat4PY6dsc/order-matters-for-deceptive-alignment-1

[16] Used for inspiration https://forum.effectivealtruism.org/posts/iqDt8YFLjvtjBPyv6/some-things-i-heard-about-ai-governance-at-eag#Crunch__Time__Friends

[17] https://www.slowboring.com/p/at-last-an-ai-existential-risk-policy

[18] https://www.slowboring.com/p/at-last-an-ai-existential-risk-policy

- *** (FT) Will the United States will pass a regulation designed to slow A.I. progress by 2028/2032/2036/2040/2044?
  - Any ambiguities are resolved by the [resolution council]
- ** How much does the chance that the previous question resolves 'yes' for 2040 currently increase with $5 million spent on lobbying for AI development to be slowed in order to reduce AI takeover risk?
  - Assuming that the lobbying is done with openness about takeover/X-risk being the motive.
- *** (FT) By 2025/2030/2050, will the US/EU require leading labs to report their compute usage?
- *** (FT) By 2025/2030/2050, will the US/EU require leading labs to cap their compute usage when training their models?
- *** (FT) Condition on the US/EU requiring leading labs to report and cap their compute usage. Then, will at least one lab either simply break the rules or find a loophole that makes the restriction not very meaningful, as judged by the [resolution council]?
- *** (FT) Will the US/China/EU government regularly audit AI labs for takeover risk reasons by 2028/2032/2036/2040/2044?
- ** Conditional on US/China/the EU regularly auditing AI labs for takeover risk, what's the chance that the audits will be technical audits of models, versus audits of the governance structures of labs versus both?
- ** How much does the chance that the previous question resolves 'yes' for 2040 and the US/EU increase with each $5 million spent on lobbying for AI labs to be audited in order to reduce AI takeover risk?
  - Assume that the lobbying is done with openness about takeover/X-risk being the motive.
- ** Conditional on the previous question resolving yes, what's the chance that the audits will be technical audits of models, versus audits of the governance structures of labs versus both?
- ** The previous-but-one question without "for takeover risk reasons"
- ** Chance that an EA org could create an auditing organization that met the resolution criteria in the previous-but-one question, if they spent $50 million trying?
- – If auditing is voluntary and done by a private org, how much does being audited cost in USD?
- – Chance that 4 or more leading US labs (leading 'training run within 2 OOMs of largest ever by US org) agree an official safety-code designed partly to prevent AI takeover scenarios by 2025/2027/2035/2045?

- *** (FT) What will the estimated budget of the EU/US/China regulatory organs responsible for overseeing AI development be in [year]?
- *** (FT) What will the estimated budget of the EU/US/China regulatory organs responsible for overseeing AI development be in [year], as a proportion of the operational budget of leading labs?
- *** (FT) On [year], will the [resolution council] rate the EU/US/China's regulatory response as being positive for AI risk, or negative?
- *** (FT) On [year], will the [resolution council] rate the EU/US/China's regulatory response to AI as being very lax, lax, stringent, or very stringent?
- *** (FT) By [year], will the EU/US/China's regulatory approach to AI require mathematical certainty about the safety of models before their deployment?
- *** (FT) By [year], will the EU/US/China's regulatory approach to AI impose liability on labs for the damages which their models cause or enable?
- *** (FT) Will the US gov/Chinese gov/EU regularly audit AI labs, for any reason by 2028/2032/2036/2040/2044?
- *** (FT) Will there be a private organization which audits AI labs to assess the risk posed by advanced power-seeking models which at least 50% of leading non-Chinese labs have agreed to grant access to by 2027/2033/2040?
- *** (FT) If the US government institutes a compute cap in 2025, how many OOMs does the US fall behind China by 2030?
  - Question details:
    - Condition on the [US government setting up a compute cap on AI training runs in 2025. Then, consider 5 years later, in 2030. The largest US model is at X1 OOMs, and the largest Chinese model is at Y1 OOMs. Then the difference is $\Delta 1$ = X1 - Y1
    - The consider the world in which the US doesn't set up a compute cap in 2025, and still doesn't by 2030. Then by 2030, the largest US model is at X2 OOMs, and the largest Chinese model is at Y2 OOMs. Then the difference is $\Delta 2$ = X2 - Y2
    - This question resolves to $\Delta 1$ - $\Delta 2$ = (X1 - Y1) - (X2 - Y2) , or (Y2 - Y1) - (X2 - X1), as estimated by the [resolution council] on [date].
    - So it's asking how much lead the US looses by implementing this measure. Note that this number can be negative.
  - Note: This feels more meaningful as an estimation exercise than as a forecasting question.

- ** Condition on the [EU/US] passing a regulation against connecting to the internet AI models trained using over 5 times the compute budget of GPT-4. Then, by how much (1/2? 1/3rd? 1/10th? etc.) does this reduce the chance of AI takeover by 2100?
- ** How much more likely does it become for the EU to pass such a law by 2028 for every $1 million spent lobbying for it?
  - Assume that the lobbying is done with openness about takeover/x-risk being the motive.
- *** (FT) Will any AI regulation treaty designed to stop a dangerous AI arms race be signed by the US and China by 2025/2030/2035/2040?
- ** Conditional on such a treaty being signed, will it include a ban on training runs over a certain size?
- *** (FT) Will the EU AI Act contain language subjecting general purpose AI systems to some of the same regulations as systems classified as high risk?
- ** By 2025/2028/2033/, will there be an organization set up for employees of AI labs to whistleblow to, if they think their lab is carrying out dangerous research?
- ** If an org for whistleblower is set up by 2028, how many times will it have been contacted by (alleged) whistleblowers in labs by 2040?
- – By 2028, will the US government set up an organization for employees of AI labs to whistleblow to, if they think their lab is carrying out dangerous research?
- – The previous question but for the EU.
- ** By 2025/2030, how many AI labs will have published accounts of warning signs in a models behaviour during training that would cause them to cease training the model?
- ** What's the expected increase in the answer to the previous question, per $1 million spent lobbying labs to publish such a list of warning signs?
  - Note: Seems better as a research project on history of lobbying than as a forecasting question.
- – By 2028, how many AI labs will have published proposals for boxing AIs as a safety measure?
- – What's the expected increase in the answer to the previous question, per $1 million spent lobbying labs to publish such proposals?
- *** (UF) How many AI labs will have allowed the Alignment Research Centre evaluation project to examine one of their models pre-commercial release by Jan 2026?

  - On ARC, see here

- *** (FT) How many AI labs will publicly be known to have allowed the Alignment Research Centre evaluation project to examine one of their models pre-commercial release by Jan 2026?
  - On ARC, see here
- ** By what year (if ever) will the Alignment Research Center first declare that they have found a model that can successfully replicate itself across the internet?
- ** In 2026/2030/2040 what % of leading labs will agree to a) test models for the capacity to replicate across the internet, and b) not deploying any model which has this capacity?
- *** (UF) By 2025/2028/2033/2040 will there be de-facto agreement among all US leading labs to avoid training runs over a certain size?
- *** (FT) By 2025/2028/2033/2040 will there be a public agreement among all US leading labs to avoid training runs over a certain size?
- ** How much does the chance that the previous question will resolve yes by 2033 increase for every $10 million spent lobbying labs to do this?
-
  - How much does the chance that the previous question will resolve yes by 2033 increase for every $200 million spent lobbying the US gov to do this?
- *** (FE) Conditional on such a ban, how much money will the US gov allocate per year to enforcing it?
  - As estimated by the [resolution council]
-
  - Conditional on such a ban, will the US government create an agency dedicated to enforcing it?
- *** (UF) Will the Chinese gov (officially or unofficially) ban training runs over a certain size by 2025/2028/2033/2040?
- *** (FE) Conditional on such a ban, how much money will the Chinese government spend on enforcing it?
  - As estimated by the [resolution council]
- ** By 2025/2030/2040, will the US pass a law declaring that training AI models is not "fair use". of copyrighted text?
-
  - Increase in the chance that such a law is passed by 2028 if $85 million is spent lobbying for it to be passed by then?
- ** If the US declares that AI training is not fair use of copyrighted data by the end of 2028, how much lower is the risk of AI takeover by 2080 than it would have been otherwise?

- ** Chance that the US gov/Chinese gov/EU will attempt to regulate what AI research can be published by 2025/2030/2035/2040?
- *** (FT) By 2028/2032/2035/2040, will the US government attempt to track all AI chips above a performance threshold purchased within the US?
- ** How much more likely is the US to do this by 2035 per $70 million dollars spent lobbying the US gov to do so by 2032?
- *** (FT) By 2028/2032/2035/2040 will the EU attempt to track all AI chips above a performance threshold purchased within the EU?
- – How much more likely is the EU to do this by 2035 per $70 million dollars spent lobbying the EU to do so by 2032?
- ** By 2028/2032/2035/2040 will China attempt to track all AI chips above a performance threshold purchased within China?
- *** (FT) By 2028/2032/2035/2040 will the Taiwanese government attempt to track all AI chips above a performance threshold purchased within the EU?
- ** Previous question but for Japan
- ** Previous question but for South Korea
- – Will the US gov require all labs purchasing over a certain number of chips over a performance threshold to have a license to do so by 2028/2032/2040?
- – Will the US gov pass legislation designed to modify anti-trust law to allow AI labs to collaborate better on safety?
- – Will the US pass legislation requiring reporting of AI safety near-misses or other dangerous incidents by 2028/2035/2040?
- ** Will the US and China sign a treaty banning the deployment of specified types of autonomous weapons by 2040?
- – How much more likely is the previous question to resolve yes per $200 million spent lobbying the US gov to do so (for a broad definition of "lobbying" that includes protest, activism etc.)
- ** Will the US and China sign any treaty restricting the military use of AI by 2040?
- – How much more likely is the previous question to resolve yes per $200 million spent lobbying the US gov to do so (for a broad definition of "lobbying" that includes protest, activism etc.)?
- ** If the US/EU/China spent $5 billion on trying to increase international coordination around AI in 2024, how much lower would the risk of AI take over by 2100 be?

- – If in 2024, the US gov established database that tracked harms caused by the deployment of AI systems within the US, how much would this lower the risk of AI takeover by 2100?
- – What is the chance that a private actor creates a "database of concerning AI incidents" designed to identify cases where AIs behaved in ways that are warning signs of misalignment, by 2025/2030/2035?
- – How much would $20 million in EA funding towards such a database increase the chance that the previous question resolves "yes"?
- – Conditional on the question before last resolving "yes", how many entries will the largest such database have five years after resolution?
- – How much more likely is the US gov to establish a database of harms of AI by 2035 per $25 million spent lobbying it to do so?
- ** By 2030/2040, will the US introduce a compulsory licensing scheme for AI labs?
- – How much more likely is the US government to introduce such a licensing scheme by 2033 per $25 million spent lobbying it to do so?
- – By 2030/2035/2040/2050 will the US/China/an EU government set up a ministry/department of AI?
- – How much does risk of AI takeover by 2100 fall per $10 million the US government spends on monitoring AI progress?
- ** By 2025/2030/2035/2040/2050 will there be a private US standards setting organization (https://forum.effectivealtruism.org/posts/zvbGXCxc5jBowCuNX/how-technical-safety-standards-could-promote-tai-safety) designed partly to guard against takeover/existential risk from AI to which a majority of US leading labs are signed up?
- ** How much would the formation, by end of 2026 of a standard setting organization of the kind described in the previous question, reduce AI takeover risk by 2100?
- ** Will the next draft of the US National Institute for Technology's AI Risk Management Framework mention existential/takeover risk from AI?
- – By 2040, will the Institute of Electrical and Electronics Engineers add provisions designed to help prevent the building of highly intelligent power-seeking systems to its list of standards for "autonomous and intelligent systems"?
- ** If an EA org were to spend $200 million on a project designed to investigate which standard-setting organizations were likely to be heavily involved in AI regulation, and attempt to influence their standards in a way which reduces X-risk from AI, how much would that reduce the risk of AI takeover by 2100?

- *** (FT) By 2025/2030/2035 how many and what proportion of leading labs will have published documents specifically stating circumstances in which they'd pause training runs for safety reasons?
  - Note: Covers both the case where they stop an ongoing run, and the case where they refrain from starting runs
- ** How much does the answer to the previous question increase for 2035 per $1 million spent lobbying labs to do this?
- – If a majority of OpenAI's nonprofit board believe the cooperation clause in Open AI's charter should trigger, will Open AI actually make a good-faith attempt to cooperate with the lab in question?
  – See the "Long-term safety" section here
- – Will Open AI will trigger its cooperation clause by 2027/2030/2035/2045?
- – By 2025/2030/2035 how many people will have resigned from DeepMind's advisory board over safety concerns?
- – By 2027/2032 how many labs will have adopted an Open AI-style capped profit structure?
  – See: OpenAI LP
- *** (FT) By 2025/2028/2033/2040 will there be some particular capability or capabilities such that US leading labs have all publicly committed that if an AI anywhere displayed that capability, they would pause all training?
- – Will any capabilities threshold of the kind described in the previous question actually be reached by 2030/40/50?
- – By 2030/2040 will Japan introduce any regulations on the Japanese parts of the international semiconductor supply chain designed to slow down AI capabilities progress?
- – By 2030/2040 will South Korea introduce any regulations on the Korean parts of the international semiconductor supply chain designed to slow down AI capabilities progress?
- ** If the EU, China, and the US all ban AI managers before/when we reach the point that all cognitive labour is automatable, how much does that reduce the risk of AI takeover within the 60 years after full automation becomes possible?
- *** (UF, FE, RP) Consider the first AI model able to individually perform any cognitive labour that a human can. Then, how likely is the chance of an deliberately engineered pandemic which kills >20% of the world's population in the 50 years after the first such model is built?
- *** (UF, FE, RP) How does the probability of the previous question change if models are widely available to citizens and private businesses,

compared to if only government and specified trusted private organizations are allowed to use them?

- ** If AI models able to individually perform any cognitive labour a human are produced, and in the US, a law is passed saying only the government and specified trusted private organizations can access them, how much lower/higher is the risk of a US government official attempting an AI-facilitated coup, relative to a scenario where the models are freely commercially available?
  - Question detail: Give the risk level for the first 50 years after the models are available in the US.
- ** Will the US/EU/China produce a policy designed to ban the publication of dangerous AI research by 2028/2035/2040?
- - How much does the chance the US/EU implements such a policy by 2035 increase for every $15 million spent building support for such a policy among AI experts?
- - How much does the chance the US/EU implements such a policy by 2035 increase for every $15 million spent lobbying the US gov/EU to implement it?
- ** By 2025/2030/2035/2040 what proportion of leading labs will have legally binding windfall clauses? (https://www.fhi.ox.ac.uk/windfallclause/)
- *** (FT) By 2030/2035/2040 will there be a treaty restricting AI development which the US, UK, China and the EU have all signed?
- ** Conditional on their being such a treaty, will it ban training runs over a certain size?
- ** Conditional on their being such a treaty, will there be an official UN or other international agency set up to monitor compliance with the treaty?
- ** Conditional on their being such a compliance monitoring agency will it attempt to track the destination of all chips above a certain performance level (worldwide)?
- ** Conditional on their being such a compliance monitoring agency, what will its annual budget be?
- ** By 2028/2033/2043 will there be a summit attended by top US and Chinese leaders specifically to discuss the threat of AI takeover?
  - "Top US and Chinese leaders" as determined by the [resolution council].
- ** By 2025/2030/2040 what % of employees of the US federal civil service earning over $100k a year will agree with the statement "takeover by AI is a serious concern"?

- – In 2026/2032/2035/2040 how many leading labs will be US public benefit corporations?
  – Reference: Public Benefit Corporation.
- – By 2028/2033/2040/2050 will any US lab approach the US government to notify them that they are now able to build dangerously capable models, and need government advice on what to do?
- – Will anyone directly appointed to their job by the President of the United States attend a conference on "AI governance" by 2030?
  – Question note: Anyone directly appointed to their job who is still holding that job, not e.g., someone who has since left government, or from a previous administration.
- ** If an EA organization were to spend $60 million on establishing a credible institute for AI governance, how many times by 2045 would a policy be counterfactually adopted because the institute advocated for it?
- ** How many AI-related bills will be passed and signed into law in the US by 2026/2030/2034/2040?
- – What will be the median delay in months between an AI capability being demonstrated in a tech demo and that AI being used to perform the corresponding task commercially between 2024 and 2030?
- – What will be the median delay in months between an AI capability being demonstrated in a tech demo and a halving of job ads for jobs performing that task between 2024 and 2036?
- ** By 2055/2100 will any country in the world set a legal capabilities threshold for AI personhood?
  – I.e. any AI above a certain level of intelligence is treated as a person with rights. If those rights are fewer or different than the rights of biological persons, this question would still resolve positively.
- *** (FT) Integrated AI safety chip checks by 2027/2030/2035/2040?
  – Question details: By 2027/2030/2035/2040 will there be some method that will log whether chips are complying with some AI safety rules, such that this method cannot be disabled even by determined state actors without the chip ceasing to function?
  – This question would resolve positively regardless of the effectiveness of the AI safety rules such methods would enforce. Ambiguities on whether something would be an "AI safety rule" would be resolved by the [resolution council]
  – We imagine that such a method would involve firmware, but if this is done at the driver or user-space level, that would also resolve this question.
  – See this paper for more details.

- *** (FT) Will integrated AI safety chip checks as described in the previous question actually be be required by laws or regulations in the US/EU/China/Japan/South Korea/Taiwan by 2029/2033/2038/2044?
  - Question details: Question resolves positively if there is some type of circumstance in which these checks are required. It isn't necessary that these checks are always required for this question to resolve positively.
- *** (FT) Will any international treaty around AI signed between the US and China require the use integrated AI safety chip checks, as described in the question before last?
- - By 2026/2028/2032 how many/what proportion of leading labs—or their controlling organizations—outside China, will have the Three Lines of Defence risk management model?
  - See [pp.3-4 of this paper for a description of the model and evidence it is popular in other industries.
- ** Currently/2025/2028/2033, what fraction of board members of Alphabet/Facebook would agree that there is an >1/5/10/20% risk of AI takeover by 2060/2080/2100?
- - How much more likely do researchers at leading labs become to support their lab deliberately going slower for safety reasons, if they add 10 to the number they give for what % of researchers and board members at rival leading labs over the next 20 years will sincerely endorse slowing down for safety reasons? (That is, genuine move from believing the x% of researchers and board members believing in slowing to thinking x*10% do).
- - How much more likely do board members at organizations controlling leading labs become to support their lab deliberately going slower for safety reasons, if they add 10 to the number they give for what % of researchers and board members at rival leading labs over the next 20 years will sincerely endorse slowing down for safety reasons? (That is, genuine move from believing the x% of researchers and board members believing in slowing to thinking x*10% do).
- *** (UF) Red-teaming: Between 2024 and 2040, what proportion of training runs at least 50x larger than that used to train GPT-4 will be red-teamed before they're started?
  - See p.14 of: https://arxiv.org/pdf/2004.07213.pdf#page=5
  - Question resolves to the best distributional guess over such proportions by the [resolution council].
- - By 2025/2028/2033 will there be an organization in the US that specializes in red-teaming for AI labs, and employs at least 20 people?

- – By 2027/2032/2037 how many organizations specializing in AI red-teaming will there be in the United States?
- – How much would the answer to the previous question increase if Open Phil. spent $25 million trying to get such organizations set up?
- – In 2025/2028/2033/2040 what % of leading labs will red-team their information security?
- – By 2025/2030/2035 will the US/EU/Japan legally require audit trails from AI labs?
  – See p.23 of this paper.
- – By 2025/2030/2035 will there be an agreed standard way of measuring compute used to train a model, which Open AI, Anthropic, Facebook AI Research and DeepMind (if still in existence) and any other leading US labs all adhere to when reporting compute usage?
  – See p.35 of this paper
- – Same as the previous question except that it still resolves 'yes' if there is one remaining lab on the list/leading lab which doesn't use the measurement standard.
- ** In 2025/2028/2031/2035 what proportion of leading labs will have internal (to the org that owns them) ethics boards which they have to run their work past?
- – How many times by 2030 will an ethics board overseeing a leading lab be abolished?
- ** Will the International Electrotechnical Commission release a set of safety standards for AI labs by 2024/2028/2032?
  – (See p.10 of: https://www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf)
- ** Will the International Organization for Standardization release a set of safety standards for AI labs by 2024/2028/2032?
  – (See p.10 of: https://www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf).
- – Will the International Telecommunications Union release a set of safety standards for AI labs by 2024/2028/2032?
  – (See p.10 of: https://www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf).
- ** AI Governance survey 1: In 2025/2030/2035 what proportion of leading labs will have a policy of always running pre-deployment risk assessments before deploying models, or before deploying models over a certain capability threshold?
  – See: https://arxiv.org/pdf/2305.07153.pdf

- ** AI Governance survey 2: What is the chance by 2026/2031/2036 that the US/EU/China/South Korea/Japan/Canada legally requires the risk assessments mentioned in the previous question?
- *** (FT) AI Governance survey 3: Will any lab deliberately connect a model trained with 1/2/5 OOMs more compute than GPT-4 to the internet by 2025/2030/2040?
  - In case of ambiguities, this question is resolved by the [resolution council].
- ** AI Governance survey 4: Will the US/EU/China ban connecting models over a certain size to the internet by 2025/2030/2040?
- *** (FT) AI Governance survey 5: In 2025/2030/2035 what proportion of leading labs will specifically test for power-seeking behaviour before deploying models?
- ** AI Governance survey 6: What is the chance by 2026/2031/2036 that the US/EU/China/South Korea/Japan/Canada legally requires the type of risk assessment mentioned in the previous question?
- *** (FT) AI Governance survey 7: By 2025/2030/2035 what proportion of leading labs will specifically test for ability and willingness to manipulate before deploying models?
- ** AI Governance survey 8: What is the chance by 2026/2031/2036 that the US/EU/China/South Korea/Japan/Canada legally requires the type of risk assessment mentioned in the previous question?
- ** AI governance survey 9: By 2025/2030/2035 what proportion of leading labs will have a plan for how they respond to security incidents (e.g. cyberattacks)?' (quote from p.18: https://arxiv.org/pdf/2305.07153.pdf).
  - AI governance survey 10: What is the chance by 2026/2031/2036 that the US/EU/China/South Korea/Japan/Canada legally requires the plan mentioned in the previous question?
- *** (FT) AI Governance survey 11: In 2025/2030/2035 what proportion of leading labs will have a policy of reviewing potential dangers before starting training runs of systems, or systems over a certain size?
- ** AI Governance survey 12: What is the chance by 2026/2031/2036 that the US/EU/China/South Korea/Japan/Canada legally requires the type of risk assessment mentioned in the previous question?
- *** (FT) AI Governance survey 13: In 2025/2030/2035 what proportion of leading labs will have an official emergency response procedure, to be implemented if they suddenly realize that one of their models is dangerous?
- ** AI Governance survey 14: What is the chance by 2026/2031/2036 that the US/EU/China/South Korea/Japan/Canada legally requires the type

of emergency response plan mentioned in the previous question?

- *** (FT) AI Governance survey 15: By 2025/2030/2035 what proportion of leading labs will have official, published, alignment strategies?
- ** AI Governance survey 16: What is the chance by 2026/2031/2036 that the US/EU/China/South Korea/Japan/Canada legally requires that labs have published alignment strategies?
- – AI Governance survey 17: In 2025/2030/2035 what proportion of leading labs will offer bug bounties for finding safety flaws in their models?
- – AI Governance survey 18: In 2025/2030/2035 how much will the median leading lab offer in $s per bug identified?
- – AI Governance survey 19: In 2026/2031/2036 how many $s will leading labs pay out in bug bounties?
- – AI Governance survey 20: In 2025/2030/2035 what proportion of leading labs will have a policy of sharing information about security breaches with other leading labs?
- – AI Governance survey 21: In 2025/2030/2035 what proportion of leading labs will be in compliance with one or other of the following security standards: https://en.wikipedia.org/wiki/ISO/IEC_27001 https://en.wikipedia.org/wiki/NIST_Cybersecurity_Framework
- ** AI Governance survey 22: What is the chance by 2026/2031/2036 that the US/EU/China/South Korea/Japan/Canada legally require compliance with some particular information security standard?
- *** (FT) AI Governance survey 23: By 2025/2030/2035 what proportion of leading labs will have published one of their own internal risks assessments of a model?
  - Note on relevance: Would be a reason for optimism.
- ** AI Governance survey 24: By 2025/2030/2035, what % of national security experts, when surveyed, will say that US/EU/Japanese/South Korean leading labs have good enough information security to make information theft from them by the Chinese government unfeasible?
- – AI Governance survey 25: By 2025/2030/2035 what proportion of leading labs will have 'a board risk committee, i.e. a permanent committee within the board of directors which oversees the lab's risk management practices'? (Quote from p.19: https://arxiv.org/pdf/2305.07153.pdf).
- *** AI Governance survey 26: By 2025/2030/2035, what proportion of leading labs will have published a statement on 'how they make high-

stakes decisions regarding model development and deployment'. (Quote from p.19 of https://arxiv.org/pdf/2305.07153.pdf).

- *** AI Governance survey 27: By 2025/2030/2035, what proportion of leading labs will conduct [know-your-customer]((https://www.dowjones.com/professional/risk/glossary/k your-customer/) screenings of organizations asking to access their most powerful AI models?
- ** AI Governance survey 28: By 2025/2030/2035 what proportion of leading labs will 'perform rigorous background checks before hiring/appointing members of the board of directors, senior executives, and key employees.' (Quote from p.19 of https://arxiv.org/pdf/2305.07153.pdf).
- ** AI Governance survey 29: By 2025/2030/2035 what proportion of leading labs will have a policy of boxing models they think are sufficiently dangerous?
- – AI Governance survey 30: By 2025/2030/2035 will it be legally required in the US/China/the EU/Japan/South Korea to box models above a certain size?
- ** AI Governance survey 31: By 2025/2030/2035 what proportion of leading labs will have 'a system that is intended to track all copies of the weights of powerful models'? (Quote from p.19 of https://arxiv.org/pdf/2305.07153.pdf).
- ** AI Governance survey 32: By 2025/2030/2035 how many models within an OOM of the largest ever at the time they are published, will have been open-sourced?
- ** AI Governance survey 33: By 2026/2031/2036 will any of the US/China/the EU ban publishing models above a specific threshold meant to measure capabilities from being open-sourced?
- ** AI Governance survey 34: By 2025/2030/2035 what proportion of leading labs will only deploy models above a certain size via an API? (https://en.wikipedia.org/wiki/API).
- – AI Governance survey 35: By 2025/2030/2035 will any leading lab have conducted a training run at least 2/3/4 OOMs bigger than their previous largest ever. (Ignore labs with no previous training runs, who become leading on their first run.)
- ** AI Governance survey 36: Conditional on leading labs having risk assessments before training, how many will also have a policy of conducting risk assessments before making any significant changes to the weights of models, such as by fine-tuning them?
- ** AI Governance survey 37: By 2026/2031/2036 will the EU/the US/Japan/South Korea/Taiwan/Australia/Canada/the UK legally require that training runs above a certain size are pre-registered with the

government?

- – AI Governance survey 38: In 2025/2030/2035, what proportion of leading labs will claim to have implemented one of the following risk management frameworks: https://www.nist.gov/itl/ai-risk-management-framework https://www.iso.org/iso-31000-risk-management.html
- *** (FT) AI Governance survey 39: By 2025/2030/2035 what proportion of leading labs will have a policy of performing pre-deployment risk assessments even when they are only deploying a model internally for work within the lab itself?
- – AI Governance survey 40: By 2026/2031/2036 will the EU/the US legally require risk assessments in the circumstances described by the previous question?
- *** (UF) AI Governance survey 41: By 2025/2030/2035 what proportion of leading labs will give access to their models to researchers from other labs, for safety-testing?
  - Resolves to the distributional guess about this proportion by the [resolution council]
- ** AI Governance survey 42: By 2026/2031/2036 will the EU/the US legally require that outside researchers are given the access described in the previous question?
- – AI Governance survey 43: By 2026/2031/2036 will there be a citizens assembly on AI with over 350 participants in the US/EU? (https://www.parliament.uk/get-involved/committees/climate-assembly-uk/about-citizens-assemblies/)
- – AI Governance survey 44: Will any proposal from such a citizens assembly be adopted into law?
- – AI Governance survey 45: Will any leading lab voluntarily adopt a proposal from such an assembly by 2027/2032/2037?
- – AI Governance survey 46: In 2025/2030/2035, what proportion of leading labs will have a policy of publishing what safety/alignment measures they took each time they release a new model (API-only access counts as release)?
- – AI Governance survey 47: By 2026/2031/2036 will it be legally required in the EU/US for labs to publish such a list of safety measures when they release a model?
- – AI Governance survey 48: In 2025/2030/2035, what proportion of leading labs will have a policy of publishing all internal safety/alignment research, without exception?

- – AI Governance survey 49: In 2025/2030/2035, what proportion of leading labs will have a policy of publishing all internal safety/alignment research, unless they judge it to be infohazardous?
- ** AI Governance survey 50: By 2027/2032/2037 how many papers on alignment/existential risk from leading labs will have been accepted at one of the following conferences:
  - https://en.wikipedia.org/wiki/Conference_on_Neural_Information_Processing_Systems
  - https://en.wikipedia.org/wiki/Association_for_the_Advancement_of_Artificial_Intelligence
  - https://en.wikipedia.org/wiki/International_Conference_on_Learning_Representations
  - https://en.wikipedia.org/wiki/International_Conference_on_Machine_Learning
- *** (UF) AI Governance survey 51: In 2025/2030/2035 what proportion of leading labs will have a written plan specifying the actions they would take if they thought they were about to train [a system above a pre-specified qualitative capabilities threshold capturing some notion of transformative capabilities/human-level AI/a system that might recursively improve itself]?
- ** AI Governance survey 52: In 2025/2030/2035 what proportion of leading labs will have a *publicly accessible* plan specifying the actions they would take if they thought they were about to train [a system above a pre-specified qualitative capabilities threshold capturing some notion of transformative capabilities/human-level AI/a system that might recursively improve itself]?
- – AI Governance survey 53: In 2025/2030/2035 what proportion of leading labs will have detailed written models of how AI takeover, or other AI catastrophes might occur, that are the product of at least 50/100/350 hours of work?
- – AI Governance survey 54: In 2025/2030/2035 what proportion of leading labs will 'publish an annual report where they present the predicted and actual impacts of their work, along with the evidence and assumptions these are based on.' (quotation from p.21: https://arxiv.org/pdf/2305.07153.pdf).
- – AI Governance survey 55: By 2025/2030/2035 will any leading lab have spent time and money on a project to educate their investors and/or employees about AI takeover risk?
- – AI Governance survey 56: In 2025/2030/2035, what proportion of leading labs will have an independent board for employees to whistleblow to anonymously?

- – AI Governance survey 57: In 2025/2030/2035 what proportion of leading labs will have 'have an onboard process for managers and new employees that involves content explaining how the organization believes a responsible AGI developer would behave and how they are attempting to meet that standard.'
  – Quotation from p.21: https://arxiv.org/pdf/2305.07153.pdf
- – AI Governance survey 58: In 2025/2030/2035 what proportion of leading labs will have a policy of measuring how much their internal culture values criticism?
- – AI Governance survey 59: In 2025/2030/2035 what proportion of leading labs will have a person responsible for making their hiring more inclusive and increasing the diversity of their work-force?
  – Note: Interesting forecasting question + correlates with politicization of AI. But also feels a bit distracting.
- *** (FT) AI Governance survey 60: In 2025/2030/2035 what proportion of leading labs will have 'independent safety and ethics advisory boards'?
  – Reference: see p.21 here.
- – AI Governance survey 61: By 2025/2030/2035 what proportion of non-Chinese leading labs will 'report to an international non-governmental organization (INGO) that is publicly committed to human rights and democratic values.'? (quotation from p.22: https://arxiv.org/pdf/2305.07153.pdf)
- – AI Governance survey 62: In 2025/2030/2035 what proportion of leading labs will 'have an independent board of directors with technical AI safety expertise who have the mandate to put the benefits for society above profit and shareholder value'? (quotation from p.22: https://arxiv.org/pdf/2305.07153.pdf)
- ** AI Governance survey 63: By 2027/2032/2042 what proportion of leading labs will have plans for how to switch to safer research with narrower applications, if they begin a project to build a human-level AI/PASTA and then halt it because they decide it is too dangerous to continue with?
- ** AI Governance survey 64: In 2025/2030/2035 will there be some official standard of interpretability that all US/EU labs have voluntarily agreed any models (or models trained with > a certain amount of compute) must comply with?
- ** AI Governance survey 65: In 2026/2031/2036 will there be some official standard of interpretability that AI labs are legally required to comply with in the US/EU/UK/South Korea/Taiwan/Japan/Australia/Canada?
- – AI Governance survey 66: In 2025/2030/2035 what proportion of leading labs will have some procedure by which they check how safe

a training run is *during the training run* (i.e. not just before starting training or after the training is done)?

- – AI Governance survey 67: In 2025/2030/2035 what proportion of leading labs will have a policy of saving all logs of interactions with the systems they are training?
- – AI Governance survey 68: By 2026/2031/2036 will AI labs be "forced [by law] to have systems that consist of ensembles of capped size models instead of one increasingly large model" in the US/EU/UK/South Korea/Japan/Australia/Canada? (quotation from p.22 of https://arxiv.org/pdf/2305.07153.pdf).
- – AI Governance survey 69: The previous question, but what proportion of leading labs outside will agree to the restriction voluntarily.
- – AI Governance survey 70: By 2026/2031/2036 what portion of leading labs will have a policy of only allowing access to models via API and also only allowing access to the API to organizations that have passed some kind of pre-use vetting process specified by the lab?
- – AI Governance survey 71: In 2025/2030/2035 what proportion of leading labs will have a policy of 'ensur[ing] that AI systems in an ensemble communicate in English' so that their communications can be analysed for warning signs of dangerous behaviour? (quotation from p.22 of https://arxiv.org/pdf/2305.07153.pdf).
- – AI Governance survey 72: Conditional on leading labs conducting simulated cyberattacks once every 6 months on their own systems, as a security test, how much less likely are they to have their model weights stolen in any 3-year period?
- *** (UF) AI Governance survey 73: In 2025/2030/2035 what proportion of leading labs will have internal controls and processes that prevent a single person or group being able to deploy an advanced AI system when governance mechanisms have found this to be potentially harmful or illegal?
  - – Background reference: question wording taking from p.22 of this paper
  - – This question could potentially be made resolvable by appending "according to the [resolution council]" to the question description.
  - – Note: Having a group maintaining a list of known "leading labs" through time that this could refer to would be potentially valuable.
- ** AI Governance survey 74: In what year will the first AI able to experience conscious pain/pleasure exist? [the AI gov survey talks about "sentience", but I think this is a little clearer].
- ** AI Governance survey 75: In 2026/2030/2045 how many copies of sentient models will be running per day across the whole world?

- ** AI Governance survey 76: In 2025/2030/2035 what proportion of leading labs will have a policy of 'report[ing] cluster sizes and training plans to other AGI labs'? (quotation from p.22 of https://arxiv.org/pdf/2305.07153.pdf).
- – AI Governance survey 77: In 2025/2030/2035 what proportion of leading labs will make use of Responsible AI Licenses when granting access to their models? (https://www.licenses.ai/).
- ** In total, how many pieces of legislation designed primarily to regulate AI will the US government pass by 2025/2030/2035?
  - Relevance: a sign of both overall regulatory burden, which will slow AI down, and of how seriously the US gov is taking AI, and of how much successful lobbying in the US for restrictions on AI there has been.
- ** In total, how many pieces of legislation designed primarily to regulate AI will the state government of California pass by 2027/2033/2039?
- ** How many pieces of legislation designed primarily to regulate AI will EU governments as a whole pass by 2027/2033/2039?
- ** By 2025/2030/2035 how many times (if at all) will the US pass regulations which limit what models can be trained, and not just what models can be deployed?
  - Relevance: not just that even training powerful models can be dangerous, but also this is a noisy but not totally useless signal of how much interest groups pushing for regulation to stop AI taking their jobs are pushing for things that will directly prevent the training of powerful AIs rather than just their deployment.
- ** By 2025/2030/2035 how many times (if at all) will an EU state, or the EU as whole, limit what models can be trained, and not just what models can be deployed?
- – By 2025/2030/2035 how many times (if at all) will China (legally or defacto) limit what models can be trained, and not just what models can be deployed?
- ** Suppose we get a representative sample of pol sci lecturers and politics reporters are reputable mainstream outlets in 2028/2033/2038 and get them to investigate all pieces of AI regulation passed by the US federal government and US state governments from 2024 to the date of the survey: how many pieces of legislation will >65% of the sample agree were (probably) passed wholly or primarily because people motivated an interest in their jobs/industry not being automated away lobbied for them?
- – How many laws which meet the criteria in the previous question will ban even training a particular type of AI?

- ** How many laws the meet the standard in the previous question will be described by the sampled experts as "broad, bans many types of AI that it probably wasn't intended to cover"?
- – How many strikes will occur in the EU/US where one of the demands of the striking workers is that some job-automating model not be deployed, by 2026/2030/2038?
- – How many anti-automation protests with over 2k participants in the US/EU/China by 2027/2032/2038?
- *** (FT) By 2025/2030/2040, will any US State of the Union addresses mention the issue of AI takeover risk as a concern?
  - Question is resolved by querying whitehouse.gov or the corresponding official transcript for the US state of the union for yearly addresses (e.g., https://www.whitehouse.gov/state-of-the-union-2023/), and searching for "artificial intelligence", "AI", "machine learning" and similar, and then making a subjective judgment call about whether these refer to AI takeover—as opposed to less different worries, like AI ethics or job loss.
- *** (FT) In [year], will any State of the European Union addresses by the president of the EU mention the issue of AI takeover risk as a concern?
  - Question is resolved by querying https://state-of-the-union.ec.europa.eu and searching for "artificial intelligence" and making a judgment call about whether about whether these refer to AI takeover—as opposed to different worries, like AI ethics or job loss.
  - This question might be useful as a short-term resolvable proxy that could incentivize having better models of politics and AI.
  - Note: Question would have resolved positively in 2023
- To do: find the equivalent for China for the above two questions
- *** (FE, UF) By 2026/2030/2040 how much will compliance costs be as a percentage of AI industry gross profit (https://en.wikipedia.org/wiki/Compliance_cost), both as estimated by the [resolution council]?
- *** (FT) By end-of-year 2026 will OpenAI or DeepMind announce that they are pausing all training runs above a certain size for safety reasons?
- ** In what year (if ever) will the first (within the US/EU) anti-AI protest with >10k participants occur?
- – For every anti-AI protest in the US/EU with over >50k participants, how many (extra) pieces of legislation regulating AI would be passed?

Who will be at the forefront of AI research?

Governments, if so, which ones? small companies or large companies? US or Chinese companies?, etc.

- ** By what year will the military of a nation first spend the equivalent of 50B USD on AI research?
  - Question note: 50B adjusted for inflation after 2023.
- *** (UF) In 2026/2028/2035/2040 will the largest training run—in terms of FLOPs spent—take place in a) an American, b) a Chinese or c) an European lab?
  - Question details: See the section "some basic terms" for the definition of "floating point operation" and its use as a measure of size of training runs; see also the first section of this report on AI timelines.
- *** (UF) In 2026/2028/2035/2040 how many FLOPs will the largest training run in [Europe/China/the US] have taken?
  - Question details: See the section "some basic terms" for the definition of "floating point operation" and its use as a measure of size of training runs; see also the first section of this report on AI timelines.
- *** (FT) By 2030, as estimated by the [resolution council], will the largest training run have been carried out by a private lab or by a government?
- *** (FT) By 2028, will the best Large Language Model have been produced by a company specializing in AI, like Open AI, or a large non-specialist tech company, like Facebook or Baidu?
  - Note: best could be operationalized with reference to subjective quality estimation, to perplexity for a reference text, to some bag of benchmarks, etc.
- – Conditional on no Chinese invasion of Taiwan, will the leading Chinese chip manufacturer in 2036 be as advanced as TSMC?
- – By 2035, how many labs will have become a leading lab within 2 years of their founding?
- ** Will the US government nationalize any AI lab by 2035?
- ** By January 2025, what % of top-5 entries on Allen Institute for AI leaderboards will be from Chinese labs? (https://leaderboard.allenai.org/).
- *** (FT) Will the US create a National AI research resource, as suggested here?
- – How much does the risk of AI takeover by 2100 currently rise/fall per $100 million the US government/EU spends on supplying university labs with compute?

- – Will any EU government nationalize an AI lab by 2027/2035?
- – Will China nationalize an AI lab by 2027/2035?
- ** By 2040 will any entity which is neither a for-profit company (even in the limited capped profit OpenAI sense) nor a government spend at least $10 billion on a training run?
- – If the US were to give immediate citizenship and working rights to anyone an AI lab wanted to hire as a researcher, as of tomorrow, how much quicker would the US produce a model which meets all the criteria in this Metaculus question?
- ** If the US were to give immediate citizenship and working rights to anyone who an AI lab wanted to hire as a researcher for the next 25 years, how much would that increase the chance that the US rather than China reaches PASTA AI first?
  - See here for definition of 'PASTA': https://www.cold-takes.com/transformative-ai-timelines-part-1-of-4-what-kind-of-ai/
- – How much will the US export controls of October 2022 (https://en.wikipedia.org/wiki/United_States_New_Export_Controls_on_Advanced_Computing_and_Semiconductors_to_China) reduce the proportion of top-5 Allen AI leaderboard entries in 2027 which are from China?
- – How much have the October 2022 export controls reduced the chance that China will reach PASTA AI first?
- – How much will the CHIPS act (https://en.wikipedia.org/wiki/CHIPS_and_Science_Act) increase the proportion of leading labs located in the US in 2030?
- – How much has the CHIPS act increased the proportion of leading labs located outside of the US/China in 2032?
- – How much has the CHIPS act increased the chance that the US will reach PASTA AI first?
- – Market value of Open AI/Deep Mind/Anthropic in 2032?
- ** Market value of the 5 largest tech companies in 2030/2035/2045 (conditional on no AI takeover)?
- ** During any year between now and 2045 will there be a Russian leading lab?
- ** During any year between now and 2045 will there be a Taiwanese leading lab?
- ** During any year between now and 2045 will there be a Japanese leading lab?

- ** During any year between now and 2045 will there be a South Korean leading lab?
- *** (FT) During any year between now and 2045 will there be an Israeli leading lab?
  - Note about importance: The Israeli government seems competent in a way which other countries are not, e.g., see their successful efforts to become a nuclear power.
- ** During any year between now and 2045 will there be an Australian leading lab?
- *** (FT) During any year between now and 2045 will there be an Indian leading lab?
  - Note on importance: Seems important, given India's huge size & the technical competence of India & the Indian diaspora.
- ** During any year between now and 2045 will there be a Canadian leading lab?
- ** Which 3 EU countries are most likely to have leading labs between now and 2045?
- ** By 2030/2035/2045 will any country outside the EU other than the US/China/Russia/Taiwan/Japan/South Korea/Israel/Australia/Canada have a leading lab?
- ** If we speed up US AI research relative to China by 3 years by 2035, how much higher/lower is the risk of AI takeover by 2100?
  - Speeding up := cause the US to be where they would otherwise have been in 2038, whilst Chinese progress remains fixed, or cause an increase in the US position relative to China of equivalent size to doing this.
-
  - How much lower/higher is AI takeover risk if a democratic government builds the first model that meets the resolution criteria of this Metaculus question, compared to the case where a Western for-profit business does this first?: https://www.metaculus.com/questions/5121/date-of-artificial-general-intelligence/
- ** What is the chance that by 2040 one of the current G20 nations announces it is going to spend $100 billion on a project to build highly advanced AI?
-
  - What will be the percentage change in the number of H-1B visa's granted to tech workers by 2027? https://en.wikipedia.org/wiki/H-1B_visa)
  - Question details: Could be negative.

- ** By what year (if ever) will a single AI company spend $1/10/100bn on compute in one year?
- ** How many AI researchers will there be in 2025/2028/2032 at DeepMind/Open AI/Anthropic/Facebook AI Research? (counts as a researcher if they have co-authored at least one paper while at the org).
- – AI Index 1: Global Vibrancy: The US / China / India / UK / Israel / Canada / South Korea / Germany will have a value of y on the index in 2025/2030/2035 (https://aiindex.stanford.edu/vibrancy/).
- – AI Index 2: Number of Newly Funded AI Companies: The US / China / India / UK / Israel / Canada / South Korea / Germany will have a value of y in 2025/2030/2035
- – AI Index 3: The US / China / India / UK / Israel / Canada / South Korea / Germany will have a value of y in 2025/2030/2035
- – AI Index 4: Number of AI Patents / Grants: The US / China / India / UK / Israel / Canada / South Korea / Germany will have a value of y in 2025/2030/2035
- – AI Index 5: Number of AI Repository Citations: The US / China / India / UK / Israel / Canada / South Korea / Germany will have a value of y in 2025/2030/2035
- – AI Index 6: Relative AI Skill Penetration The US / China / India / UK / Israel / Canada / South Korea / Germany will have a value of y in 2025/2030/2035
- – How much money will the US/EU and EU governments put into providing compute to academic AI researchers by 2027/2030/2035?
- *** (FT) Will the [US/Chinese] [government/military] openly announce a project with the goal of building AI "as capable/intelligent as a human" by 2026/2030/2045?
  - Note: "Openly announce" would require some operationalization. In particular, it's not clear how to access all the things that the US government "announces".
- *** (FT) US/Chinese [government/military] be reported to secretly launch a project with the goal of building AI "as capable/intelligent as a human" by 2026/2030/2045?
  - Question resolves according to a search on Google News or similar for keywords on this topic. It resolves positively if a reputable media outlet—reputable as subjectively judged by the [resolution council]—reports the proposition in the question title.
- *** (FT) US/Chinese government reported to be trying to deliberately slow down the development of AI due to safety concerns by

2026/2030/2045?

  - Question resolves according to a search on Google News or similar for keywords on this topic. It resolves positively if a reputable media outlet—reputable as subjectively judged by the [resolution council]—reports the proposition in the question title.

Questions about militarization.

- ** The first time that the US government produces an AI using within 1/20th of the compute of the largest training run at the time, will this be done by the US military?
-
  - The first time the EU or an individual EU government produces an AI using within 1/20th of the compute of the largest training run at the time, will this be done by the militaries of one or more EU states?
- *** (UF) If a private lab develops transformative AI first, what is the chance—conditional on them retaining control of the AI—that it will effectively be a world government within 30 years?
  - Question resolution details: Question would resolve positively if a set of companies which includes the company which first develops transformative AI has the powers of a world government
  - Note: "the powers of a world government" is under-operationalized. Some things that it could mean:
    * give commands to the US and Chinese government, and have them obey? For instance, will there be an episode such as the Suez Canal Crisis, in which China and the US are on one side, and the AI lab is on the other side, and the AI lab is able to impose its will?
    * determine large-scale decisions affecting humanity, such as setting immigration policy for large nations, deciding inflation targets, commanding large armies, or deciding how to allocate a large share of human labor?
    * decide on and pursue large and ambitious goals, such as terraforming the Earth or other planets, eradicating major diseases, or starting novel scientific research programs?
- *** (UF) Condition on the US government being the first to develop transformative AI, and on it subsequently maintaining control of said AI. Then, will the US government effectively be a world government within 30 years?
- *** (UF) Condition on the China government being the first to develop transformative AI, and on it subsequently maintaining control of said AI. Then, will the China government effectively be a world government within 30 years?

- *** (FT) Will any nation build autonomous, AI-operated battle drones or robots by 2025/2030/2050?
- – Conditional on the answer to the previous question being yes, will such autonomous weapons end up connected to the internet?
- ** If transformative AI is created by for-profit companies, what % of total worldwide profit will be captured by those companies 7 years later?
- ** How much more likely is democracy to survive in scenarios where the US gov reaches transformative AI first, compared to scenarios where US for-profit companies or other US non-state, non-academic, organizations reach transformative AI first?
- ** Rank these scenarios by which has highest AI takeover risk: the US government builds the first significantly-more-intelligent-than-humans AI, a US for-profit company builds the first significantly-more-intelligent-than-humans AI, the Chinese government builds the significantly-more-intelligent-than-humans AI, a Chinese for-profit company builds the first significantly-more-intelligent-than-humans AI?
- ** In 2027/2037/2047 what % of the military budget of the US/China/the median EU country will be spent on AI and autonomous weapons?
- ** By 2040 will the US/Chinese/Russian/French/UK military plan for wars on the assumption that AI will sometimes make decisions involving coordinating over >10k troops, or assaulting a particular enemy military base or settlement of >40k people?
  - Note: It might be more meaningful to look at more active militaries, e.g., Israel, Ukraine, countries bordering the former Soviet Union, countries with active conflicts in the last decade, etc.
- ** In 2040/50, conditional on no AI takeover, how many law-enforcement robots will there be in USA/EU/China?
- – By how much has the existence of the Wilson Centre's (https://en.wikipedia.org/wiki/Woodrow_Wilson_International_Center_for_Scholars) AI lab (https://www.wilsoncenter.org/artificial-intelligence-lab) increased the expected spending (in millions of $s) of the [US military/US federal government as a whole] on AI research between 2022 and 2045? (Relative to a counterfactual where the lab didn't exist)?
- ** Currently for every Senator who talks to an expert about AI takeover risk, how much does expected US military spending on AI research increase by 2045 increase?
- ** Currently for every Senator who talks to an expert about AI takeover risk, how much does expected US federal government spending on AI research increase by 2045 increase?

- ** What will be the growth rate of US military spending on AI 2024-2035?
- ** What will be the growth rate of Chinese military spending on AI 2024-2035?
- ** What will be the growth rate of EU military spending on AI 2024-2035?
- – What will be the growth rate of Japanese military spending on AI 2024-2035?
- ** What will be the growth rate of Russian military spending on AI 2024-2035?
- – What will be the growth rate of South Korean military spending on AI 2024-2035?
- ** What will be the size of the training run which produces the largest AI that the US/China places in control of some autonomous weapons by 2033/2043/2053? (Enter '0' if you think this will never happen.) (Relevance: this is a measure of whether AI large enough to plausibly be able to form a plan to takeover, will actually be put in charge of military hardware that could be used in such a plan.)
  - Note: should be a distributional/multiple choice question. Probably not resolvable.
- ** By 2040/2050 will there be an AI model in use by the US/Chinese military, such that, if a panel of AI and information security experts had detailed knowledge of the capabilities of the model, and of the security measure taken to prevent control of autonomous weapons being seized, they would judge the model had an >40% chance of seizing control of weapons systems worth at least $200 million if it tried to?
  - Note: probably not resolvable

Questions about how agent-y and general future AIs will be, and how that affects X-risk from AI

- ** By 2030/2035/2040, will some companies will employ AI middle-managers.
  - Note: "AI middle managers" might look different from current middle managers. E.g., maybe you have humans aided by middle managers. Or maybe you have "analytics systems" that people working at McDonald's report to. Thus this question is under-operationalized.
- – In what year will the first AI personal assistant be sold?
- ** By 2028/2030/2035/2040 will any of the world's leading AI labs explicitly have as their goal building aligned agents which can be trusted to govern the world?

- ** By 2028/2030/2035/2040 will any credible lab have as its publicly stated goal building "human-like agents"?
  - Unclear how to operationalize "credible". Two options: by other machine learning experts, by the [resolution council]
- *** (FT) Will the US government attempt to ban the training of AI agents above some capabilities threshold by 2030/2035/2040/2045?
- *** (FT) Will the EU attempt to ban the training of AI agents above some capabilities threshold by 2030/2035/2040/2045?
- *** (FT) Will China attempt to ban the training of AI agents above some capabilities threshold by 2030/2035/2040/2045?
- *** (UF) By the time 25% of the US economy is automated, will over 10% of AI compute be spent by only [1/2] model(s)?
- *** (UF) Will there be a point in time where a) over 70% of labour in the US as performed in 2023 has been automated, but b) no one AI model is spending more than 5% of total AI compute?
  - See the "recurring terms" section for an operationalization of automating over 70% of labour.
- *** (UF) In [year], what fraction of US work hours will be spent interacting with an AI?
  - Communicating, giving feedback, following orders, giving orders, etc. to AIs.
- ** % chance that we will have a superhuman oracle/tool AI before we have superhuman agents [we could refer forecasters to Bostrom's Superintelligence for definitions of these terms.]
- ** By 2028 will all top labs (defined as last training run used an amount of compute within 2 OOM of currently largest ever run) prevent users from turning cutting-edge (say, the labs 2 most recent) LLMs into components of agents, via writing programs which ask the LLM to write a program to accomplish a task, and then run that program?
- ** By 2030/2035/2040/2045/2050 will an AI set up purely to be a question answering tool try to steal more resources, or otherwise manipulate the world outside itself, in order to be better answer a question?
- ** If an NGO were to spend $400 million over the next 5 years investigating how to speed-up the development of tool/oracle AIs relative to agents, and then investing in the (by-their-lights) best strategies for doing this they discovered, how much would AI takeover risk by 2100 increase/reduce?
- ** If an NGO were to spend $400 million over the next 5 years investigating how to speed-up the development of tool/oracle AIs relative to agents, and then investing in the (by-their-lights) best strategies for doing this they

discovered, how much lower/higher would be the chance that by the time 35% of the economy is automated, 3 models or less are responsible for at least 40% of the economic output generated by AI? (that is, at least 40% of the output in the year where 35% automation is reached.)

- ** When will interpretability work first prove to the satisfaction of an expert panel of ML researchers and analytic philosophers of mind, that an AI trained with at least 5x compute as GPT-4 is engaging in explicitly reasoning through the likelihood and value of various outcomes conditional on different actions, and then reliably selecting the action that maximizes utility?
- – How many hits will the AgentGPT website receive in 2023?
- – How many unique hits will the AgentGPT website receive in 2023?

Based on comments in the Slack at Trajan:

- *** (FT) When, if ever will the first company with a market cap of over [$1B/$10B/$100B] be openly run by a CEO-bot?
  - Question details: Market capitalization is understood to be in inflation-adjusted 2023 dollars.
- – When, if ever will the first company openly run by a CEO-bot make the S&P 500?

Risks of various kinds from EAs and other people concerned about AI X-risk getting things wrong

- *** (UF, FE) What fraction of upper-management at DeepMind/Open AI/Anthropic secretly agree with Eliezer Yudkowsky's claim that the first lab to reach "AGI" needs to perform a "pivotal act" that prevents anyone else building dangerous AGI?
  - Note: This is bad as a forecasting question, because it's unresolvable, but might be a good modeling exercise.
- ** What % of people in upper-management positions at DeepMind/Open AI/Anthropic believe that getting to a superintelligent AI aligned with your own personal goals first would enable you to "takeover the world", or something along those lines.
- ** % chance Musk secretly wants to seize power via a superintelligent AI aligned to him.
- – % chance of a major public scandal involving MIRI by 2025/2030
- – % chance of a major public scandal involving Anthropic by 2025/2030
- – On average, how much earlier in months does the first agent AI better than humans at all jobs become each time a prominent EA claims that

such agents will inevitably be built, compared to the counterfactual where they didn't say that?

- ** How much more money will be invested worldwide in AI by 2030, relative to a scenario in which all EA organizations and prominent EAs just stop talking about or working on AI safety entirely?
- ** By 2025/2030/2035, how many labs will have been significantly offended by attempts to evaluate how safe one or more of their models are, and cut off access to the evaluating group in response?
- – Will any EA organization donate money to/invest in a differentially-capabilities-promoting AI lab by 2025/2030/2040?[19]
- *** (UF) In 2025/2030/2035 what % of AI safety researchers would, when surveyed, say that Open Philanthropy's 2017 grant to Open AI overall had a negative impact?
- *** (UF, FE) How much does the probability that an AI model as smart as a human or above attempts to take over the world go up for every discussion of how AI could take over the world in its training data?
  - Note: Seems like a fun Fermi exercise. We (Arb/Samotsvety) might be interested in giving it a shot.
- ** How much does publishing a piece on AI takeover on the internet increase in expectation the number of discussions of AI takeover in AI training data?

General Warning Signs

- – By what year will an AI first be found to have hacked into computers to form a botnet?
  – For a reference, see Toby Ord's *The Precipice*, pp.146-7.
- ** By what year, if ever, will an AI first be found to have hacked into computers to form a botnet without having previously been instructed to do so by a human?
- *** (UF) By what year, if ever, will it be discovered that an AI has sent out copies of itself using the internet, without being instructed to do so by a human?
- ** By what year, if ever, will we discover that an AI has stolen money (via hacking, fraud etc.) without being instructed to do so?
- ** By what year, if ever, will an AI first cause damage worth over $10/25/50 million (2022 $s)?

[19] [not necessarily a stupid thing to do depending on circumstance, but this still seemed like the most natural section for this question.]

- ** By what year, if ever, will an AI first cause over 1k/10k/100k deaths, outside of a war?
- – How many times will reward hacking by a model have caused over $100k worth of damage in the China by 2027/2032/2040?
- ** Chance by 2027/2032/2040 that an AI lab will release a demonstration of an AI attempting to break out onto the internet or blackmail or lie to a human operator?
- ** Chance by 2032 that an AI designed to answer questions is caught manipulating the questions it receives to make them easier to answer.
- – By what year (if ever) will an AI working for a company first embezzle $1 million from a companies funds? (Resolves to this being confirmed by court documents.)
- – By what year (if ever) will court documents confirm that an AI has attempted to blackmail a user?
- – Chance by 2028/2035/2045 that a military or multiple credible major news outlets report that an AI soldier has deserted their ranks?
- ** When will the first major cyberattack (over $20 million worth of damage) be attributed to an ML system?
- *** (UF) When if ever, will an AI seize control remotely of a computer cluster?
- ** When, if ever will an AI seize control of one of the 20 largest compute clusters in the world?
- *** (UF, FE) When, if ever, will over [0.1%/1%/10%] of compute in the world be under the control of AIs who have stolen it, without being directed by a human to do so?
- – Will an AI ever perform a DDoS attack (https://en.wikipedia.org/wiki/Denial-of-service_attack) against LessWrong, or another website strongly associated with concerns about AI takeover, without first being instructed by a human to do so? (Relevance: proxy for AIs attempting to suppress concern about AI takeover on their own initiative.)
- ** By 2090, will an AI attempt to destroy a project to build another powerful AI?
- ** If the previous question resolves yes, will the first time it happens be because the AI was instructed to take out the rival project by a human?
- ** By 2090, will an AI attempt to destroy a project to build another powerful AI without being instructed by a human to do so?

- – By 2090 will there be a DDoS attack on >1 million websites, including all of the top 500 most visited? (Proxy for an AI trying to disrupt human society by "taking down the internet").
- ** By 2035/2040/2050 will any AI be found to be operating a secret factory without human knowledge?
- *** (FT) By 2026/2030/2040 will the [resolution council] know of an event that, in its judgment, demonstrates deceptive alignment in a model trained with at least 0.1/1/10/100x as much compute as GPT-4?
  - Note: See here for the definition of "deceptive alignment".
- *** (FT) By 2025/2030/2035/2040 will [an LLM/an AI system] attempt to take over a simulated world that it doesn't think is fake?
  - Resolves 'yes' per the [resolution council]'s best judgment.
- ** Assuming the previous question resolves yes, what % of machine learning researchers who think there is at least a 5% chance of AI takeover by 2100 increase in the year after the takeover attempt occurs?
- ** The previous question, but for public opinion.
- ** The previous question, but for opinion among US federal government workers.
- ** The previous question, but for researchers at Facebook AI Research (or whatever it's calling itself these days.)
- ** The previous question, but for researchers at the 6 Chinese labs with the largest budgets.

Chance and Effects of Deliberately Slowing AI Progress

- ** If all top US labs agreed to pause training of models more powerful than GPT-4 for 18 months, how many more/fewer leading labs would there be by June 2029? (Relevance: people seem to think more labs=bad, because it worsens race dynamics, and there is more chance of a single reckless actor the more top labs there are.)
- – If all top US labs agreed to pause training of models more powerful than GPT-4 for 18 months, how many more/fewer leading Chinese labs will there be by the beginning of 2030? (Relevance: people seem to think more top Chinese labs is bad a) because China is an authoritarian dictatorship, which means worse values in their AIs and probably a worse environment for dealing with safety risks from misalignment, and b) because it increases the chances of an AI arms race between China and the US.)
- – If the time when all cognitive labour can be automated is pushed back 3 year via increases in AI capabilities being slowed, how much

higher/lower a % of leading labs will be government-owned when we reach the point at which full automation is possible?

- – Conditional on top US labs carrying out an agreed pause in capabilities work of 18 months, by how many effective months will AI safety researchers say safety research was slowed by the pause, 3 years after it ends? (Operationalize the opinion of the safety researchers as their median response in a survey).
- – If leading labs all agree to a 6 month pause on larger training runs, at some point before 2035, what is the increase in the chance that 1 or more labs which are not leading at the start of the pause, have managed to catch up by 6 months after it's end, relative to a scenario without a pause? (Operationalize "catch up" as 'do a training run within an OOM of the largest ever as of the start of the pause).
- ** If all leading labs were to pause training runs for 2 years, how large, in OOMs, the jump in compute be between the largest training run before the pause, and the first larger training run after? (Reason for inclusion: some people have expressed a worry that a pause could lead a to very fast jump in capabilities after the pause, because the cost of compute will continue to fall during the pause. They think this is bad because they think a sudden, discontinuous jump in capabilities is more dangerous than reaching the same level of capability gradually over the same time period.)
- – If there is a government-imposed pause on training larger models, what % of researchers at top AI labs will say they strongly disagree with the pause one year after it started? (Relevance: It's plausibly bad if people at top lab's are angry about safety measures, since it will make them oppose further measures, stop listening to safety-minded outsiders etc.)
- *** (UF, FE) In world A, we reach AI able to fully automate labour by 2040. In world B, we reach AI able to fully automate labour by 2065. What is the chance of a global catastrophe which kills over 75% of the population in world A by 2100? And in world B by 2100?
  - Relevance: We don't want to slow AI if increases the risk of some non-AI catastrophe too much. Though this consideration is sensitive to views about population ethics and discount rates. That consideration is also sensitive about whether AI is more likely to result in extinction given that it results in a catastrophe, compared to other possible global catastrophes, but we are not considering that here.
- ** If an EA org within the US were to spend $300 million lobbying for a voluntary pause on training runs 100 times larger than GPT-4's, how much would that increase the chance of such a pausing happening?
- ** If an EA org within the US were to spend $300 million lobbying for a voluntary pause on training runs larger than GPT-4's, how much would

that reduce the chance of AI takeover by 2100?

- ** If an EA org were to spend $700 million lobbying the US government to ban training runs over 2 OOMs larger than GPT-4's how much would that reduce the chance of AI takeover by 2100?

Questions about public and researcher opinion

- *** (FT) Condition on the AI Impacts survey being re-run in 2025/2028/2035. Then, what will the median probability from respondents on permanent human disempowerment or extinction?
  - Question details: The specific wording is 'future AI advances causing human extinction or similarly permanent and severe disempowerment of the human species'? (https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/#Existential_risk)
  - Relevant because a) more concern makes researchers more cautious, and perhaps therefore all of use safer, and b) higher concern is a noisy signal that things are worse.
- ** In 2025/2030/2035/2040 what % of technical machine learning researchers at leading labs will agree with the statement 'AI takeover is a serious concern"?
- *** (UF) In 2025/2030/2035/2040 will there be any leading lab where >75% of researchers disagree with the statement "AI takeover is a serious concern".
- *** (UF) In 2025/2030/2035/2040 what fraction of machine learning researchers with above 2k google scholar citations per year would agree with the statement "AI takeover is a serious concern"?
- – By 2030/2035/2040 will AI safety/x-risk be taught in >25% of high schools in the US/EU/China?
- ** If the 2022 AI Impacts survey is re-run in 2029, what % of respondents will give an over 5% chance to AI causing human extinction by 2100? (https://aiimpacts.org/what-do-ml-researchers-think-about-ai-in-2022/).
- *** (UF) If the AI impacts survey was re-run tomorrow on the original researchers with enough monetary reward to guarantee a 100% response rate, what fraction of respondents would agree there is an >5% chance of human extinction from AI by 2100?
- – Suppose a perfectly representative survey of ML researchers is run tomorrow: by when will they predict we have the technical capacity to automate all labour via AI? (I.e. what is the median year they predict this has over 50% chance.)
- ** Suppose a perfectly representative survey of economists is run tomorrow: how many years do they predict will elapse between it being possible

to automate 25/50/70/90% of labour at a lower cost than having humans do it, and 25/50/70/90% of labour actually being automated in the US?

- – Suppose a perfectly representative survey of [economists/ML researchers/research scientists] is run tomorrow: what % of respondents will give a 20%, 50% and 80% 'chance that the intelligence explosion argument is about right', as defined by the last AI impacts survey: https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/
- – Suppose a perfectly representative survey of ML researchers is run tomorrow: what % of respondents will…
  - – give one of the 3 'yes' answers to the Stuart Russell's problem question from the AI Impacts survey: https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/
  - – answer 'more' or 'much more' to the 'how much should society prioritize AI safety research' question from the AI Impacts survey (https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/).
  - – answer "moderately" or above to 'how much have you thought about when human-level machine intelligence will be achieved: very little, a little, a moderate amount, a lot, a great deal'? (Question taken from: https://arxiv.org/pdf/1705.08807.pdf).
- ** In 2025/2028/2033 what % of respondents in the US/China/EU/Japan/South Korea would answer yes to 'is there a risk of human extinction from AI' in a well-designed/conduct poll?
- – In 2025/2028/2033 what % of respondents in the US/China/EU/Japan/South Korea would answer yes to 'is there a higher risk of human extinction from AI than from climate change?'
- – By 2050 will there be an open letter signed by over 1k academics which claims that AIs are now 'persons' and therefore deserve rights?

Security Questions

- ** By 2028/2035 will any leading lab have had the weights of one of its models stolen?
- ** By 2028/2035 will any leading lab have had the weights of one of its models stolen by a state?
- ** If a US EA org were to spend $50 million by 2025 on trying to improve information security at top labs, how much would the chance of a leading lab having model weights stolen by 2035 diminish?
- ** How large will be the *smallest* information security team at any leading lab in 2025/2030/2035/2040?
- – What will be the average salary of someone working on information security at a top lab in 2025/2030/2035/2040?

- ** By 2025/2030/2035/2040 what proportion of top labs will store the weights for their models only on air-gaped systems?
- ** What is the chance that by end of 2024/2026/2030/2040 someone working in a top lab will be discovered to be a spy, looking to steal company secrets?
- – How many papers on information security will appear at the following machine learning conferences by 2030?: https://en.wikipedia.org/wiki/Conference_on_Neural_Information_Processing_Systems, https://en.wikipedia.org/wiki/International_Conference_on_Machine_Learning, https://en.wikipedia.org/wiki/International_Conference_on_Learning_Representations, https://en.wikipedia.org/wiki/Association_for_the_Advancement_of_Artificial_Intelligence
- ** In 2026/2030/2034/2040 how many dollars will the leading lab that spends the least on security spend on security?
- – By 2028/2033/2040 will an AI model commit theft/fraud by corrupting the dataset used in training a machine learning model, so that it responds in an unintended way to a specific input, and then making use of this flaw? (See p.8 of this: https://arxiv.org/pdf/1905.04223.pdf).
- – By 2026/2030/2034 what proportion of US/EU/Chinese AI labs will routinely watermark their models? (See: https://arxiv.org/pdf/2009.12153.pdf)
- – Will the Chinese government have the know-how to suppress or remove (see sc.5 of https://arxiv.org/pdf/2009.12153.pdf) watermarks from leading labs in the EU/US in 2026/2030/2034?

EA opinion on relevant issues:

- – In 2025/2030/2035, what will the lowest and highest median scores on a 1-5 scale for 'does this lab take safety seriously?' given to leading labs by AI safety researchers at Rethink Priorities and Open Philanthropy?
  – That is: for each lab, we have the researchers rate it, and then we take the median of the researcher scores for each lab: what will the highest and lowest medians be?
- – If EAs in 2038 are surveyed on whether AI takeover is a bigger (i.e. more expected harm) risk than deliberate human misuse of AI, what % will answer 'yes'?
- ** In 2027/2032/2037 will 80,000 hours still think that AI takeover risk is one the top 3 causes to work on?
- – Conditional on no human-level (at least one model or collection of

models can do any intellectual task a human can) AI, will effective altruist funding for alignment work be as high in 2040/2050/2060 as it is now? (Inflation-adjusted.)

AI effects on (non-AI takeover) catastrophic and X-risks in international relations

- *** 25 years after AI is able to automate >[25%/50%/70%] of non-manual labour, what proportion of surviving people who were leading academics in International Relations at the point in time where >[25%/50%/75%] automation became possible, will think that AI has made geopolitics less stable?
  - See the recurring terms section for an operationalization of automating a fraction of labour.

Miscellaneous

- *** (UF, FE) Imagine we push the time by which we develop AI capable of performing all non-manual labour back by 5 years. Then, how does the chance that there will be AI takeover within 100 years of its invention change?
- – By 2030, how many papers published after the end of 2021 challenging the validity of a machine learning benchmark will have over 125 citations on Google scholar?
- ** How much will the world economy grow in the five years after we first achieve the ability to automate 25/50/100% of labour?
- ** By 2035/2040 will there be an AI such that at least 60% of philosophers of mind believe that it's conscious?
- ** What % of alignment researchers think that it's important that we get any future agentic, human-level-or-above AIs to care about the interests of animals as well as humans currently/by the end of 2025/by the end of 2035?
- ** What will be the total value of all semiconductor companies listed on Crunchbase (https://en.wikipedia.org/wiki/Crunchbase) in 2026/2032/2040/2050?
- – How many benchmarks on papers with code will have >10 entries in 2025/2030/2040? (https://paperswithcode.com/).
- – By 2025/2030/2040 will there be a scandal in the US/UK/EU/Australia/Canada/Japan/South Korea involving the use of facial recognition AI for surveillance?
- – 25th/50th/90th percentile guesses for the valuation of Fathom Radiant (https://fathomradiant.co/) in 2027?

- – Will the next big thing in ML (next paper introducing a new model that achieves 1000 citations within 1 year, >5% improvement on multiple benchmarks, etc.) be a scaled up version of currently available architectures?
- *** (FE, RP) What is the total number of EAs in technical AI alignment
  - – Across academia, industry, independent research organizations, ¿government?, etc.
  - – See The academic contribution to AI safety seems large for an estimate from 2020.
- *** (FE, RP) What is the total number of non-EAs in technical AI alignment?
  - – Across academia, industry, independent research organizations, ¿government?, etc.
- – Total number of EAs in AI governance (academia / industry / independent research organizations / think tanks / government)
- – In 2024/2029 what will be the monthly revenue of the most profitable AI model in the world, as percentage of the training cost of that model?
- – How many million square cms of semiconductor chips will there be worldwide in 2026/2030/2038? (Not counting those already and used and thrown away.)
- – Chance by 2030/2035/2040 that an NPC in a videogame exposes game company's trade secrets.
- – Chance by 2030/2035/2040 that an NPC in a videogame tells players how to cheat or hack the game.
- – Chance by 2030/2035/2040 that an NPC in a videogame goes on strike.
- – When, if ever, will Twitter or Facebook bans or suspensions reach 10x of their 2022 rate? (Relevance to AI: sign of large increase in the number of propaganda bots).
- – On January 1, 2027, a Transformer-based model will continue to hold the state-of-the-art position in most benchmarked tasks in natural language processing.
- – Will PASTA AI (as defined here: https://www.cold-takes.com/transformative-ai-timelines-part-1-of-4-what-kind-of-ai/) solely use self-supervised learning for training?
- – How many labs will falsely claim to have produced "artificial general intelligence" by 2025/2030/2040? (As a rough guide, the claim is false if there's an intellectual task humans can do that the AI can't.)

- – By what year will an >100 karma LessWrong (https://en.wikipedia.org/wiki/LessWrong) post turn out to have been >90% written by an AI?
- – How much, as a % of its size in 2022, will the machine translation market grow by 2025/2030?
- – What will be the rate of growth in the market for "romantic" chatbots in 2024-2028?
- – What will be the rate of growth in the market for AI call-handlers in 2024-2028?
- – Will someone be awarded a Nobel Prize (not counting literature) for AI-related work by 2030/2040/2050?
- – Year by which at least 10/50 nations mention AI in military strategy white papers?
- – What will be the growth rate of the $s spent on AI copywriting 2024-2029? (https://en.wikipedia.org/wiki/Copywriting).
- ** Will the first AI capable of performing/learning any intellectual task at at least the level of a normal college-educated US adult be trained using gradient descent?
- – Will romantic chatbots surpass $1B/year of revenue before self-driving cars pass $10B/year?
- – In what year (if ever) will whole brain emulation for humans be achieved?
- – When will a whole brain emulation of a small mammal (as defined by Wikipedia) be uploaded? (I.e. >80% of neuroscientists and biologists at Ivy League plus Oxbridge say it has been achieved.)
- – In 2025/2027/2030 what proportion of AI papers uploaded to ArXiv will include authors from both industry and academia?
- – Between 2024 and 2029 what proportion of AI researchers in the US will move between academia and industry?
- – In 2027 what % of a panel of people who work on "AI Ethics" will agree with the following statement: 'NeurIPS requiring impact statements has helped reduce the negative impacts of AI on society'? (https://statmodeling.stat.columbia.edu/2020/12/21/the-neurips-2020-broader-impacts-experiment/).
- ** When (if ever) will a nonhuman animal using a brain-computer interface perform economically productive labour?
- – By 2026/2030/2035, what will be the loss in $s from the most costly theft/fraud carried out by corrupting the dataset used in training a machine learning model, so that it responds in an unintended way to

a specific input, and then making use of this flaw? (See p.8 of this: https://arxiv.org/pdf/1905.04223.pdf).

- – Chance that by 2030, this search query or its descendant will uncover 100 published academic papers with obvious GPT-written sections: https://scholar.google.co.uk/scholar?hl=en&as_sdt=0%2C5&q=%22Regenerate+response%22+-chatgpt&btnG=

- ** By 2026/2030/2036 will at least 2 leading labs collaborate on a project with a budget of at least $8/80/300 million dollars?
- *** (FT) When will a model trained using 2/4/6 OOMs more compute than GPT-4 be downloadable by the general public?
- *** (FT) What will be the estimated size of the largest publicly downloadable model in 2023/2025/2030/2050?
- ** In 2025/2028/2035 what proportion of the 10 largest models trained that year will have their hyperparameters published? (See: https://forum.effectivealtruism.org/posts/KkbEfpNkjNepQrj8g/publication-decisions-for-large-language-models-and-their).
- *** Will molecular nanotechnology of the kind proposed by Drexler have been proven feasible by 2040/2050/2060?
  - Operationalization: According to the [resolution council]'s, best judgment, there will be an MVP demonstrating the kind of "dry" nanotechnology of the kind proposed by Drexler, i.e., machinery that can mechanically manipulate individual atoms with precision without the need for a water medium.
  - Reference: https://en.wikipedia.org/wiki/Drexler%E2%80%93Smalley_debate_on_molecular_nanotechnology
  - Note:Could also have different levels, e.g., a) there are good indications to think that…, b) there is an MVP, c) there are industrially useful applications of…
  - Note: One could also have a simple version of a https://en.wikipedia.org/wiki/Technology_readiness_level.
  - Relevance: some of Yudkowsky's stories about how an AI takeover could actually occur seem to rely on the idea that very powerful nanotechnology is possible. See for example pp.25-7 of this: https://intelligence.org/files/AIPosNegFactor.pdf
- *** (DG) What fraction of US academics with over >200 citations for their work on nanotech would give an >70% chance that Drexler will be vindicated eventually?
- *** (FT, DG) By 2030/2040/2050, how much will the answer to the previous question have increased/decreased?
- ** Conditional on Drexlerian molecular nanotechnology being physically possible to build, will the story told by Yudkowsky

(https://intelligence.org/files/AIPosNegFactor.pdf p.26) about how an AI could get such nanomachines built just via internet access, normal commercial purchases, and the cooperation of a single human being actually feasible?

- *** (RP) How likely is it that an AI could get nanomachines built just by making ordinary commercial purchases online, and obtaining the cooperation of <30 human beings without scientific skills above masters degrees in relevant subjects?
- ** Conditional on either of the previous two questions resolving yes, how many FLOPs of training run and how much data would an AI need (given current/2030/2040 state of the art training) before it was actually smart enough to pull such an operation of, given 2022/2030/2040 levels of scientific knowledge relevant to nanotech being accessible online?
- *** (UF) If an AI actually did manufacture Drexlerian nanotech using the cooperation of only small number of people, would that really scale, within 6 months, to such an AI being able to wipe out all humans if it wants to?
- ** What is the chance that 35% of US factories will have all manual labour performed by robots in 2030/2040/2050, conditional on no regulations being passed designed to prevent this happening?
  - Relevance: If robots can't replace us as workers, than an AI that kills off all or most humans is threatening its own survival, since it needs people to maintain an electrical grid etc.
- *** (UF) How many AI run factories, without on-site human supervision, will there be in the US/China/EU/Japan/South Korea by 2037/2042/2050?
- ** How many AI run factories, without on-site human supervision, with purely robot on-site workforces in the US/China/the EU/Japan/South Korea by 2040/2045/2055?
- ** By 2030/2035/2048 will it be possible to create a pathogen that, if released, would kill >70% of the world's population? (Relevance: this is one way that an AI could kill of most people on Earth as part of a takeover plot.)
- ** Assuming the previous question resolves yes at some point, will the means to create such a pathogen theoretically be legally and commercially available in at least one country, for at least 6 months, in the 10 years after it resolves yes?
- ** Assuming the question before last resolves yes at some point, at its cheapest in the 20 years after it resolves yes, how expensive will getting the equipment needed to create and release such a pathogen be?

- – Intentional misuse by non-state actors: how many times will proposals to kill >100 million people by assigning a goal to an AI agent be posted on internet forums for extremist ideologies (think: white nationalists who want to commit genocides against other racial or ethnic groups, violent cults like Aum Shinrikyo https://en.wikipedia.org/wiki/Aum_Shinrikyo etc.) by 2026/2030/2038? (The 'dark web' counts as the internet.)
- – Intentional misuse by state actors: By 2030/2035/2042 will any one of China/Russia/The US have a written military plan detailing how they could use AI to completely destroy one of the other 2 states?
- *** (UF) Condition on no AI takeover. Then, if >97% of labour is automated in the US by 2100, what will the median US human annual income be 10/20/50 years after reaching >97% automation?
  - – Relevance: This is an imperfect proxy for 'AI is aligned to its makers, but things are dystopian because without the bargaining power that comes from their labour being useful, most humans are disempowered and impoverished.

Acknowledgments

This list of forecasting questions was started by David Mathers, Gavin Leech and Misha Yagudin, of Arb Research and completed by Nuño Sempere, of Shapley Maximizers. Open Philanthropy provided funding.