IAPS | Institute for AI Policy and Strategy

MAY 2025

The US Government's Role in Advanced AI Development: Predictions and Scenarios

Bill Anderson-Samways and Oscar Delaney

Abstract

There has been significant recent speculation about whether the US government (USG) will lead a future project to build and acquire advanced AI, or continue to play a more arms-length role. Such speculation warrants rigorous assessment, given its implications for research priorities and geopolitical dynamics.

We conducted a forecasting workshop on this question, employing the IDEA protocol ("Investigate, Discuss, Estimate, Aggregate") to elicit predictions from six professional forecasters and five experts on US AI policy. This included presenting participants with reference-class data addressing whether past analogous innovations (e.g., the computer, the atomic bomb, and historical AI breakthroughs) were developed via USG-led or purely private projects.

On average, participants gave a 34% median probability that a USG-led project builds and acquires the first AI system capable of accelerating AI R&D tenfold, which we call *AIR-10*. Participants also estimated the probability that AIR-10 is developed via several granular scenarios for a USG-led project: the USG leads a consortium of laboratories to build the system (14%), the USG engages a single private contractor to build the system (9%), a USG laboratory directly builds the system (4%), the USG nationalizes a company midway through building the system (4%), and the USG uses legal means to compel a company to build the system (3%). Conditional on a USG-led project developing the first AIR-10, participants estimated a 46% probability that it would be controlled by the US military or intelligence community (e.g., the Department of Defense or National Security Agency), as opposed to a civilian agency (e.g., the Department of Energy).

However, participants also expressed considerable uncertainty. The central 34% forecast, for example, has a 90% confidence interval of 11-61%. Therefore, AI policy development efforts should span a portfolio of strategies that ensures preparedness across both USG-led and purely private development scenarios.

Executive Summary

- **We ran a structured 3-hour workshop to forecast whether a US government (USG)-led project will build and acquire the first AI system that accelerates AI R&D tenfold (AIR-10).** Participants consisted of six professional forecasters and five experts on US AI policy.
    - We defined a "USG-led project" as one where the USG both (a) decides whether to start and stop AI training and (b) acquires the final model for government purposes.
    - The workshop followed the IDEA protocol, a variant of the Delphi technique.
- **Workshop participants predicted, on average, a 34% median probability that a USG-led project builds and acquires the first AIR-10**, conditional on AIR-10 first being developed in the US by 2035. The average prediction among forecasters was 28%, compared to 40% for subject-matter experts.
    - There was **considerable uncertainty in these predictions**, with participants expressing an average 90% confidence interval of 11-61%.
    - The most likely *scenario* for a USG-led project building and acquiring AIR-10 was thought to be a **government-led consortium (14%)**, followed by a **single private contractor (9%)**, a government lab (4%), a nationalized private company (4%), and legal compulsion of a private company (3%).
    - Conditional on a USG-led project building and acquiring AIR-10, participants estimated a **46% probability of military or intelligence community control over the project** (versus 54% for a civilian department such as the Department of Energy).
- To forecast complicated questions, it is useful to look at past data covering similar cases. **While AI is unique, some historical analogs can still be informative (e.g., the computer, the atomic bomb, and historical AI developments).**
    - We provided workshop participants with a dataset of 35 past US technological innovations, categorized into five reference classes, and analyzed whether each innovation was developed via a USG-led or purely private project. The proportion of USG-led projects within each class was as follows: general-purpose technologies (40%), ambitious STEM projects (63%), dual-use technologies (57%), megaprojects (78%), and past AI developments (23%).
    - The most common forms of USG-led projects were consortia and private contracts.

- Our qualitative analysis of forecasters' and experts' statements suggests the following key considerations for whether a USG-led project builds and acquires the first AIR-10:

    1. **National security** factors, especially the military-technological threat of China, may compel the USG to launch a project to build and acquire AIR-10.
    2. **Historical precedents** show the USG has often been the driving force behind high-risk, ambitious technology advancements.
    3. **Rapid AI development,** combined with the AIR-10 threshold's lack of attention-grabbing features, may prevent the US government from acting quickly enough to launch a successful project.
    4. **Political considerations,** such as pro-market principles or a lack of in-house expertise, may lead the USG to favor private-sector development.
    5. **Budgetary constraints** make large appropriations for AI development difficult.

- The relative probability of different *scenarios* for a USG-led project was most impacted by point (4) above: both ideologically and practically, the USG would likely prefer less coercive options enabling it to draw upon a wide range of private-sector talent.
- **The workshop had little impact on participants' predictions:** their final estimates remained closely aligned with their initial estimates (R^2=0.9). In other words, discussing these questions with other forecasters and experts mostly did not move participants significantly away from their initial beliefs. Combined with participants' wide confidence intervals, this suggests that **there is considerable, difficult-to-resolve uncertainty around whether a USG-led project will build and acquire AIR-10.**
- **Therefore, AI policy development efforts should span a portfolio of strategies** to ensure preparedness across both USG-led and purely private development scenarios.

Table of Contents

Background and literature review

Many researchers and forecasters think that AI R&D could be considerably or completely automated within a decade.[1] This could lead to a rapid acceleration in AI capabilities,[2] such that within a few years of R&D automation, AIs could perform nearly all remote work tasks, and/or provide military-strategic capabilities that are decisive in any conflict. As such, AI R&D automation is an important target for forecasting work.

Various prominent predictions have been made regarding a USG-led project to build and acquire powerful AI systems. Notably, Aschenbrenner (2024) claims that by 2027 or 2028 there will be "some form of government AGI project," where AGI is defined similarly to AIR-10: a system that can "compress a human-decade of [AI] algorithmic progress into less than a year". Similarly, Kokotajlo et al. (2025) predict that by 2027, the USG will have entered into a close partnership with the leading US AI developer, involving government control over key training and deployment decisions. Conversely, the forecasting platform Metaculus (2025a) currently puts the likelihood that "transformative AI" is developed via a "government project" at just 13%.

However, it is difficult to assess the credibility of these claims regarding USG-led projects. No existing probabilistic predictions on this topic make use of structured forecasting methods or expertise specific to the US government. Cheng and Katzke (2024) thoughtfully assess the plausibility of different scenarios, but the analysis is deliberately non-probabilistic, drawing upon the authors' intuitions rather than structured protocols involving a wider group of forecasters or experts.

Our research aims to build on this work, using best-practice forecasting methods to estimate the probability of a USG-led project building and acquiring AIR-10. Numeric predictions, though still subjective, are more precise than qualitative statements such as "plausible," "likely," or "unlikely," which imply very different numeric probabilities to different people (Friedman et al. 2018). Additionally, group forecasts have been shown to be a better predictor of future events than the forecasts of any one individual (Hemming et al., 2018).

---

[1] AI companies are working on automating AI R&D, with coding assistants already proving very useful for software engineering tasks (Pichai, 2024). Current AI agents are more capable than human experts at a range of AI R&D tasks over a 2-hour time frame, but humans perform better over longer time horizons (Wijk et al., 2024). Especially time-intensive parts of the AI R&D process, such as implementing, debugging, and analyzing experiments, will likely be automated first (Owen, 2024).

[2] Expert views are mixed as to whether partial or complete automation of AI R&D will rapidly lead to strongly superhuman AI capabilities (Erdil and Besiroglu, 2023; Erdil, Besiroglu, and Ho, 2024; OECD, 2023), and if so on what timeline (Owen, 2024), but this seems very plausible.

Better forecasts can help shape research allocation around downstream questions, for example whether to conduct research on policies to improve the security and safety of a USG-led or private AI project. Policy researchers are already making implicit bets about the plausibility of a USG-led project: several authors (Zelikow et al. 2024; Katzke and Futerman, 2024; Aschenbrenner, 2024) all assume that a USG-led project is quite plausible and make recommendations accordingly.

Our research can also help avoid overconfident claims about the imminence or implausibility of a USG-led project. Such claims risk creating a "self-fulfilling prophecy," for example leading other nations to launch their own projects, which could spark an arms race (Hendrycks, Schmidt and Wang, 2025).

Methodology

Selection of method

In our forecasting workshop, we relied on the IDEA protocol—an expert elicitation process based on the Delphi method. Expert elicitation is a useful substitute for complex questions where strong data do not already exist (Hemming et al, 2017). Whether a USG-led project will build and acquire advanced AI is one such question.

We explain the IDEA protocol in more detail below. However, our reason for selecting the IDEA protocol over the original Delphi method is that the former does not require experts to arrive at a consensus probability, which would likely be implausible for such a complex question, and would give a false sense of certainty. Like Delphi, however, IDEA still allows participants to discuss and update their probabilities, which generally results in greater accuracy than a simple survey (Hemming et al, 2017). Though the IDEA protocol is too new for its long-run accuracy to be assessed, there is some limited evidence that the Delphi method can be accurate for forecasts over decades-long time periods (Parente and Anderson-Parente, 2011; Ono and Wedemeyer, 1994).

Functionally, our process drew on many insights from "superforecasting" (see Tetlock and Gardner, 2016). We deliberately included many participants with strong forecasting track records. We also presented participants with reference-class data (described below), which superforecasters frequently use. We chose to supplement our method with superforecasting techniques because superforecasters demonstrably produce better predictions than subject-matter experts over timescales of one year (Tetlock and Gardner, 2016), and there is some limited evidence of superforecaster accuracy over longer time-periods (Tetlock et al., 2023).[3] However, forecasters are

[3] However, it is difficult to assess the relevance of this evidence to our specific forecasting question.

unlikely to possess deep knowledge on all of the factors relevant to our question; as a result, we also included US AI policy experts in the workshop.

Forecasting definitions and questions

Before undertaking the workshop, we drafted our forecasting questions and key definitions, and received several rounds of feedback from AI policy researchers and forecasters.

We ultimately settled on the following conceptualization of advanced AI systems:

> *We define a **"10x AI R&D multiplier" AI system (hereafter, AIR-10)** as a system that accelerates 10x the research progress of leading AI R&D teams focused on algorithmic improvements and novel architectures (as opposed to teams focused on engineering large training runs, or designing improved hardware). In principle, this could be determined by randomizing R&D teams to either spend 10 weeks working unaided by post-2022 AIs, or one week working with the candidate AI system, and having a blinded panel of judges assess each team's research output quality.*[4]

Our definition matches well with this comment in *Superintelligence Strategy* (Hendrycks, Schmidt, and Wang, 2025, p.11):

> *Even if an intelligence recursion achieves only a tenfold speed up overall, we could condense a decade of AI development into a year. Such a feedback loop might accelerate beyond human comprehension and oversight. With iterations that proceed fast enough and do not quickly level off, the recursion could give rise to an "intelligence explosion."*

As previously mentioned, our AIR-10 definition also fits with Aschenbrenner's definition of AGI in *Situational Awareness* (2024).

We considered many other options for defining advanced AI systems. However, we ultimately decided on AIR-10, as other posited forms of "advanced AI" were either too vaguely defined for forecasting purposes (e.g., AI Impacts, 2022; Open Philanthropy, 2016; Karnofsky, 2021), may not correspond to an AI system that is highly transformative (e.g., Metaculus, 2025b; OpenAI, 2023), or captured only *one* dimension of advanced AI systems' impact, such as military or economic impacts (e.g., Metaculus, 2025c). Our conceptualization of advanced AI is somewhat similar to

[4] In practice, such a test may never be done, in which case the question would resolve based on the subjective judgment of a panel of relevant AI experts familiar with the candidate system, who could use nonrandomized data about the relative pace of progress before and after each research team gains access to the system.

Anthropic's (2025) notion of an AI system that accelerates twofold the rate of effective compute growth from 2018-2024, but employs a higher and more specific bar for acceleration.[5]

We also specified our definition of a "USG-led project". Increased USG involvement in AI development could take many forms, from light-touch security assistance to directly building frontier AI systems within a national lab. However, we focus on a binary division of whether the USG builds and acquires the first advanced AI system:

> ***Specifically, we define an AI system[6] as being built and acquired by a USG-led project[7] if the USG both:***
>
> - *Decides whether and when to start/stop developing the AI model.[8]*
> - *Acquires the final model, and thus decides how (or if) to deploy it.[9]*

This is quite a stringent definition, meaning that other forms of USG involvement in frontier AI development—such as the USG acquiring AIR-10 after it is developed, developing a later, more powerful system, or implementing strong regulations and oversight—would not count.[10]

---

[5] We decided to raise the bar to a tenfold acceleration to ensure that the AI system is truly transformational, and we excluded compute growth, instead focusing just on algorithmic progress, which has faster feedback loops.

[6] For assessing whether a USG-led project has built and acquired an AI system, the most critical subsystem or component (the foundation model, in the current paradigm) will be considered. For instance, if the USG developed and acquired the foundation model, which is scaffolded to use an off-the-shelf AI code linter built by a private company for some tasks, the USG would still have "built and acquired" the overall system.

[7] The USG includes any set of one or more federal government agencies (e.g., DoD + DoE, but also novel federal agencies), Congress, and the President.

[8] This does not necessarily require government involvement in day-to-day decisions in the training process, e.g., which datasets and hyperparameters to use. It is sufficient for the USG to have official decisionmaking authority over when to start/stop the training process, even if a private company actually implements the training.

[9] The USG might not have the only copy of the model, e.g., a corporate contractor that developed the model may also have a copy. But the USG must be able to deploy its copy as it sees fit, rather than just regulating the corporate partner's deployment. It is not necessary that all parts of the USG gain access to the model.

[10] For instance, neither of these hypothetical scenarios would meet the criteria in our definition:

A. The White House is sent a daily report on the training run of a leading AI company, and can force training to stop if the President is concerned about safety or security failures. The President does not intervene, and the company trains and deploys their advanced AI unhindered.

B. The USG is not involved during the model's training, but there is a law stipulating that if a company trains a model with certain (e.g., militarily important) characteristics, they must share a copy with the USG.

"A" does not meet the second criterion of USG acquisition, and "B" does not meet the first criterion of USG-led development. A combination of these scenarios could meet our overall definition.

We selected this definition because AI risks could arise both during development and during deployment (Delaney and Acharya, Forthcoming). Thus, USG influence over both stages is important (Hendrycks, Mazeika, and Woodside, 2023). We considered alternative definitions, but these were not suitable for forecasting purposes: they were either too vague or understood the existence of a USG-led project as a continuous rather than a binary variable (which would have made eliciting numeric forecasts difficult).

Using these definitions, the overall question that we asked participants was as follows:

> ***Imagine that AIR-10 is first developed in the US by December 31st, 2035.[11] What is the probability that this system is built and acquired via a USG-led project?***

We chose to condition on AIR-10 being developed by 2035 because nearer-term AI development scenarios are more action-guiding for current policymaking. Additionally, it is more feasible to forecast scenarios where the geopolitical and technological landscape is broadly similar to today; longer-horizon forecasts are more fraught.

We were also interested in the *form* that a successful USG-led project to build and acquire advanced AI would take. Moreover, breaking down a forecasting question into smaller sub-categories can lead to improved accuracy (Tetlock and Gardner, 2016). In our taxonomy, there are five mutually exclusive and collectively exhaustive (MECE) scenarios for a USG-led project, shown in Figure 1 (p. 11). For this, we introduced another definition:

> *A **"project"** comprises the beginning of the training run, through any post-training enhancements, to the system reaching AIR-10 level.[12]*

For each scenario, we instructed participants to "Imagine that AIR-10 is first developed in the US by December 31st, 2035. What is the probability that the system is built and acquired via the following subforms of a USG-led project?":

1. ***Government-led consortium**: The USG coordinates multiple labs in a centralized project.[13] This can include consolidating multiple labs into a single unified entity.*

---

[11] An AI system is said to be **developed in the US** if the primary decision-makers for the AI project are US persons (natural persons or companies) or the resulting intellectual property is owned by a US entity. This may be despite the physical data centers used for AI training being partly or fully outside the US, or the funding for the project coming from a mixture of countries.

[12] If AI training works very differently by the time AIR-10 is developed, such that this conception of discrete models being trained makes less sense, a "project" will simply mean the final concerted development push that eventuates in AIR-10, whatever that corresponds to in the new paradigm.

[13] It does not matter whether the labs are government-owned or privately-owned, e.g., if one government-owned lab, one nationalized lab, and one private lab under government contract all collaborated in a USG-led project, this would count as a "consortium". The category covers any project with two or more labs, regardless of the form of government influence over each constituent lab.

2. ***Government lab**: The project occurs from beginning to end in a single government lab.*[14]

3. ***Nationalization:** The project begins at a single private lab, but is transferred to and then completed at a single government lab.*[15]

4. ***Private contractor:** A single private lab voluntarily enters into a contract with the USG to develop AIR-10; the project occurs from beginning to end in the private lab.*[16]

5. ***Legal compulsion:** A single private lab is compelled by the USG to develop AIR-10; the project occurs from beginning to end in the private lab.*[17]

As a rough consistency check, we asked participants during the workshop whether they thought that any other scenarios needed to be added in order to make our breakdown MECE; they did not. We can moreover sum the forecasts for the five scenarios, and compare this to the overall forecast for each participant. Of the 11 participants, seven had an exact match between the sum of these scenarios and the overall forecast, three had a discrepancy of one percentage point, and one had a discrepancy of eight percentage points. So overall, participants agreed and understood that these subforms of a successful USG-led project sum to the overall likelihood of a successful USG-led project.

---

[14] A "government lab" is a lab that is majority-owned by a government agency, or where government officials comprise the majority of the lab's management or board of directors. For example, the Naval Research Laboratory is both owned and operated by the Department of the Navy, whereas Lawrence Livermore National Laboratory is owned by the Department of Energy but operated by a private sector partner; both qualify as "government labs".

[15] The private → public transfer could take several forms, such as the whole private lab being incorporated into the USG, just the AIR-10 project team and resources, or just the in-development system's model weights.

[16] This could take the form of a tender or procurement process where the USG rigorously defines the system's specification, and different companies bid for the contract. Alternatively, it may look more like a collaborative public-private partnership between the government and a (single) AI company to design, build, and possibly finance the AI system together, as long as the overall "USG-led project" definition is still met.

[17] The USG must use some legal instrument to compel production, such as Title I of the Defense Production Act. If the USG uses soft power or various forms of pressure to incentivize the company to sign a contract that is favorable for the USG, this would still count as a company contract, not forced production. Again, the overall "USG-led project" definition must still be met in order to count.

Moreover, the line between "compulsion" and voluntary "contracting" of a private company is sometimes blurry. For example, under the prioritization power granted by Title I of the Defense Production Act, the President can compel companies to prioritize the fulfillment of existing government contracts. For our purposes, we consider that the form of government project control should resolve as "private contractor" if the company willingly entered into the initial contract with the government (even if the government then forced the company to prioritize it); the scenario should resolve as "legal compulsion" if the government used its powers to force the company to accept the initial contract.

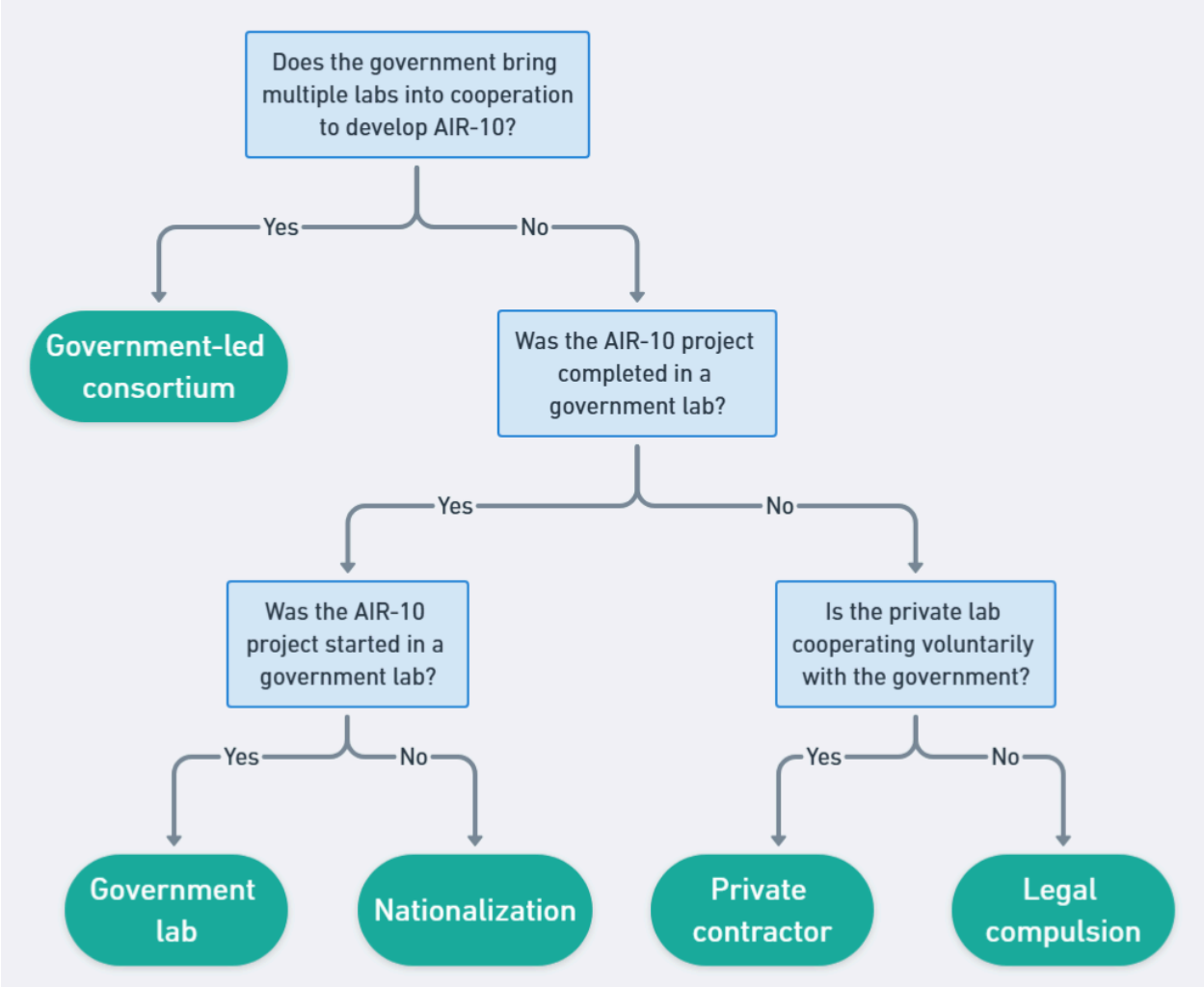

Figure 1: The relationship between different scenarios for a USG-led project to build and acquire AIR-10.

Finally, we were interested in whether the military/the intelligence community (IC) will control AI development, as opposed to a civilian agency such as DOE. As a result, we asked:

> *Conditional on a USG-led project building and acquiring the first AIR-10 system by December 31st, 2035, what is the probability that said project is primarily controlled[18] by the*

[18] By "primary" control we mean that the organization in question takes the majority of decisions regarding (a) whether and when to start/stop developing the AIR-10 system and (b) whether and how to deploy the resulting AIR-10 system. For example, even if the President and Congress authorize a project to construct and deploy AIR-10, and even if DOE is involved via providing specialized compute or data, if DoD officials take the majority of decisions about when to start/stop development and how to deploy the resulting system, then DoD is said to exercise "primary control".

Department of Defense or a member of the Intelligence Community (as opposed to a different part of the USG)?[19]

For all of the above questions, we asked participants for their median probability and 90% confidence intervals. The idea of an individual possessing a 90% confidence interval over the probability of a one-off event is difficult to operationalize, but we chose to define it as follows:

> *Imagine that there was an idealized superforecaster who was perfectly rational, with all available information one could gather on a particular topic and the best possible ways of interpreting that information. What is the interval within which you think their answer would fall (90% confidence)?*

This definition encourages participants to think about new, counterfactual information that could possibly change their minds, and alternative ways of weighing that information. As such, it corresponds to their epistemic uncertainty about their forecast. We factored this operationalization into our calibration training (see Forecasting workshop section).

The full handout of background information, definitions, and questions that we provided to participants is available here.

Workshop participants

We were interested in participants with some combination of the following backgrounds and areas of expertise:

- Knowledge of the US government, especially the Department of Defense (DoD), the Intelligence Community (IC), and the Department of Energy (DOE).
- Knowledge of AI/AGI development pathways and prospects, with at least one participant being an *expert* in AI development, e.g., resourcing requirements.
- A strong forecasting track record.

We assembled a longlist of 82 potential participants by searching for authors of relevant publications, think tank staff with relevant experience, and former government officials. We also drew upon snowball sampling and our own professional networks.

We then cut this longlist down to a shortlist of 26 individuals whom we invited to the workshop, using a profile matrix capturing the above-mentioned criteria. Participants were assigned a score according to the number of criteria they fulfilled. However, as our main concern was getting a strong distribution of expertise within the group (as opposed to any one individual), we

[19] Recall that the USG includes any set of one or more federal government agencies (including novel federal agencies), Congress, and the President.

supplemented our list with individuals who performed strongly on certain individual criteria, especially those with a strong proven forecasting track record (as this is the best predictor of forecasting performance, and difficult to find in combination with all of the other desired characteristics).

Of the 26 invitees, 11 accepted and attended the online workshop. This was in line with our target, as more than 6 to 12 participants leads to minimal improvements in group accuracy (Hemming et al. 2017).[20] Of our 11 participants, six were professional forecasters and five possessed expertise on US AI policy. The AI policy experts had prior experience in the Department of Defense (DoD), a Department of Energy (DOE) national laboratory, the Intelligence Advanced Research Projects Agency (IARPA), the Central Intelligence Agency (CIA), and the public sector divisions of two major AI companies. All participants had experience researching or working on AI development or acquisition. As desired, one participant was an expert in technical AI development, including resourcing requirements.

The workshop followed the Chatham House rule, so we are not disclosing the identities of participants (except where they consented to this—see Acknowledgements).

Reference class data

Forecasts are often improved by considering reference class data (Tetlock and Gardner, 2016). A reference class is a set of historical cases that share some relevant feature with the phenomenon of interest, which in this case is AIR-10 (Franklin, 2010). A feature is relevant if it is hypothetically correlated with the outcome variable (Franklin, 2010)—in this case, whether or not a technology is developed by a USG-led project (or via a particular subform of said project).

We identified five reference classes, each of which contained past technological developments that occurred in the United States. Our reference classes were:

- **General-purpose technologies (GPTs; 40% USG-led):** We used a list from Lipsey, Carlaw and Bekar (2005, p. 132) who define a GPT as a technology which "initially has much scope for improvement and eventually comes to be widely used, to have many uses, and to have many spillover effects [across the world economy]".[21] One important feature of AIR-10 is that it will affect many industries, so GPTs are a useful reference class. This

[20] Indeed, one consultant with relevant experience informed us that more than 12 participants leads to *diminishing* returns.

[21] Since we were focused on the US, we excluded GPTs developed elsewhere, in particular all pre-industrial revolution GPTs, and also the steam engine, factory system, electricity, steamship, railway, internal combustion engine, automobile, lean production, and nanotechnology, which were developed internationally.

feature likely reduces the probability of a USG-led project, as the development of such technologies may be very profitable for private actors.

- **Ambitious STEM technologies (63% USG-led):** We used a list from Davidson (2021) of "ambitious but feasible technolog[ies] that a serious STEM field [was] explicitly trying to build".[22] Some theories suggest that where a technology is difficult to develop and therefore the pay-off is more uncertain, governments are more likely to develop the technology than profit-seeking actors (Naudé, Gries and Dimitri, 2024). AIR-10 is certainly an ambitious technology in terms of difficulty, though (as previously noted) it is at least plausible within the coming years, and a large amount of effort is being put into building AIR-10 by serious researchers and engineers.

- **Key dual-use technologies (57% USG-led):** We used a list from Williams-Jones, Olivier, and Smith (2014, p. 79) who define dual-use technologies as developments that "have both beneficial and harmful applications or consequences".[23] Advanced AI systems will have many beneficial and harmful applications, so AI development is a paradigm case of dual-use research. This feature could either raise or lower the probability of a government-led project, as such technologies will interest both public and private actors.

- **Megaprojects (78% USG-led):** We used a database from Potter (2024) and included projects that cost more than 0.1% of US GDP in the year(s) in which they occurred. It is unknown how expensive the first AIR-10 system will be to develop, but rising AI infrastructure investments suggest it may require vast financial resources (Sevilla et al, 2024). One might expect that very costly projects are more likely to be government-led, because the government has greater financial resources than any single private actor.

- **Historical AI developments (23% USG-led):** We used the "notable AI models" dataset from Epoch AI (2025) and filtered these to include only models that were developed in the US, had >1000 citations, and were deemed "historically significant" by Epoch. Future AI progress may be very different from past breakthroughs, but historical AI developments are another relevant reference class for AIR-10. For example, there may be path dependencies wherein the fact that most past AI models have been privately developed makes it more likely that future models will be privately developed as well.

For each reference class, we computed the proportion of cases that fit our definition of being built and acquired by a USG-led project, as well as the proportion that fit each subform of a USG-led project (e.g., consortium, government lab, etc).

---

[22] Again, we excluded non-US-only projects (the steam engine, the periodic table, DNA structure, DNA sequencing, DNA editing with CRISPR, antibiotics and the Standard Model of particle physics).

[23] Again, we excluded non-US-only projects (chlorine gas, sulfur mustard, mousepox viral enhancement, the Human Genome Project, rational protein design, H5N1 host range research, and nanotechnology).

The raw reference-class data is available here. To show the uncertainty in the data, we developed this interactive tool, which allows the user to assign weights to each reference class based on their subjective assessment of its relevance to AIR-10, and provides a resulting uncertainty distribution. The tool uses a simple linear regression (with flat priors) to estimate the effect that each feature has on the outcome variable, which is then used to predict the value that a new technology (in our case, AIR-10) has on the outcome variable. We provided this tool to participants during the workshop (see the next section). If all weights are set to 1—implying that each reference class is equally informative—the tool yields a 59% probability that a USG-led project builds and acquires AIR-10, with a 90% confidence interval of 21-88%.

Forecasting workshop

We ran a 3-hour online forecasting workshop based on the IDEA protocol (Hemming et al. 2017), a variant of the Delphi technique, including our 11 participants, one expert facilitator from the RAND Corporation, and several IAPS note-takers and supporters. The workshop proceeded as follows:

- **Introduction to forecasting and calibration training** run by professional forecaster Peter Wildeford, designed particularly for experts on the USG who were forecasting novices.[24]
- **Clarification on the forecasting definitions and questions.** Participants had already read the questions and definitions handout before the workshop, but we clarified various edge cases and uncertainties during the workshop.
- **Initial individual forecasts and rationales.** Participants provided their personal forecasts and rationales using our forecasting app. Other participants could not see their forecasts or rationales.[25]
- **All participants' forecasts and rationales were revealed** to each other (using pseudonyms), and participants were allowed to update their forecasts in response.
- **Participants were then shown the reference class data** and interactive tool (not included in the pre-workshop materials) and could update their initial forecasts.
- **A facilitated discussion** followed, where participants responded verbally to each other's forecasts and rationales, and discussed any disagreements.

[24] To calibrate experts on the idea of a 90% confidence interval over the probability of a one-off event, we asked them to provide confidence intervals over individual Metaculus forecasts—for example, providing them with the Metaculus question, "Will Iran possess a nuclear weapon before 2026?" and then asking them to provide a range within which they were 90% confident that the Metaculus forecast would fall.

[25] RAND provided a custom R Shiny app for the workshop, but it is no longer online.

- **Participants made their final forecasts**, incorporating any updates based on the discussion, and the workshop ended.

Data analysis

Our analysis and visualization code is available on GitHub here. For most of our analysis, we report the arithmetic mean of participants' median forecasts, as is standard for the IDEA protocol (Hemming et al. 2017). We also provide the arithmetic mean of participants' 90% confidence intervals, as we believe that our definition of these confidence intervals (capturing participants' epistemic uncertainty) is reasonably strong.[26]

As a robustness check, we also employed various more sophisticated statistical techniques (in brief, a beta regression involving only participants' medians, an unweighted aggregation of metalog distributions, and a Bayesian hierarchical model with a flat prior—see the Appendix for details). These did not produce dramatically different results.

For our qualitative analysis, we combined comments written in the forecasting app with those from the verbal discussion, and categorized each relevant comment according to whether it suggested a higher or lower forecast for a given question. In many cases, multiple participants made similar comments, in which case we grouped and thematized these comments.

Results

This section presents our results for each question. Most importantly, participants predicted a 34% probability that a USG-led project builds and acquires the first AIR-10 system (conditional on it being developed in the US by 2035). Partitioning by participant type, forecasters estimated a 28% probability compared to 40% for experts.

Figure 2 shows the distribution of forecasts for each question—both for the group overall, and divided by participant type (i.e., expert or forecaster). Participants' individual forecasts are visualized here. All participants gave wide confidence intervals, and similar rankings of scenarios in terms of relative likelihood, but sometimes quite different overall probabilities. The raw data for each participant's forecast, and the extent to which they updated their forecasts throughout the workshop, are available here. The following subsections provide the thematized rationales participants gave for each question. When multiple participants cited similar rationales, we grouped these together. We use parentheses to note the number of participants that cited each rationale.

[26] See the Forecasting definitions and questions section.

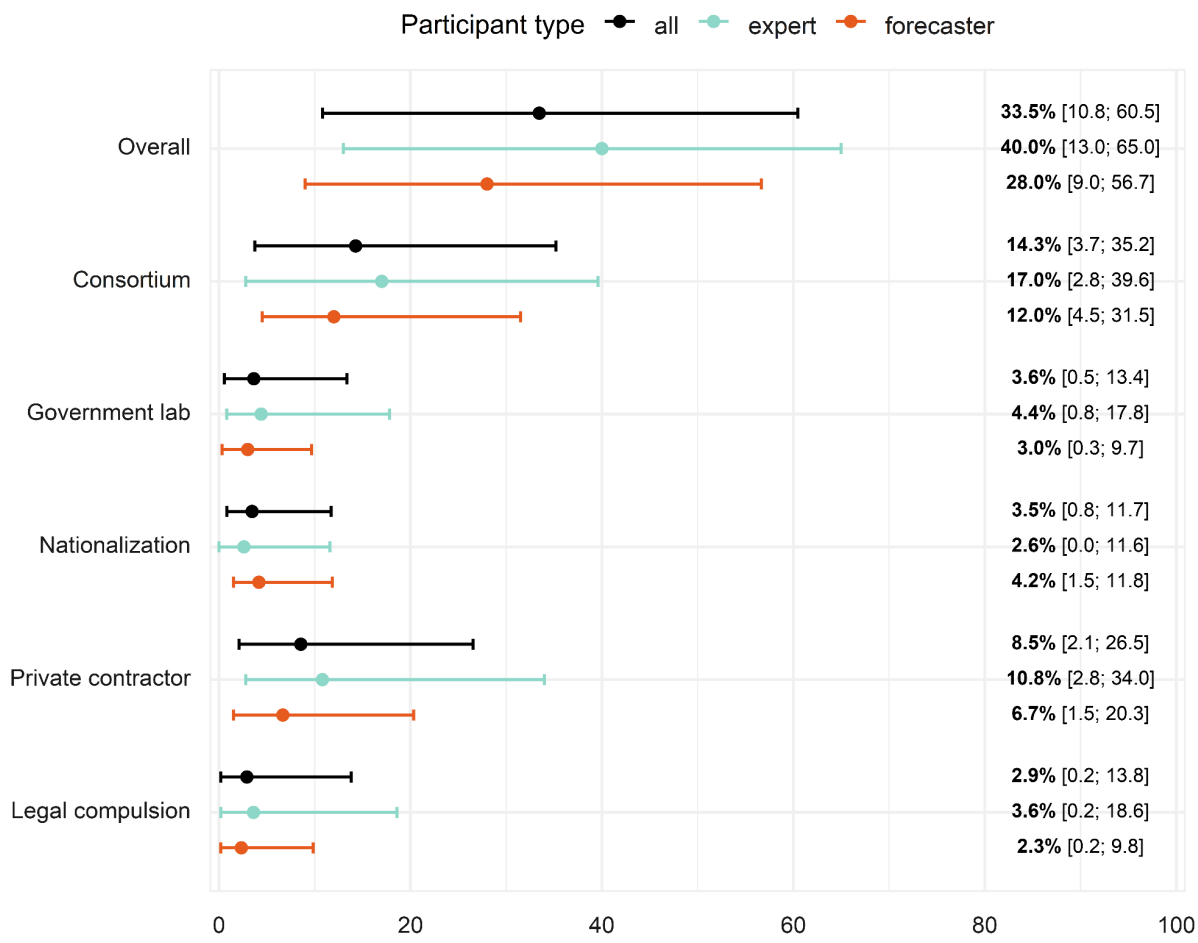

Figure 2: Arithmetic means of participants' forecasts, by question and respondent type.

Overall likelihood that a USG-led project develops AIR-10 (34%)

As a reminder, the question wording was as follows:

> *Imagine that AIR-10 is first developed in the US by December 31st, 2035. What is the probability that this system will be built and acquired by a USG-led project?*

Participants' reasons for:

- (3 participants) National security imperative: Particularly if China were seen to be (close to) matching or exceeding US AI supremacy, a government-led project would become more likely.
    - (1 participant) Already, the U.S.-China Economic and Security Review Commission (2024) has called for a Manhattan Project for AGI in their report to Congress.

- (1 participant) Historically, the USG has been very involved in high-risk tech advancements.
    - (1 participant) However, there have been relatively fewer cases of USG-led technological projects since the Cold War.

Participants' reasons against:

- (4 participants) The USG will likely prefer to play a supporting role and let the private sector continue taking the lead on AI development, for instance because of pro-market principles or a lack of in-house expertise.
- (4 participants) AIR-10 may be achieved in the next several years. If so, it is less likely that the government, which tends to be more slow-moving, will have enacted a plan to build and acquire advanced AI systems before we reach AIR-10.
- (3 participants) The USG is more likely to acquire an AI system once it has already been created, and the USG realizes it is strategically important.
- (1 participant) The USG may not find AIR-10 capabilities attention-grabbing, instead focusing on later capabilities milestones.
- (1 participant) It is unlikely that the USG will want to deploy the financial resources necessary to compete with frontier AI companies.

Participants also noted that a volatile political environment makes forecasting harder, and that elections, particularly the 2028 and 2032 presidential cycles, amplify uncertainty.

Likelihood of subforms of a USG-led project

Government-led consortium (14%)

This scenario was defined as follows:

> ***Government-led consortium****: The USG coordinates multiple labs in a centralized project.[27] This can include consolidating multiple labs into a single unified entity.*

Participants' reasons for:

- (4 participants) The USG will likely not want to "pick a winner" from all of the companies currently on the path to developing AIR-10. Instead, it will prefer to make use of the expertise spread across the private sector, via a consortium.

[27] See the Forecasting definitions and questions section for further information on this scenario.

- (2 participants) "Consortium" is defined very broadly (any combination of two or more labs, whether public or private), which makes this scenario more likely than other, more specific scenarios.
- (2 participants) Historically, this has been one of the most common forms of a USG-led project.
- (1 participant) In most other scenarios, Congress would need to approve massive spending on the development of AIR-10. But a consortium may not require government money, just coordination.
- (1 participant) OpenAI's "merge and assist" clause provides some theoretical basis for AGI cooperation.

Participants' reasons against:

- (2 participants) The current administration may be less inclined to coordinate a large public-private collaboration.
- (2 participants) Rivalries between companies may make working together effectively difficult.
- (1 participant) A consortium may be especially complicated to set up if AIR-10 was developed soon.

Private contractor (9%)

This scenario was defined as follows:

> ***Private contractor**: A single private lab voluntarily enters into a contract with the USG to develop AIR-10; the project occurs from beginning to end in the private lab.*[28]

Participants' reasons for:

- (2 participants) Private AI companies may well favor working with the USG because:
    - They are already trying to build AIR-10, so working with the USG would not jeopardize their mission.
    - The USG would provide funding for the contract.
    - Working with the USG may provide some legal protection.

[28] See the Forecasting definitions and questions section for further information on this scenario.

    - Safety-focused companies would also want to work with the government, in order to have more say over AI development speed and precautions.
    - Companies may worry that not working voluntarily with the USG could lead to more coercive measures at a later date.
- (1 participant) Ideological and practical considerations may lead the USG to favor private development, while providing financing to speed up progress. As noted above, coordinating multiple private labs may be complicated, so a single private lab may be preferred.

Participants' reasons against:

- (2 participants) The USG will not want to "pick a winner" as many companies are competitive in AI development.
    - (1 participant) On the other hand, even if the USG picks one contractor, talent from other companies could then be hired into that leading company.
- (1 participant) Congress would need to grant funding for such a contract, but Congress can be slow to act.
- (1 participant) Private contractors will likely have poorer cybersecurity measures than government (including nationalized) labs.

Government lab (4%)

This scenario was defined as follows:

> ***Government lab**: The project occurs from beginning to end in a single government lab.*[29]

Participants' reasons for:

- (1 participant) The USG could ban AGI research at private companies, and then hire up the spare talent from said companies into a government lab.
- (1 participant) The USG could nationalize a company *before* it begins the final project to develop AIR-10. If this project then succeeds at building AIR-10, this would count as a "government lab" rather than "nationalization" according to our definitions (see the Forecasting definitions and questions section).

Participants' reasons against:

[29] See the Forecasting definitions and questions section for further information on this scenario.

- (4 participants) Government labs are far behind the industry frontier in AI (in compute, and especially talent), which could prevent them from catching up in time to reach AIR-10 first.
    - Moreover, there are no existing efforts by government labs to catch up to the industry frontier.
- (1 participant) There is little historical precedent for government labs being at the frontier of information technology.
    - Moreover, today's government labs are smaller and less well-resourced than they used to be,[30] so a government lab developing such a consequential technology today is even less likely than in the past.
- (1 participant) Government labs are largely funded directly by the USG, and fiscal constraints will limit the amount that the government is willing to spend directly on an AI project.

Nationalization (4%)

This scenario was defined as follows:

> ***Nationalization****: The project begins at a single private lab, but is transferred to and then completed at a single government lab.*[31]

Participants' reasons for:

- (1 participant) A sudden disaster could lead to the USG wanting to act quickly and decisively, which could motivate nationalization.
    - (1 participant) However, the government is likely to be well aware of internal progress at AI companies domestically and (to some extent) internationally, so would likely not be caught by surprise in this manner.

Participants' reasons against:

- (3 participants) The political climate does not support nationalization. The general public and private lobbyists would likely be against it.
- (2 participants) It is more likely that the government would nationalize a company after a training run is completed and the government sees very impressive results, rather than

[30] Here we are simply reporting a participant's claim, without supporting evidence.

[31] See the Forecasting definitions and questions section for further information on this scenario.

during AI development (which is a necessary condition for our definition of "nationalization"). The USG may not realize that a given training run is on track to achieve AIR-10.

- (1 participant) The USG may fear that nationalization would stifle private sector innovation.
- (1 participant) Nationalization may backfire if many employees disapprove and resign in protest.
- (1 participant) Nationalization was absent in the reference class data.

Legal compulsion (3%)

This scenario was defined as follows:

> ***Legal compulsion**: A single private lab is compelled by the USG to develop AIR-10; the project occurs from beginning to end in the private lab.*[32]

Participants' reasons for:

- (5 participants) If a US military adversary develops very advanced capabilities, this option would become more likely as a quick reactive intervention.
- (1 participant) One plausible mechanism, Title I of the Defense Production Act, was used during the COVID-19 pandemic, so there is recent precedent (albeit at a smaller scale).

Participants' reasons against:

- (3 participants) The political climate (including public opinion and private lobbying) does not support legal compulsion currently, and this seems unlikely to change.
- (2 participants) Companies are already looking to build AIR-10, so legal compulsion will not be necessary.
- (1 participant) Employees might move to a different private company or a foreign competitor if they do not want to work for the USG.

Military or IC control over the project (46%)

The question wording here was as follows:

> *Conditional on a USG-led project building and acquiring the first AIR-10 system by December 31st, 2035, what is the probability that said project is primarily controlled by the*

[32] See the Forecasting definitions and questions section for further information on this scenario.

Department of Defense or a member of the Intelligence Community (as opposed to a different part of the USG)?[33]

Participants gave a median of 46% for this question (experts = 53% and forecasters = 40%).[34] A simple average of the reference classes suggests a 77% chance of military/IC control; however, we did not present this information to participants during the workshop.[35]

Participants' reasons for:

- (4 participants) The main reason the USG would want to build and acquire advanced AI in the first place is to protect national security, so housing the project in a national security-focused agency seems likely.
- (3 participants) DoD has a large budget and more bipartisan support for spending.
- (1 participant) DoD has lots of experience with cutting-edge tech R&D (e.g., through DARPA).

Participants' reasons against:

- (1 participant) The USG may be worried about the optics of housing a national project within DoD, so might prefer DOE or a new agency instead.
- (1 participant) Involving DoD/IC might prevent foreign nationals from being involved in the project, which could be a substantial cost given the composition of current leading labs.
- (1 participant) DOE has more high-performance computing infrastructure than DoD, and more experience with certain kinds of public-private partnerships (via running the national labs).
    - If DoD provides assistance to an AIR-10 project housed at a DOE facility, it is unclear whether DoD or DOE would have ultimate control. (Two participants took opposing positions on this point).
- (1 participant) The AIR-10 project may be set up quite hurriedly, which could require going outside of the normal military chain of command.
- (1 participant) The *first* AIR-10 system will likely not be military—it is more likely that DoD would control a subsequent AIR-10 system fine-tuned for military use cases.

---

[33] See the Forecasting definitions and questions section for further information on this scenario.

[34] This question is excluded from the main figures because it employs a different condition (namely that a USG-led project *has* built and acquired advanced AI).

[35] We had not analyzed our reference class data by military/civilian control at the time of the workshop.

Forecast updates during the workshop

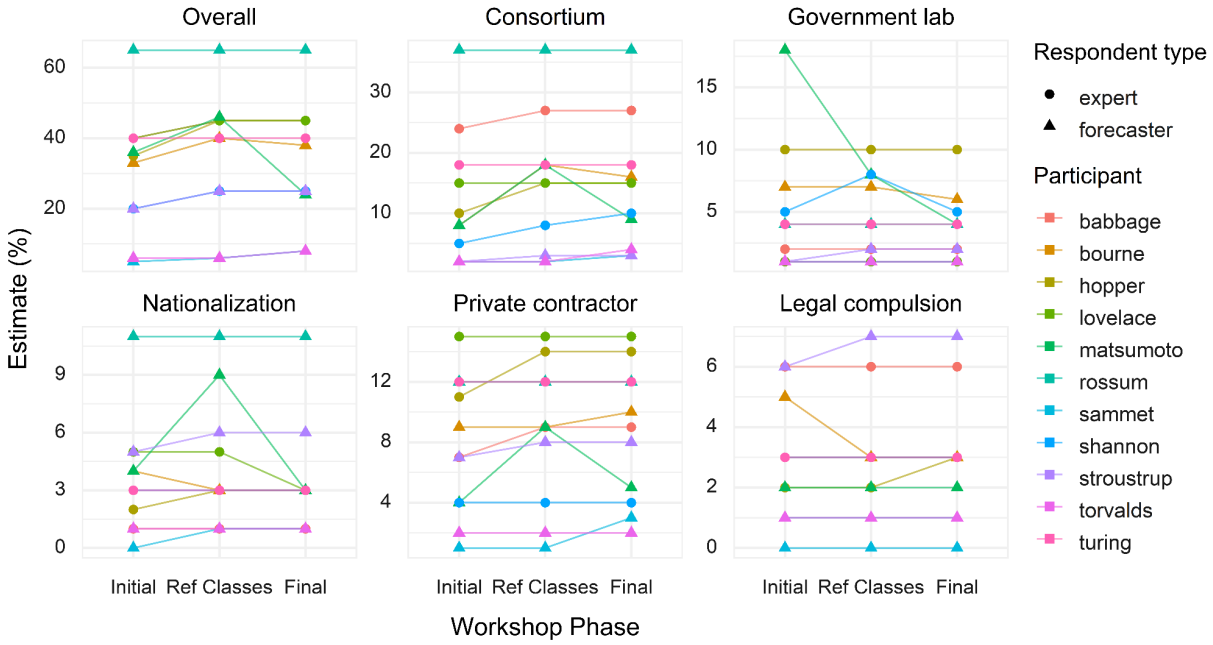

Figure 3: Evolution of participants' forecasts across the three workshop phases by respondent type.

Participants were allowed and encouraged to make updates to their forecasts throughout the workshop, and they did so, with particular spikes in updates when participants were making their initial forecasts and when they were shown the reference class data. Figure 3 shows the evolution of participants' forecasts across three key workshop phases: the initial estimates participants gave while working independently, their estimates after being shown the reference class data,[36] and their final estimates following the group discussion. The raw movements of participants' central estimates (not grouped into three phases) can be seen here.

Participants tended to increase their forecasts when exposed to the reference class data, which accords with the reference classes showing a higher probability than most participants initially estimated. Indeed, based on the written comments people provided when updating their forecasts, we know that eight of the eleven participants updated upwards on the overall question because of the reference class data. Participants did not always mention which exact reference classes were more important for their updates, but of those that did, megaprojects were noted three times, dual-use technologies twice, and ambitious STEM projects once.

[36] At this stage participants also gained access to the forecasts and comments of other participants, so other than by looking at people's text comments we do not know whether updates in this phase were due to other participants or the reference class data.

Participants generally gave fewer and shorter comments when updating their specific scenario forecasts, so it is hard to tell which reference class data, if any, were most useful to them here. However, we can see from their forecast updates that “consortium” is the scenario for which participants updated most, with five participants moving upwards during the reference classes phase, whereas three or fewer participants updated their estimates for the other scenarios.

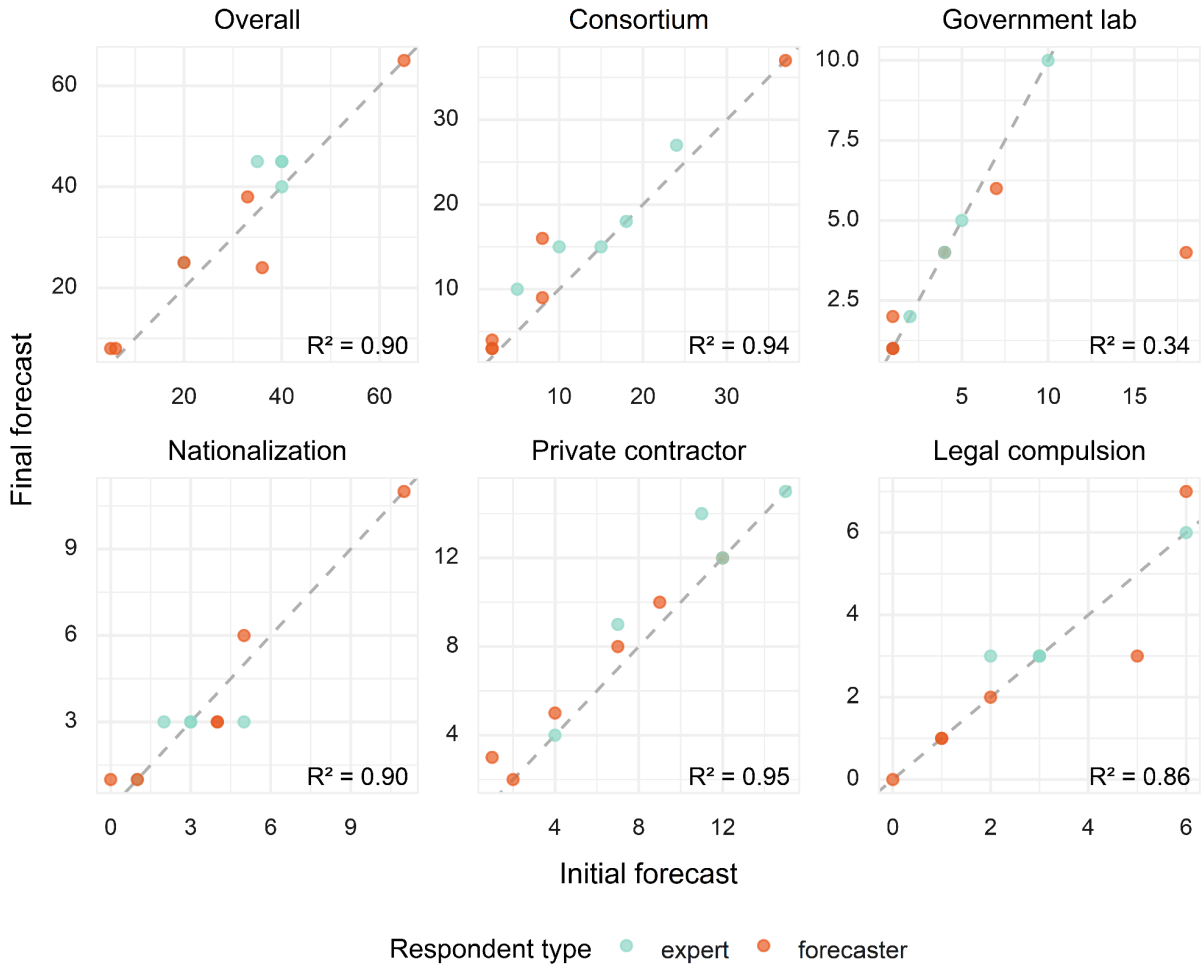

Figure 4: Correlations between the initial and final forecasts of participants, by respondent type.

Overall, however, the workshop seems not to have changed participants' minds much. Neither the reference class data nor interacting with other participants appeared to significantly influence their predictions. Except for the “Military vs. civilian control” question (where several participants initially did not appreciate that the question was *conditioning* on a USG-led project, meaning that their early answers were invalid), participants’ initial forecasts are a strong predictor of their final

forecasts. The overall question has an R^2 value of 0.9 (see Figure 4).[37] All participants made updates to their forecasts during the workshop, but the updates were fairly small on average. This is similar to the findings of the Existential Risk Persuasion Tournament, where forecasters' and experts' views did not majorly converge during the tournament (Karger et al., 2023).

Most participants did not become especially more or less certain during the workshop, with the average width of confidence interval changing from an initial 48.9% to a final 50.1% for the overall question.[38] The other questions ranged from a reduction of 3 percentage points to an increase of 8 percentage points in average confidence interval width.

Figure 4 provides correlations between the initial and final forecasts of participants, grouped by respondent type (i.e., forecaster vs. expert).

Discussion

Key themes

Across all of the questions, several qualitative themes emerge. Each of these themes influenced participants' predictions regarding whether a USG-led project will build and acquire AIR-10, as well as their predictions regarding the shape of said project.

- **National security**: The USG may not trust its geopolitical interests to be adequately served by private AI development. For instance, it may worry about adversaries stealing AI models, or private companies not developing military or strategic capabilities rapidly enough for the government's purposes. More coercive options such as nationalization and legal compulsion become more likely in futures where the US perceives itself to be *losing* an AI race.
- **Historical precedents**: In the past, the USG has played a key role in the development of many strategically significant technologies, often directly developing the technology itself. In addition to suggesting that a USG-led project is more likely, these historical analogs imply

[37] Similarly, the correlation between participants' overall forecasts from the initial stage and the reference class stage was 0.96, and the correlation between their forecasts from the reference class stage and the final stage was 0.86.

[38] Interestingly, the two participants with the lowest final median forecasts started out with quite narrow 90% confidence intervals (2-16% and 0-25%), and by the end of the workshop had considerably wider intervals (0-35% and 1-40%, respectively). Conversely, some other participants' confidence intervals shrunk during the workshop.

that specific scenarios—such as a government-led consortium (along the lines of the Manhattan Project or ARPANET) or a private contractor (akin to Lockheed Martin developing fighter jets)—are more likely.

- **Rapid AI development**: AIR-10 may be developed very soon, perhaps within the next few years. Moreover, the AIR-10 threshold is not necessarily very "attention-grabbing," compared to thresholds around (for example) CBRN development. Given these assumptions, the USG may not initiate a project in time to build and acquire the first AIR-10 system. As such, short AI timelines make it more likely that the USG will acquire an AIR-10 (or more powerful) system *after* it has been developed. This factor also influences the likelihood of different subforms of a USG-led project. If the USG were to build and acquire AIR-10 within the next few years, it would almost certainly need to heavily leverage private sector talent, e.g., via a private contractor or a consortium. It is only if AIR-10 arrives later that the USG could conceivably build up the talent and resources necessary to develop AIR-10 in-house.

- **Political considerations**: USG-led AI projects in general, and more coercive mechanisms in particular (nationalization and legal compulsion), are outside of the political orthodoxy today. For these options to become more likely, radical changes to the political environment would be necessary. On the other hand, future geopolitical or national security events could plausibly provide a motivating shock.

- **Budgetary constraints**: Financial considerations are double-edged. Under the status quo, the large budgets needed for frontier AI development militate against a USG-led project, as Congress may be reluctant to authorize such a project. However, after a major national security shock, government willingness to spend on AI development could increase massively, causing the USG to outstrip private industry's already considerable spending. In particular, DoD's large budget could plausibly support an AIR-10 project—a fact which also makes military/IC control over the project more likely.

Model uncertainty and limitations

Our results should be treated with caution. As noted in the Results section, workshop participants were often highly uncertain about their forecasts. Moreover, *model uncertainty*—in other words, uncertainty about the accuracy of our selected method—provides further reason for caution.

As previously noted, there is some evidence that the Delphi method can be accurate over the timescales on which our forecast operates (namely, years to decades). However, this evidence is limited to two studies (Parente and Anderson-Parente, 2011; Ono and Wedemeyer, 1994). There is also some evidence of individual superforecaster accuracy over longer time periods, but this evidence is even more limited—based on the results of one study employing data that was not intended to assess long-run accuracy (Tetlock et al., 2023). The study also limited itself to

"slow-motion variables with low base-rates of change" (Tetlock et al., 2023, p. 2), which does not seem to apply to our variable of interest (whether a USG-led project builds and acquires a given technology).

Overall, once one accounts for this model uncertainty, the all-things-considered 90% confidence intervals could be even wider than those given in the results section.

Another limitation of our results is that, by necessity, we focused on just one AI capability threshold (AIR-10). We endorse this threshold as an important progress marker, but several participants noted that the USG may care more about later thresholds, perhaps closer to superintelligent systems.[39] Moreover, as noted earlier, participants rightly pointed out that the USG may come to acquire an AIR-10 system *after* it has been developed, which would not meet our stringent definition for a USG-led project (including both development and deployment). Thus, future research directions could consider cases where the USG acquires but does not develop AIR-10, or develops/acquires the first superintelligence but not the first AIR-10.

Strategic implications

As noted above, our results reveal considerable uncertainty regarding whether a USG-led project will build and acquire AIR-10, with large disagreements between participants, and most participants expressing wide confidence intervals. The fact that there was relatively little convergence throughout the workshop suggests that these differences are hard to reconcile following discussion.

The lack of convergence could be used to dismiss the results, with readers instead relying on their personal intuitions. This would be a mistake: if experts and professional forecasters struggle to find agreement, it is likely that the questions posed in this report are simply very difficult to answer. Thus, an intuitive guess is all the more likely to be flawed.

Given the substantial uncertainty around whether a USG-led or purely private project will develop the first AIR-10, AI policy development efforts should span a portfolio of strategies to ensure preparedness across both scenarios. For instance, working to improve the security of model weights and algorithmic secrets will likely be useful regardless of whether AIR-10 is built and acquired by a government-led or purely private project (whereas other recommendations may only be relevant to either USG-led or private projects).

---

[39] The two most relevant quotes are:

- "I don't think the AIR-10 threshold is the relevant one. It's not the attention-grabbing one. So I expect labs to easily break it without any government attention."
- "Overall, I expect the government to miss the ball on this in the early stages, being largely focused on other matters."

Acknowledgements

We are grateful to the following people for providing valuable feedback and insights: Ashwin Acharya, Will Aldred, Ryan Beck, Deric Cheng, Tom Davidson, Shaun Ee, Rose Hadshar, Eli Lifland, Will MacAskill, Jenny Marron, Angus Mercer, Malcolm Murray, Javier Prieto, Nuño Sempere, Zach Stein-Perlman, Benjamin Tereick, Lizka Vaintrob, Peter Wildeford, and Zoe Williams. Jamie Elsey and Alex Rand provided valuable statistical advice and analyses.

We would like to extend special thanks to Dulani Woods at the RAND Corporation for running the workshop, as well as to our workshop participants Vidur Kapur, Brodi Kotila, David Manheim, Dewey Murdick, Philipp Schoenegger, and six other anonymous participants.

Bibliography

Acharya, A., and O. Delaney. Forthcoming. "Managing Risks from Internal AI Models."

AI Impacts. 2022. "Human-Level AI." https://perma.cc/ULZ2-V6LM.

Anthropic. 2025. "Responsible Scaling Policy 2.1" https://www.anthropic.com/rsp-updates.

Aschenbrenner, L. 2024. "Situational Awareness: The Decade Ahead." https://situational-awareness.ai/.

Cheng, D., and C. Katzke. 2024. "Soft Nationalization: How the US Government Will Control AI Labs." *SuperIntelligence - Robotics - Safety & Alignment* 1(1). https://doi.org/10.70777/si.v1i1.10931.

Davidson, T. 2021. "Semi-Informative Priors Over AI Timelines | Open Philanthropy." https://perma.cc/ND94-JNGG.

Dimitri, N., T. Gries, and W. Naudé. 2024. "Investing in Artificial Intelligence: Breakthroughs and Backlashes." In *Artificial Intelligence: Economic Perspectives and Models*, 173–198. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781009483094.007.

Epoch AI. 2025. "Data on Notable AI Models." https://perma.cc/GDC6-G8XQ.

Erdil, E., and T. Besiroglu. 2023. "Explosive Growth from AI Automation: A Review of the Arguments." arXiv:2309.11690. https://doi.org/10.48550/arXiv.2309.11690.

Erdil, E., T. Besiroglu, and A. Ho. 2024. "Estimating Idea Production: A Methodological Survey." arXiv:2405.10494. https://doi.org/10.48550/arXiv.2405.10494.

Franklin, J. 2010. "Feature Selection Methods for Solving the Reference Class Problem: Comment on Edward K. Cheng, 'A Practical Solution to the Reference Class Problem.'" *Columbia Law Review* 110: 12–23. https://www.columbialawreview.org/wp-content/uploads/2016/07/Franklin.pdf.

Friedman, J. A., J. D. Baker, B. A. Mellers, P. E. Tetlock, and R. Zeckhauser. 2018. "The Value of Precision in Probability Assessment: Evidence from a Large-Scale Geopolitical Forecasting Tournament." *International Studies Quarterly* 62(2): 410–422. https://doi.org/10.1093/isq/sqx078.

Hanea, A. M., M. Burgman, and V. Hemming. 2018. "IDEA for Uncertainty Quantification." In *Elicitation: The Science and Art of Structuring Judgement*, edited by L. C. Dias, A. Morton, and J. Quigley, 95–117. Springer International Publishing. https://doi.org/10.1007/978-3-319-65052-4_5.

Hemming, V., M. A. Burgman, A. M. Hanea, M. F. McBride, and B. C. Wintle. 2017. "A Practical Guide to Structured Expert Elicitation Using the IDEA Protocol." *Methods in Ecology and Evolution* 9(1): 169–180. https://doi.org/10.1111/2041-210X.12857.

Hemming, V., T. V. Walshe, A. M. Hanea, F. Fidler, and M. Burgman. 2018. "Eliciting Improved Quantitative Judgements Using the IDEA Protocol: A Case Study in Natural Resource Management." *PLoS ONE* 13(6). https://doi.org/10.1371/journal.pone.0198468.

Hendrycks, D., M. Mazeika, and T. Woodside. 2023. "An Overview of Catastrophic AI Risks." arXiv:2306.12001. https://doi.org/10.48550/arXiv.2306.12001.

Hendrycks, D., E. Schmidt, and A. Wang. 2025. "Superintelligence Strategy: Expert Version." arXiv:2503.05628. https://doi.org/10.48550/arXiv.2503.05628.

Karger, E., J. Rosenberg, Z. Jacobs, M. Hickman, R. Hadshar, K. Gamin, T. Smith, B. Williams, T. McCaslin, S. Thomas, and P. E. Tetlock. 2023. "Forecasting Existential Risks: Evidence from a Long-Run Forecasting Tournament." FRI Working Paper 1. https://forecastingresearch.org/xpt.

Karnofsky, H. 2021. "Forecasting Transformative AI, Part 1: What Kind of AI?" https://perma.cc/37WK-MAZZ.

Katzke, C., and G. Futerman. 2024. "The Manhattan Trap: Why a Race to Artificial Superintelligence is Self-Defeating." arXiv:2501.14749. https://doi.org/10.48550/arXiv.2501.14749.

Kokotajlo, D., S. Alexander, T. Larsen, E. Lifland, and R. Dean. 2025. "AI 2027." AI Futures Project. https://ai-2027.com/.

Lipsey, R. G., K. I. Carlaw, and C. T. Bekar. 2005. *Economic Transformations: General Purpose Technologies and Long-Term Economic Growth.* Oxford: Oxford University Press. https://perma.cc/Y88M-STXJ.

Metaculus. 2025a. "Group to Develop First TAI." https://perma.cc/DYS7-TUNW.

Metaculus. 2025b. "Date of Artificial General Intelligence." https://perma.cc/PP4K-GBFH.

Metaculus. 2025c. "Transformative AI Date." https://perma.cc/2W93-69JE.

Naudé, W., T. Gries, and N. Dimitri. 2024. "Investing in Artificial Intelligence: Breakthroughs and Backlashes." In *Artificial Intelligence: Economic Perspectives and Models*, 173-198. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781009483094.007.

Ono, R., and D. J. Wedemeyer. 1994. "Assessing the Validity of the Delphi Technique." *Futures* 26(3): 289–304. https://doi.org/10.1016/0016-3287(94)90016-7.

Open Philanthropy. 2016. "Some Background on Our Views Regarding Advanced Artificial Intelligence." https://perma.cc/NF4L-KLXG.

OpenAI. 2023. "Preparedness Framework (Beta)." https://perma.cc/VS4N-TRBC.

Organisation for Economic Co-operation and Development [OECD]. 2023. "Artificial Intelligence in Science." https://www.oecd.org/en/publications/artificial-intelligence-in-science_a8d820bd-en.html.

Owen, D. 2024. "Interviewing AI researchers on Automation of AI R&D." Epoch AI. https://perma.cc/U989-6QRL.

Parente, R., and J. Anderson-Parente. 2011. "A Case Study of Long-Term Delphi Accuracy." *Technological Forecasting and Social Change* 78(9): 1705–1711. https://doi.org/10.1016/j.techfore.2011.07.005.

Pichai, S. 2024. "Alphabet Q3 Earnings Call: CEO Sundar Pichai's Remarks." Google. https://perma.cc/39TP-JULN.

Potter, B. 2024. "Infrastructure Costs Archive." https://perma.cc/JZW5-K9MU.

Sevilla, J., T. Besiroglu, B. Cottier, J. You, E. Roldán, P. Villalobos, and E. Erdil. 2024. "Can AI Scaling Continue Through 2030?" Epoch AI. https://epoch.ai/blog/can-ai-scaling-continue-through-2030.

Tetlock, P. E., and D. Gardner. 2016. *Superforecasting: The Art and Science of Prediction.* New York: Penguin Random House. https://perma.cc/FTG8-M5RS.

Tetlock, P. E., C. Karvetski, V. A. Satopää, and K. Chen. 2023. "Long-Range Subjective-Probability Forecasts of Slow-Motion Variables in World Politics: Exploring Limits on Expert Judgment." *Futures & Foresight Science* 6(1): e157. https://doi.org/10.1002/ffo2.157.

U.S.-China Economic and Security Review Commission. 2024. "Report to Congress." One Hundred Eighteenth Congress, Second Session. https://www.uscc.gov/sites/default/files/2024-11/2024_Annual_Report_to_Congress.pdf.

Wijk, H., T. Lin, J. Becker, S. Jawhar, N. Parikh, T. Broadley, L. Chan, et al. 2024. "RE-Bench: Evaluating Frontier AI R&D Capabilities of Language Model Agents Against Human Experts." arXiv:2411.15114. https://doi.org/10.48550/arXiv.2411.15114.

Williams-Jones, B., C. Olivier, and E. Smith. 2014. "Governing 'Dual-Use' Research in Canada: A Policy Review." *Science and Public Policy* 41(1): 76–93. https://doi.org/10.1093/scipol/sct038.

Zelikow, P., M.-F. Cuéllar, E. Schmidt, and J. Matheny. 2024. "Defense Against The AI Dark Arts: Threat Assessment And Coalition Defense." *Hoover Institution.* https://perma.cc/5AK9-F2GB.

Appendix: Alternative statistical analyses

Our main dataset consists of each participant's median, 5th, and 95th percentile estimates for the overall question and the five scenarios. The correct method for aggregating this data into overall summary statistics is nonobvious. In the main text, we simply took the average of all participants' medians as our central estimate, and the average of the 5th and 95th percentiles as the bounds. In this appendix, we show results for three other possible aggregation approaches.

Beta regression

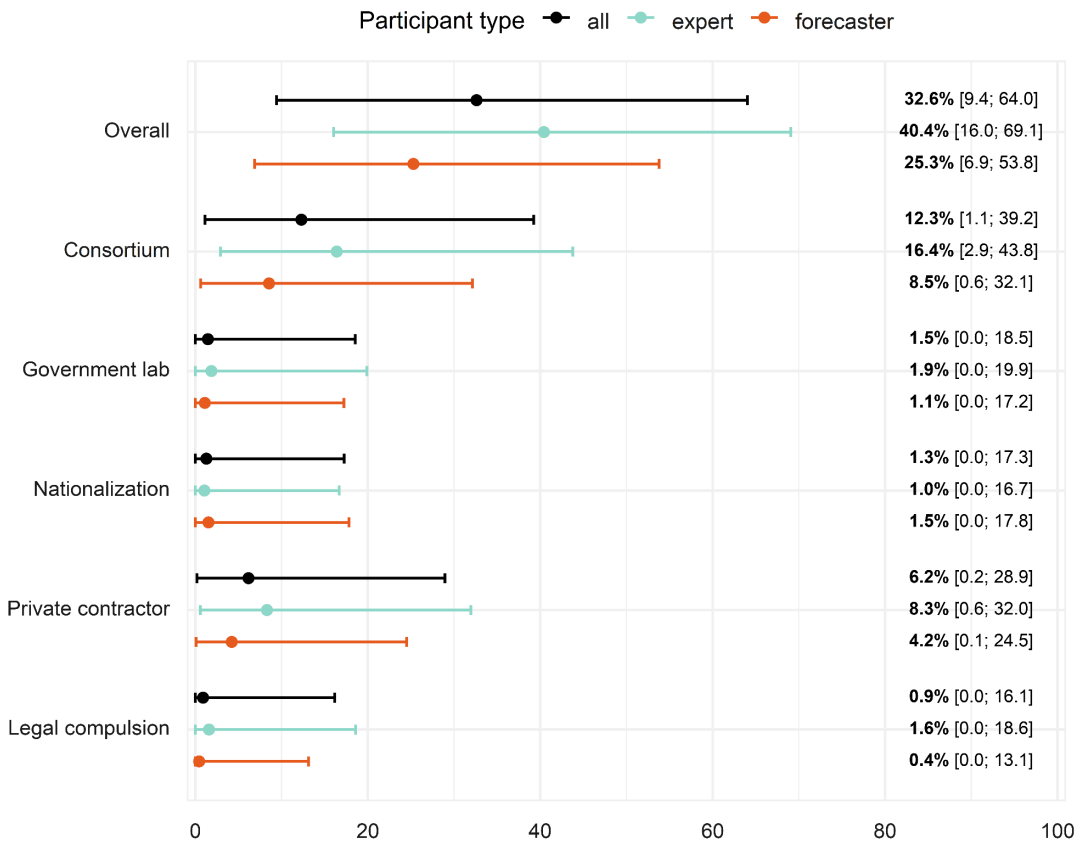

Figure 5: Beta regression aggregated forecasts by question and respondent type.

Here, we fit a beta regression, which is appropriate for modeling probabilities between 0 and 1, to the respondents' medians, without incorporating a statistical prior (Figure 5). A key feature of this method is that it does not incorporate participants' subjective confidence intervals, but instead uses the variation between participants to compute a confidence interval.

Some analysts, such as Hanea, Burgman, and Hemming (2018), note that the idea of an individual possessing a 90% confidence interval over the probability of a one-off event is difficult to operationalize. As a result, they suggest asking for such confidence intervals to encourage counterfactual thinking, but then discarding them after the workshop. We attempted to mitigate this difficulty by carefully operationalizing the 90% confidence interval for participants:

> *Imagine that there was an idealized superforecaster who was perfectly rational, with all available information one could gather on a particular topic and the best possible ways of interpreting that information. What is the interval within which you think their answer would fall? (90% confidence)*

Therefore, for the following two methods, we *did* incorporate participants' subjective 90% confidence intervals. Nonetheless, the above authors' objection is one possible reason to prefer the beta regression approach.

Metalog distributions

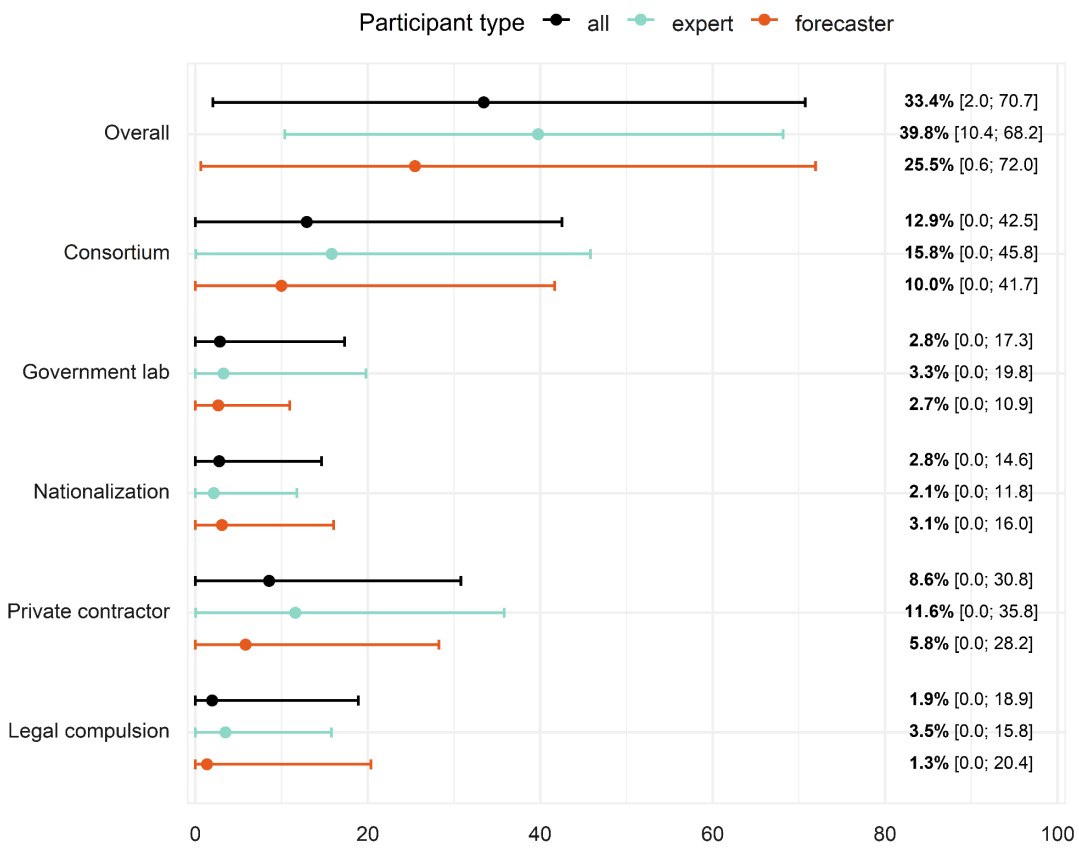

Figure 6: Metalogarithmic distribution aggregated forecasts by question and respondent type.

Here, we fit a flexible metalogarithmic distribution[40] to each respondent's median and 90% CI to capture their individual subjective uncertainty (Figure 6). Then we computed the unweighted average of these distributions to get a descriptive distribution for the group's uncertainty as a whole, and summarized the resulting distribution with its median, 5th and 95th percentiles. This approach preserves each participant's individual uncertainty assessment while generating an aggregate view.

Bayesian hierarchical model

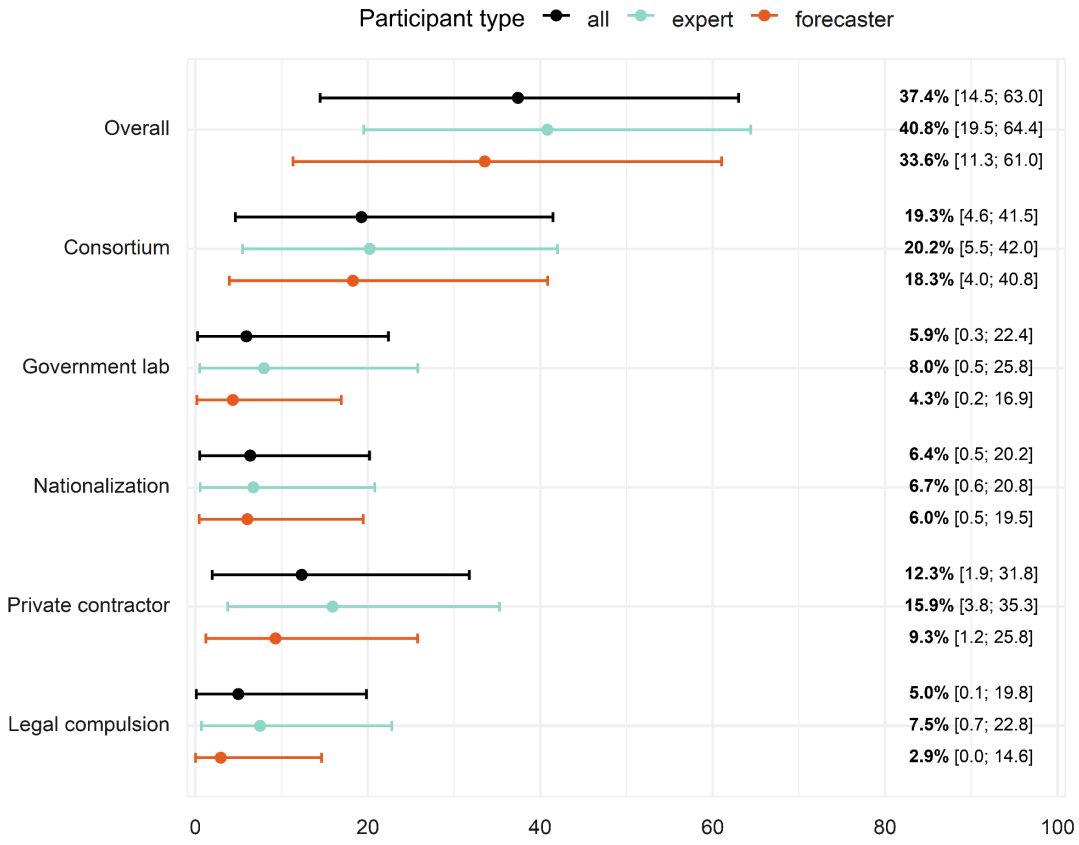

Figure 7: Bayesian hierarchical model forecasts by question and respondent type.

The distinctive feature of this approach is its use of a formal prior, which the workshop data then updates. We used a flat prior for each question, which makes the results skew towards 50% given the limited dataset (Figure 7). This method is the most complex to implement and is described in more detail here.

[40] Metalog distributions are particularly useful here because they can represent arbitrary quantile inputs without requiring binding assumptions about the form of the underlying probability distribution.