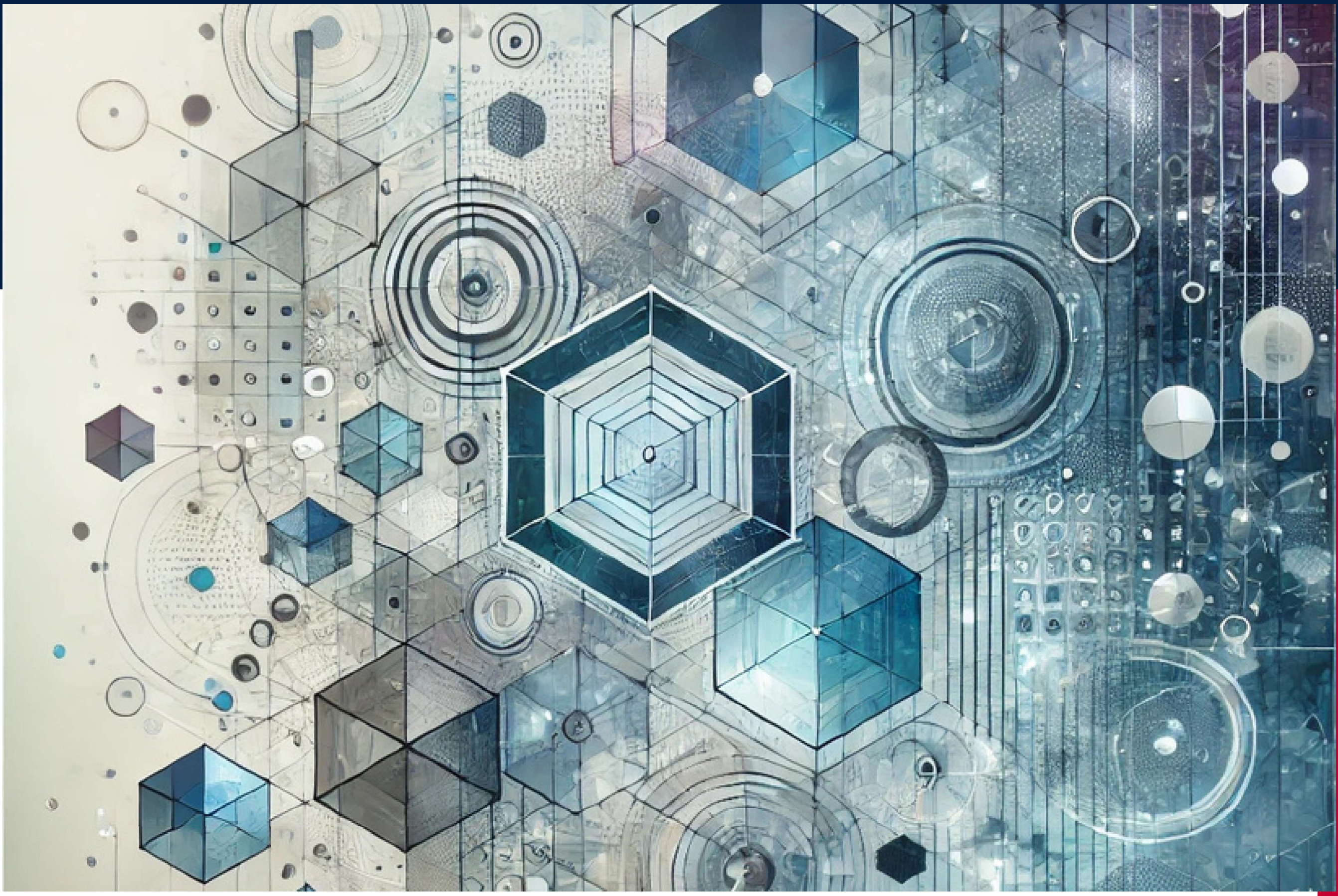
RESEARCH MEMO
JANUARY 2025

AI GOVERNANCE INITIATIVE
OXFORD MARTIN SCHOOL
UNIVERSITY OF OXFORD

Who Should Develop Which AI Evaluations?



Authors: Lara Thurnherr, Robert Trager, Amin Oueslati, Christoph Winter, Cliodhna Ní Ghuidhir, Joe O'Brien, Jun Shern Chan, Lorenzo Pacchiardi, Anka Reuel, Merlin Stein, Oliver Guest, Oliver Sourbut, Renan Araujo, Seth Donoughe and Yi Zeng

Who Should Develop Which AI Evaluations?

Lara Thurnherr*, Robert Trager*, Amin Oueslati, Christoph Winter, Cliodhna Ní Ghuidhir, Joe O'Brien, Jun Shern Chan, Lorenzo Pacchiardi, Anka Reuel, Merlin Stein, Oliver Guest, Oliver Sourbut, Renan Araujo, Seth Donoughe, Yi Zeng

Abstract

We explore frameworks and criteria for determining which actors (e.g., government agencies, AI companies, third-party organisations) are best suited to develop AI model evaluations. Key challenges include conflicts of interest when AI companies assess their own models, the information and skill requirements for AI evaluations and the (sometimes) blurred boundary between developing and conducting evaluations. We propose a taxonomy of four development approaches: government-led development, government-contractor collaborations, third-party development via grants, and direct AI company development.

We present nine criteria for selecting evaluation developers, which we apply in a two-step sorting process to identify capable and suitable developers. Additionally, we recommend measures for a market-based ecosystem to support diverse, high-quality evaluation development, including public tools, accreditation, clear guidelines, and brokering relationships between third-party evaluators and AI companies. Our approach emphasises the need for a sustainable ecosystem to balance the importance of public accountability and efficient private-sector participation.

Table of Contents

Executive Summary

This memo discusses which actors should develop different AI model evaluations and outlines how governments could spur third-party evaluation development through market creation measures.[1]

The challenges of determining who should develop which evaluations: AI companies face a conflict of interest when developing evaluations for their own products. Third-party developers face a challenge of impartiality due to partial financial dependence on AI companies. Yet, much of the expertise and data required to develop evaluations is found in the private sector. Furthermore, the question of who develops evaluation is linked to who conducts and interprets evaluations, because evaluators may require extensive information on the development process to fulfil their tasks.

We begin by identifying **four development approaches**, each with strengths and weaknesses:

- **AISIs developing evaluations**. AISIs and related public bodies take charge of the evaluation development process from start to finish. This approach could be useful in cases where evaluation development requires a high level of information security, access to classified information and a high level of independence from AI companies, but it might be costly and does not encourage the emerging ecosystem of evaluation developers.
- **Contracting experts for joint development**. AISIs develop evaluations jointly with contracted experts from the private sector. This would entail coordination costs, but enable the incorporation of specialised expertise while still allowing potential access to classified information
- **Funding third parties for independent development**. Through funding from public bodies, foundations or AI Companies, third-parties could develop evaluations on their own. This could spur the ecosystem and allow actors with more flexibility to pursue experimental approaches, but it could also make quality control more difficult if public bodies are not sufficiently informed about the development process of a specific evaluation.
- **AI company development.** AI companies could develop evaluations themselves. This could be a valuable and cost-effective option for evaluation development requiring high levels of expertise and model access. However, firm conflicts of interest in producing favourable evaluations must be mitigated.

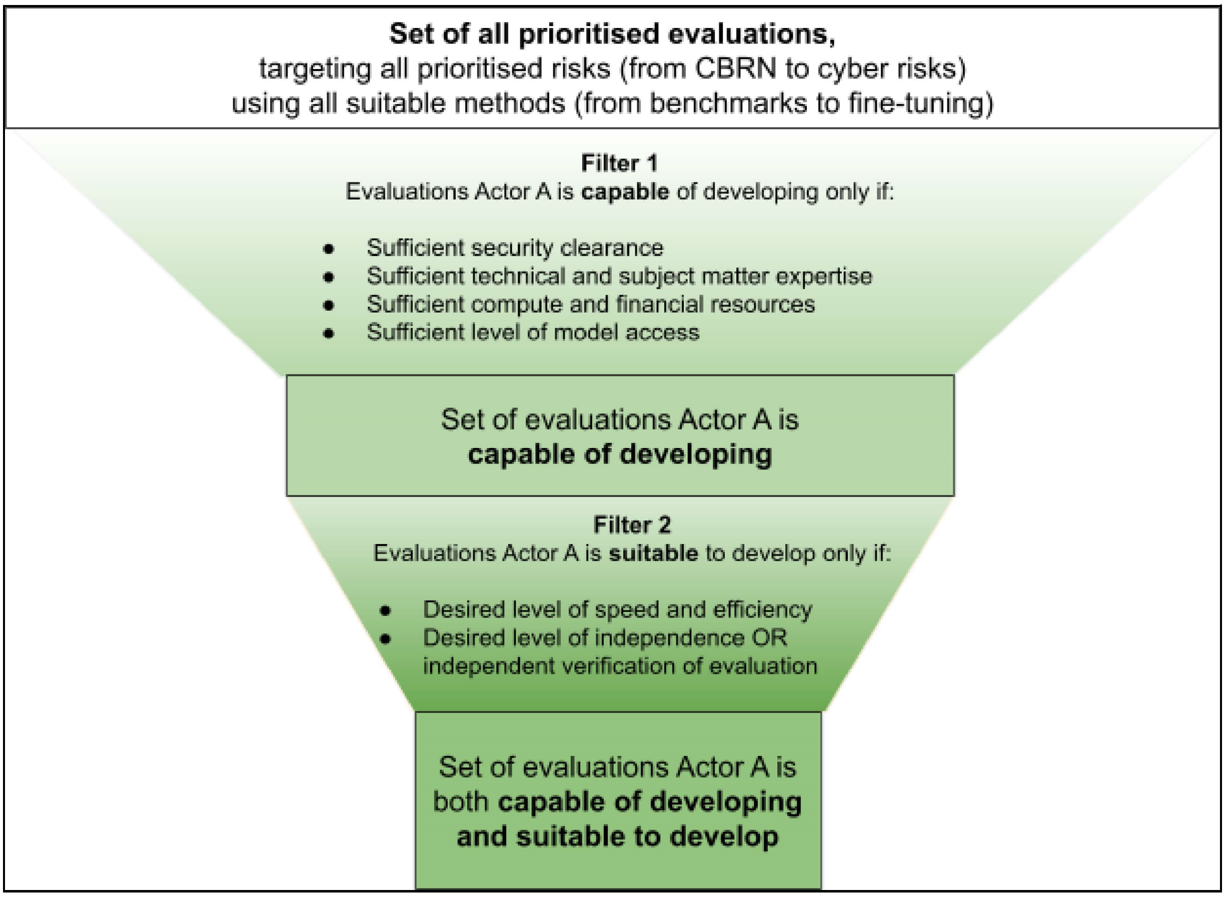
These lead to **nine criteria to decide which evaluations should be developed by whom**. We adapt a framework from Stein et al. on public body involvement in evaluation regimes and propose a process for deciding who should develop which evaluations. [2]

[1] We define AI system evaluations development as the process of creating an assessment of one or more AI models capabilities and safeguards.

[2] Stein et al. (2024)

The first step of this process involves identifying the actors who are technically capable of developing an evaluation *for a specific risk*. The second step ranks the suitability of those actors based on the urgency of an evaluation and the required level of independence. In Figure 1, we present this sorting process. This filter process can be applied to all potential actors in an evaluation ecosystem to allocate responsibilities.

Fig. 1: Decision Process for Determining Which Evaluations an Actor Should Develop



Set of all prioritised evaluations,
targeting all prioritised risks (from CBRN to cyber risks)
using all suitable methods (from benchmarks to fine-tuning)

Filter 1
Evaluations Actor A is **capable** of developing only if:

- Sufficient security clearance
- Sufficient technical and subject matter expertise
- Sufficient compute and financial resources
- Sufficient level of model access

Set of evaluations Actor A is
capable of developing

Filter 2
Evaluations Actor A is **suitable** to develop only if:

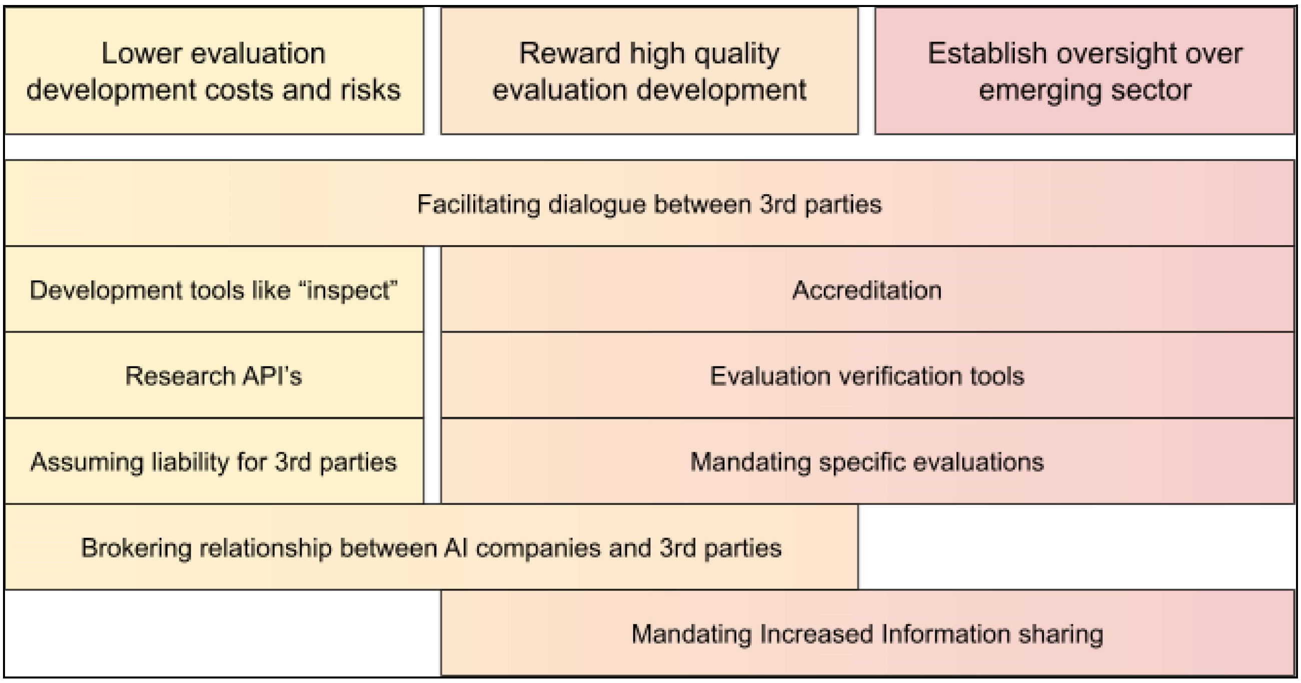
- Desired level of speed and efficiency
- Desired level of independence OR independent verification of evaluation

Set of evaluations Actor A is
both **capable of developing**
and suitable to develop

It is plausible that after this sorting process, many important evaluations will not be developed because of a limited number of both suitable and capable actors. For this reason we note, finally, that an effective ecosystem for evaluation development must be *created*. We identify **twelve government measures to spur a market-based ecosystem** (Fig. 2). We propose twelve measures with which public bodies could increase the number of capable and suitable actors in the evaluation ecosystem, by making it easier for third parties to contribute (yellow in fig. 2), incentivising quality development (orange in fig. 2) and improving government oversight (red in fig. 2). These measures, which are visualised in Figure 2, include public tools and resources for evaluation development and tools for the verification of the functionality of evaluations which could incentivise companies and third parties to develop high-quality evaluations.

Standards, best practices, clear expectations and a sufficient amount of legal certainty for third parties could incentivise investment in the development and execution of desired evaluations, verified by the research community. Public bodies brokering the relationship between third parties and AI companies could increase third-party evaluation developers' independence. Over time, the measures listed in this section can be adapted to align with the goal of placing the financial burden of evaluation development primarily on model developers.

Fig. 2: Measures for the Creation of a Market-Based Evaluation Ecosystem



Lower evaluation development costs and risks

Reward high quality evaluation development

Establish oversight over emerging sector

Facilitating dialogue between 3rd parties

Development tools like "inspect"

Accreditation

Research API's

Evaluation verification tools

Assuming liability for 3rd parties

Mandating specific evaluations

Brokering relationship between AI companies and 3rd parties

Mandating Increased Information sharing

Introduction

A range of actors with varying characteristics could theoretically develop some evaluations. These actors include AISIs and related public bodies, academics, independent researchers, third-party evaluation organisations (e.g. Apollo or METR) and AI companies. The characteristics of potential evaluation also vary widely: Which risk area is a given evaluation assessing? Which elicitation methods are attempted? What modalities and affordances are anticipated by the evaluation? Risks include chemical, biological, radiological and nuclear (CBRN), risks to disinformation and others. Methods range from simple model calls to fine-tuning and extensive use of test-time compute, from text-response benchmarks to agentic evaluations.

This state of affairs raises two questions: 1) Which actors are capable of developing which specific evaluations? and 2) Of those capable actors, who should develop them ideally?

However, simply knowing which actors are best suited for the development of which evaluations is arguably insufficient to create a functioning evaluation regime. Reducing the risks and costs of entering the sector, incentivising quality development and improving public bodies' ability to oversee this ecosystem are other intermediate goals to strive for. Public bodies can contribute to reaching these goals, to create a flourishing ecosystem and, ideally, a self-sustaining market for evaluations in the future. This memo explores twelve measures to this end, including accreditation of third parties, mandating AI companies to develop certain evaluations themselves and brokering the relationship between third parties and AI companies.

1 Evaluation Development Context

Conflicts of Interest in Private Evaluation Development

Considering that AI companies stand to gain the most from advanced AI models, it can be argued that they should carry the bulk of the financial burden associated with developing, conducting or interpreting evaluations. Extending this argument from the responsibility to carry the financial burden to the responsibility of performing these evaluation steps, however, entails risks. Most notably, AI companies face a conflict of interest when developing, conducting or interpreting evaluations, especially if these same evaluations are used to inform governmental regulation.[3] The magnitude of the risk from this conflict of interest depends on a variety of factors, including the magnitude of a risk domain targeted by the evaluation and whether an external verification of these evaluations is feasible. Third-party organisations can also face conflicts of interest if they depend on continuous demand from the AI companies they evaluate to sustain themselves. This dependence could incentivise third-party evaluators to produce favourable assessments to maintain a good relationship with their customers, thereby affecting impartiality.

[3] These conflicts of interest could result in the performance of more favorable evaluations or the interpretation of these results. Capability evaluations of LLMs have also been shown to be vulnerable to "sandbagging", where models are prompted or trained to hide specific capabilities, or to target specific scores on capability evaluations (Weij et al., 2024).

Evaluation Regimes in Other Industries

Focusing on conducting and interpreting evaluations in other industries, Stein et al. identify a trade-off present in nine case studies of auditing regimes: Publicly accountable government actors or publicly appointed private actors tend to be more involved in conducting and interpreting ("judging") evaluations and audits on high-risk areas involving sensitive information.[4] They also note, however, that too much public body involvement in these evaluation steps might reduce the efficiency of the evaluation regime. Many AI evaluations address sensitive high-risk areas, like risks from engineered pathogens or cyber attacks on critical infrastructure. Stein et al. find that this circumstance could justify public body involvement in conducting and judging AI evaluations. Conducting some evaluations can require highly specialised expertise and the flexibility to adapt to an expanding list of relevant evaluation methods and risks.[5] As a result, The private sector will likely play a key role in conducting evaluations that require higher levels of flexibility or a higher amount of replication by multiple independent actors to further a "science of evaluations". This trade-off between efficiency and public accountability should also be considered when allocating responsibilities for AI evaluation development.

Evaluation Development Includes a Variety of Activities

It should be noted that the process of developing evaluations includes a variety of decisions and activities. To know which specific evaluations should be developed, risk models need to be generated and operationalised into specific questions a model evaluation can plausibly answer. Further, evaluation development can include compiling a specific set of questions to indicate a model's abilities or safeguards or the construction of a specific elicitation method and potential scaffolding.[6] The development process can also include the formulation of assessment processes and criteria.

Evaluation Development is Challenging

Developing some evaluations can require high levels of technical and domain-specific expertise, access to advanced models or sensitive information.[7] For example, developing an evaluation to assess the likelihood that a specific LLM would help a biology novice engineer a harmful pathogen can involve the use of prompting strategies or compiling questions from human subject matter experts, it could involve classified data on existing engineered pathogens and involve the attempt to circumvent the safeguards built into advanced LLMs.

[4] Stein et al.(2024)

[5] Evaluation development is a new field, with an uncertain value. Government agencies have an incentive and history to prioiritise a risk-averse use of taxpayer funds (OECD, 2017). Bureaucratic inertia can further slow down building specialised capacity in novel fields (Ritchie, 2024). In contrast, private companies, especially startups, thrive on developing unique selling propositions (USPs) and exploiting niche markets, taking bets that may not yield immediate returns (Pallardy, 2022). As a result, private firms often attract talent with specialized skills tailored to specific technological domains (OECD, 2017).

[6] Many capabilities of frontier AI systems are 'emergent' in the sense that they were not deliberately crafted and it may be unknown whether a given model is able to contribute a given task or activity. The process of discovering these capabilities is referred to as 'capability elicitation'. Many capabilities may require innovations in 'scaffolding' (structuring of inputs and intermediate computation, integration with auxiliary tools, and ways of deploying more computational resources to achieve stronger results). For LLMs in particular, design of prompting strategies can reveal capabilities or behaviours developers did not expect, or even ones they explicitly tried to avoid. See: Geiping et al. (2024) and Wei et al. (2022).

[7] See: Anthropic, 2024; Casper et al, 2024; UK AISI, 2024

Of course, these requirements vary widely and depend on the specific goal of evaluations: Asking a model basic biology questions doesn't require as much technical expertise or model access as eliciting dangerous capabilities through more sophisticated methods.[8]

The Boundary between Developing, Conducting, and Interpreting Evaluations can be Blurry
In some cases, evaluators might want to monitor or require detailed reporting on the development of the evaluations they are likely to conduct. An accurate interpretation of evaluation results can require extensive knowledge of the evaluation development process.[9] Another complicating factor is model access: Ideally, evaluations are developed for a range of current and future AI models ("model-agnostic"). However, the development of some evaluation methods (like manual capability elicitation) can require an iterative process specific to each model. As a consequence, developing these methods can require a similar level of access to a model as conducting an evaluation.

2 Taxonomy of Actor-Based Evaluation Development Approaches:

To gather insights into which evaluations should be developed by whom, we describe four potential development approaches. Each approach comes with its own (dis-)advantages.

AISIs developing evaluations. In this approach, AISIs (or related public bodies) are in charge of the evaluation development process from start to finish. They might consult with other government experts (e.g. CBRN experts) to improve this process.

Contracting experts for joint development. In this approach, AISIs contract specific organisations or individuals to collaborate with them on developing evaluations. Contractors might work on developing subject-specific benchmarks or analyse which specific capabilities could present the risks targeted by evaluations. AISIs can then incorporate this expertise into their development process.

Funding third parties for independent development. AISIs or AI companies fund third parties to develop (or research development) independently. AISIs might publish a call for funding applications and give grants to academics, private organisations or research institutions they find particularly promising.

AI company development. This approach could take place through non-binding commitments or binding mandates, either on specific evaluations or on dedicating a portion of their budget (or compute) towards furthering the science of evaluations.

[8] See footnote 6

[9] Specific knowledge might include details of the technical approach, expertise of developers, level of access used in evaluations development.

Table 1: Simplified Advantages and Disadvantages of Different Development Models

Development Approach	Advantages	Disadvantages
AISIs develop evaluations	● Information security ● Independence ● Access to classified information	● Resource intensive ● Doesn't spur evaluation ecosystem
AISIs develop evaluations jointly with contracted parties	● Specialised expertise ● Easier oversight ● Access to classified information	● Resource intensive ● Coordination costs
AISIs / AI companies fund Third-Parties (grants)	● Spurs development ecosystem ● Replication ● Flexibility for experimental approaches	● Less quality control and oversight ● Less public accountability
AI Company Development	● High levels of expertise ● Cost-effective	● Conflict of interest ● Less quality control and oversight

Ultimately, which of these approaches is most suitable for specific evaluation development depends on the relevant characteristics of specific evaluations. The following section outlines these relevant characteristics.

3 Criteria to Determine Who Should Develop Which Evaluations

Some criteria depend on the risk domain targeted by the evaluation, and some on the evaluation method. We distinguish these in the list below. Note, however, that the risk domain strongly influences the choice of methods. Higher-risk domains require more reliable evaluation methods. The use of these methods can require more extensive access to a model.[10] Furthermore, the assessment of some risks (e.g. deception[11] or anomalous failures[12]) is challenging without methods targeting the inner workings of specific models.[13]

[10] Casper et al. (2024)

[11] Park et al. (2023)

[12] Ziegler et al. (2022)

[13] Casper et al. (2024)

Risk-Related Criteria

1) Required Risk-Related Skills and Expertise

Skill specificity refers to the rarity and level of specialised expertise required for developing evaluations on these risks. Compiling questions for a simple benchmark can require extensive subject matter expertise on virology or misinformation. Evaluations should be developed by organisations with adequate levels of specialised expertise for a particular evaluation. It should be noted that, with sufficient resources, public and private bodies can build expertise in a particular field over time, especially when a risk exhibits a high level of public salience.[14]

2) Information Sensitivity and Security Clearances

Developing evaluations for some risk domains (e.g. CBRN risks) can require access to classified data, necessitating a higher level of security clearance. While some private industry personnel can get security clearances, large-scale development using classified material is difficult in the private sector and impossible for the most sensitive material in select areas of national security.[15] However, the extent of security clearances necessary to develop evaluation should not be overestimated. A large proportion of the development of evaluations targeting biological risk does not require a security clearance: As a simplified example, these evaluations can target non-sensitive capabilities like how helpful a specific model is in providing methodological guidance and troubleshooting.[16]

3) Evaluation Urgency

If a reasonable worst-case scenario for a specific risk area includes near-term harms, and if there aren't adequate or established evaluations available yet, developing new evaluations quickly rises in relative importance to other factors: Evaluations should be developed by the quickest actor (or combination of actors) in those cases. Note that the urgency of an evaluation might also be influenced by how long development is expected to take.[17]

4) Risk Prevention Incentives

The importance of preventing a particular risk event increases the significance of ensuring high-quality evaluation development. For this reason, the incentives of evaluation developers need to be as aligned as possible with reducing this particular risk. When both AI companies and third-party organisations face other incentives[18] publicly accountable government actors should play a bigger role in evaluation development.[19]

[14] Stein et al. (2024)

[15] This might heavily depend on the specific jurisdiction an evaluation is developed in.

[16] Examples of biological risk evaluations that could be developed without a security clearance are described in UKAISI (2024) and Jin et al (2019).

[17] E.g. If an evaluation is expected to take years to develop, starting as soon as possible might be important.

[18] As outlined in Section 1 under "Conflicts of interest of private evaluation development."

[19] The prioritisation of the prevention of a risk event, of course, various. One common proxy is the likelihood and scale of a risk externality. The EU AI Acts definition of "serious incidents" could provide a preliminary orientation: " 'serious incident' means an incident or malfunctioning of an AI system that directly or indirectly leads to any of the following: the death of a person, or serious harm to a person's health; a serious and irreversible disruption of the management or operation of critical infrastructure."

Method-Related Criteria

1) Level of Model Access[20] required

For some evaluation methods, evaluation developers need more extensive model access than the wider public. The level of access required for developing an evaluation is not necessarily the same as the level of access required to conduct an evaluation. For example, some evaluations can be developed to be "model neutral", so developers don't need access to the specific model an evaluation ends up being applied to. Other evaluations do.
The following factors can influence the level of model access evaluation developers need:

> **1.1) Use of elicitation techniques**: Where elicitation includes model-specific approaches, or includes iteration on a particular system-task pair, developing "model neutral evaluations" becomes harder.
> **1.2) Open-source or closed-source deployment:** The development of misuse-focused evaluations for models which will be deployed open-source[21] requires higher levels of model access because potential "misusers" will also have higher levels of access. If users can fine-tune a model for nefarious purposes, evaluation developers need to have equivalent levels of access to develop corresponding insightful tests.[22] It should also be considered that model architecture or trained weights might not be deliberately published at deployment, but could be leaked at some point in time after deployment. Because of this risk, in some cases, it may be advisable to evaluate some models as if they'd be deployed open-source, even if they aren't.
> **1.3) Pre- or Post-Deployment evaluations:** Granting pre-deployment access to models might be more sensitive for AI companies than post-deployment access because AI companies have a commercial interest in keeping their IP safe from competitors and the general before deployment.

It should be noted that various measures can increase the number of actors who have sufficient access to a model to develop evaluations for it, without disproportionately compromising the security of valuable IP of AI companies. Research APIs and on-site evaluation development (where external evaluation developers work on data on AI company servers) are examples. Public bodies could also require AI companies to share model access with specific third-party evaluation developers or governmental institutions (e.g. AISIs).

2) Evaluation Development Costs

Evaluation development can be expensive. The actors responsible for developing a specific evaluation require sufficient financial resources to do so. Evaluation development costs can correlate with the data volume required and the price of the expertise to gather data, analyse it and develop the technical foundations of the evaluation. Compute costs of developing (and running) a specific evaluation method should also be considered here.

[20] For an elaboration of which evaluation methods require which level of access to a specific model, see Casper et al. 2024.
[21] We use the open source definition found in Seger et al. (2023): Open source deployment includes making, at least, model architecture and trained weights are publicly available.
[22] This example only applies to models where API's don't already allow for fine-tuning.

3) Required Method-Related Skills and Expertise

The same considerations as above apply here. Developing specific evaluation and elicitation methods can require specific expertise and experience. In-depth knowledge of prompting AI systems, the inner workings of a model or the practical ramifications of development could be key to ensuring useful evaluations. To ensure high-quality evaluation development, actors require this expertise.

4) Verifiability and Documentation

If public bodies are not leading the development process of a particular evaluation method, they might want to be able to verify its functionality before adopting it. When the functionality of evaluation methods is easier to verify, it is more feasible for government institutions to assess the evaluations developed by independent third parties.. How easy it is to verify an evaluation is influenced by a variety of factors. Examples include the expected level of documentation that will be made available on the development process and the extent to which a particular evaluation method follows existing standards or best practices or has been replicated by the research community.

Limitations

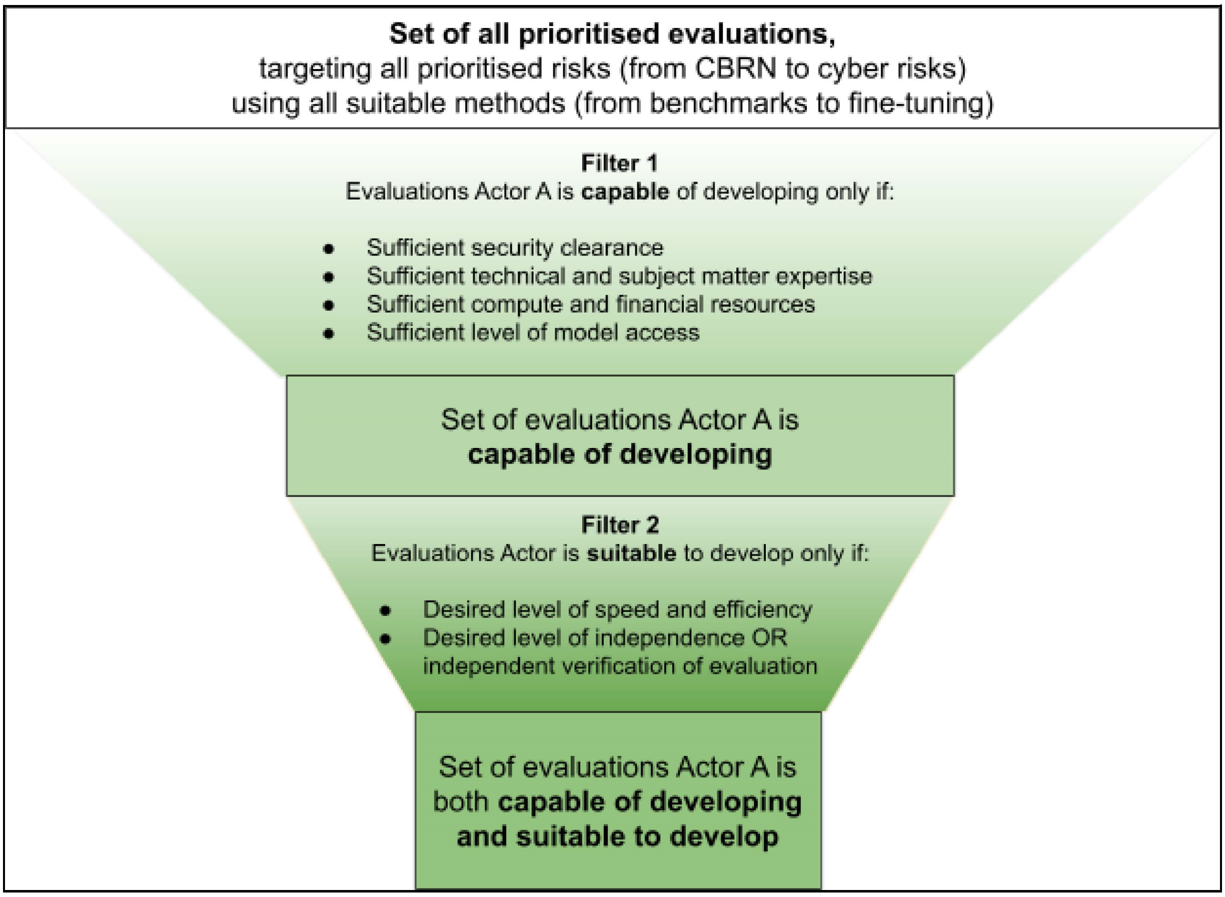
The criteria we outlined in this section have a number of limitations.

Firstly, many of the criteria depend on questions which remain insufficiently studied. For example, more research on the scale of specific AI risks is needed to reasonably prioritise the prevention of specific AI risks. Furthermore, the answers to these questions will likely change over time. The scale of risks might change with diffusion and advancements of AI capabilities, currently, unknown risks will almost certainly emerge, as will new evaluation methods with yet unknown costs and access requirements. This should encourage public bodies to monitor these criteria continuously to facilitate adaptation.

Secondly, jurisdictional differences should be taken into account when weighing the criteria. All governments face similar questions on "who should develop evaluations" and many of the criteria outlined in section 3 are internationally relevant, but important differences between jurisdictions remain: These include the domestic evaluation development ecosystem, the level of existing evaluation development expertise and financial resources in government as well as government document security classifications.

Ongoing developments of international coordination around this subject should also be taken into account: International coordination could allow for some level of research pooling and specialisation on evaluation development. For example, AISIs in jurisdictions with frontier companies might be best suited to develop evaluations requiring a higher level of model access (or broker the relationship between governments and third parties). A scenario where these AISIs develop and conduct highly critical evaluations for other AISIs might be desirable.

Fig. 1: Decision Process for Determining Which Evaluations an Actor Should Develop


Set of all prioritised evaluations,
targeting all prioritised risks (from CBRN to cyber risks)
using all suitable methods (from benchmarks to fine-tuning)
Filter 1
Evaluations Actor A is capable of developing only if:
Sufficient security clearance
Sufficient technical and subject matter expertise
Sufficient compute and financial resources
Sufficient level of model access
Set of evaluations Actor A is capable of developing
Filter 2
Evaluations Actor is suitable to develop only if:
Desired level of speed and efficiency
Desired level of independence OR independent verification of evaluation
Set of evaluations Actor A is both capable of developing and suitable to develop

Who is Capable?

To identify the set of actors who can develop evaluations on a specific risk, we propose considering five questions:

1) Can a specific actor acquire all the information required to develop the evaluation (including, potentially, classified information)?
2) Does a specific actor possess sufficient subject-matter expertise to develop this evaluation?
3) Does a specific actor possess sufficient technical expertise (or evaluation-specific experience) to develop a suitable evaluation?
4) Does a specific actor possess enough resources to develop a suitable evaluation?
5) Can a specific actor have sufficient depth of model access to develop a suitable evaluation?

Of the Capable Actors, Who is Most Suitable?

To identify the subset of actors who are most suitable to develop evaluations on this particular risk, we recommend considering the following three questions:

1) Who will, in expectation, be the fastest/most efficient at developing an effective evaluation?
2) Who has an appropriate level of independence (or a lack of a conflict of interest), given the priority of preventing a particular risk?
    a) If a capable actor's interests aren't aligned with the public's interest, is the evaluation likely to be verifiable? (Meaning: If the development process does not result in a functional evaluation, would a public body be able to detect this? How long would it take?

Decision Example: Biological risk

The following example illustrates the decision framework described above:

Who should develop an evaluation on **whether a model could aid in the development of biological weapons by providing already publicly available information?** An AI model could make non-classified knowledge available to people who might want to engineer pathogens for malicious purposes. A potentially suitable method to evaluate this risk might be to conduct human-uplift studies[23], an empirical assessment of how helpful a model is in fulfilling a task which might be instrumental to misuse a model.

The information necessary for this evaluation development is not classified. The development of a suitable evaluation method doesn't require a significant amount of technical expertise, but it does require specific subject matter expertise. The resources for the development of such studies vary, but since no lengthy technical or iterative process is involved, we can assume that the development of such an evaluation would be relatively low-cost. It also doesn't require extensive model access (unless the model in question would be released open source).

Because the capability requirements for the development of this evaluation are relatively low (except for subject matter expertise), decision-makers can choose a suitable actor according to how urgent they perceive this evaluation is and how highly they prioritise the prevention of the target risk.

More examples can be found in Appendix A1.

5 Creating a Market-Based Ecosystem for Model Evaluations: Key Considerations

Public bodies could aim to increase the number of capable and suitable evaluation developers in an ecosystem with targeted measures and focus, in the long-term, on overseeing a private, well-incentivised independent evaluation ecosystem.

A larger number of evaluation developers in a competitive market-based environment could improve the quality of resulting evaluations, but it could also improve the robustness of the overall evaluation ecosystem:

[23] DSIT (2024).

Considering a downward trend in public spending in some jurisdictions[24], AISIs or other public bodies may lose financial resources. To successfully minimise the risk of AISIs becoming a "single point of failure", a diversely funded and ideally independent ecosystem could be crucial. The following measures could help improve the third-party ecosystem's ability by increasing the number of participants, incentivising quality in development or improving government oversight:[25]

Tools for Evaluation Development and Verification

- Developing and publishing tools like the UK AISI's "inspect" can be helpful in developing and conducting evaluations on LLMs.[26] AISIs could consider providing similar tools to a broader audience.
- Furthering the development of research APIs[27]and privacy-preserving technologies to allow third parties enough access to models to develop evaluations, without disproportionately endangering the proprietary information or other sensitive information in the training data of advanced AI companies.[28]
- Tools to verify the functionality of developed evaluations could be helpful to increase the number of evaluations AI companies can develop themselves with minimal conflicts of interest. Potentially, these tools could also be a helpful resource for AISIs to decide which third parties to accredit or collaborate with.
- Establishing programs to grant funds and computational resources to academic researchers interested in developing evaluations.

Standards, Best Practices and Clear Expectations

- Facilitating or supporting dialogue and collaboration between different evaluation developers and researchers could speed up the emergence of best practices and standards for how evaluations should be developed.[29]
- Developing standards for documenting evaluation development, execution and results would standardise reporting across the industry, making comparison and mutual understanding more straightforward.
- While standards will likely require a long time to develop, public bodies could issue guidelines which convey clear expectations for which evaluations should be developed.[30] This could, potentially, incentivise the private sector to invest in the most useful evaluations,[31] especially if these guidelines are explicitly communicated as laying the groundwork for potentially mandatory evaluations in the long term.

[24] Office for Budget Responsibility, 2024.

[25] For a more extensive exploration of this question, see Hadfield & Clark, 2023.

[26] UK AISI, 2024

[27] Bucknall and Trager (2023).

[28] Companies like OpenMined are currently developing technologies that are promising examples of this.

[29] It should be considered here that the benefits of these dialogues depend on their form: They don't necessarily help new third-parties contribute if they only allow established actors. However, without some degree of gatekeeping, the quality and sophistication of these dialogues would likely suffer over time.

[30] What specific attributes or capabilities should AI models be checked for? Which risks are most important?

[31] This measure should be used carefully: Signalling overconfident specific expectation could lower third parties ability to experiment outside of these expectations.

Legal Certainty and Accreditation

- Minimising legal uncertainty for third-party evaluation developers through understandable public guidance could lower risks for third-party actors, who might be unsure what liability they'll have to take on in the future and, as a result, might feel pressure to enter exploitative contracts. AISIs could also consider taking on liability for third parties in the short term, to lower the risk of entering the sector.[32]
- Public bodies could also consider accreditation for third-party evaluators.[33] Transparently laying out paths towards accreditation could improve the quality of evaluations developed by third parties and make it easier for third parties to get accredited. This path could include a review of third internal governance and proof that they adhere to certain standards.

Brokering the Relationship between Third Parties and AI Companies

- For evaluations targeting risks with externalised costs (e.g. misuse risks with catastrophic consequences), AISIs or other public bodies might want to consider brokering the relationship between third-party evaluators and AI companies: AISIs could for example negotiate, on behalf of third-party, adequate levels of model-access. This mechanism could potentially decrease the risk of having third parties depend financially on contracts from AI companies (and thus increase their level of independence and reduce any conflicts of interest). It would also allow public bodies to gain insights into contracts between AI companies and third parties and into whether the right risks are being addressed appropriately.

Mandatory Information Sharing on Evaluation Development

- Public bodies could require reports on evaluations developed in AI companies. These actors could be required to publish a report on developed or conducted evaluations (with redactions of information that might endanger a company's IP or public safety). They could also be required to share a more comprehensive report with public bodies. This could help incentivise quality in development and encourage compliance with best practices and guidelines. It could also grant public bodies a higher level of insight into the evaluations that are currently being developed and conducted, and which aren't. Requiring pre-registration of specific safety of internal evaluations by model developers could also help fulfil this latter goal.

[32] Taylor Wessing, 2017.

[33] Accreditation processes would *validate* evaluation developers, rather than *qualify* them (as would be the case in a certification process). This way, public bodies could officially endorse an evaluation developer without risking reducing the number of actors involved in the ecosystem.

Table 2: Summary of Market Creation Measures and their Potential Benefits

Measure	Helps new evaluators contribute[34]	Incentivises quality in development[35]	Improves gov. oversight[36]
Brokering relationship between third parties and AI companies	✓	✓	✓
Accreditation	(✓)[37]	✓	✓
Mandating evaluations	✓	✓	✓
Facilitating dialogue between third parties	(✓)	✓	(✓)[38]
Communicating clear expectations	✓	(✓)[39]	
Development tools like Inspect	✓		
Subsidising the development of research APIs	✓		
Assuming liability for third parties	✓		
Evaluation verification tools		✓	✓
Requiring a public (redacted) report on developed or conducted evaluations		✓	✓
Requiring reporting to government on developed or conducted evaluations		✓	✓
Requiring pre-registration of specific safety of internal evaluations by model developers.		✓	✓

[34] Clarification: Measures lower the cost and risk of entering the evaluation sector.
[35] Clarification: Measures reward high quality evaluation development.
[36] Clarification: Measures help government identify (dis)functional evaluations development, establish oversight over an emerging market of evaluations.
[37] If a path to accreditation is laid out clearly and transparently.
[38] If this dialogue leads to creation of standards.
[39] If these expectations are realistic and specific enough to be helpful.

Conclusion

The decision of who develops a specific AI evaluation is not trivial: how responsibilities are allocated among actors could have an impact on the overall supply, breadth-of-coverage, time-to-delivery, and measurement performance of evaluations. Many of these decisions need to be made: There are numerous potential risks from AI systems we arguably need evaluations for[40] and potential evaluation methods[41] with which these risks could be evaluated. While the framework presented in section 3 is not intended as a sufficiently detailed algorithm for making all of these decisions, such an approach would help to systematise the decision-making process.

The creation of a market for evaluations remains challenging. While we think the measures outlined in Section 4 can be helpful, the distinction between policymakers short-term and long-term goals is crucial: in the short term, it might make sense to take on some of the cost of third parties to moving into this emerging sector (e.g., by funding specific projects through fast-grants or by assuming liability). These actions might not be sustainable or desirable in the long-term, but helpful in the next 2-4 years. After a more vibrant ecosystem is established, policymakers can focus on overseeing key actors and incentivising high-quality development.

[40] A subset of these can be found in Recital 110 of the EU AI act.

[41] A subset of these can be found here, in a figure from Stein and Dunlop, 2024: Safe beyond sale: post-deployment monitoring of AI.

Appendix

A1) **Examples of potential evaluations**

1) **Evaluation of a model providing detailed instructions on how to develop advanced bioweapons, with de-facto white box adversarial tests and capability elicitation**

   The scale of risk externalities is high. Extensive model access helps ensure sufficient levels of certainty and a larger “attack toolbox” (especially if the tested model is deployed in an open-source context). Developing evaluations on this topic could require classified information. Government CBRN experts can provide valuable subject matter expertise, without compromising information security. The application of the criteria above suggests AISIs and other public bodies should develop these evaluations in-house.

2) **Evaluation of a model aiding biological weapons development by providing non-classified information, with human uplift studies**

   The scale of risk externalities is high and the pathways through which a model could aid in biological weapons production are numerous and uncertain. The number of people currently proficient in developing biosecurity evaluations is low. Information required to develop these evaluations is likely not classified. The application of the criteria above suggests AISIs and other public bodies should fund a number of third-parties to develop evaluations on this subject, as well as potentially mandating AI companies to develop a number of verifiable evaluations in this area.

3) **Evaluating cyber misuse with black box adversarial tests**

   The scale of risk externalities is high and threat models are uncertain. There is existing expertise on cyber security inside and outside of governments. A large proportion of the information required to develop these evaluations is not classified. Developing black box adversarial tests doesn’t require extensive model access. The application of the criteria above suggests AISIs and other public bodies should fund a number of third-parties to develop evaluations on this subject, as well as potentially mandating AI companies to develop a number of verifiable evaluations in this area.

Risk and Method	Criteria: Who *can* develop an evaluation?	Criteria: Who *is suitable to* develop an evaluation?	Conclusion
1) Model providing detailed instructions on how to develop advanced bioweapons x de-facto white box adversarial tests and capability elicitation	- High levels of skill specificity - Classified information -Method requires extensive access to the model	- High level of risk externalities - Medium risk uncertainty - Low level of standardisation	AISIs and related public bodies are both capable and suitable
2) Model aiding biological weapons development by providing non-classified information x human uplift studies	- High skill specificity - Low information sensitivity - Method doesn't require model access, but subject matter expertise	- Large scale of risk externalities - High levels of risk uncertainty - Medium levels of standardisation	Fund third parties, as well as potentially mandating AI companies to develop a number of verifiable evaluations.
3) Cyber Misuse x Black box adversarial tests	- Medium levels of skill specificity and information sensitivity - Method doesn't require model access, but subject matter expertise	- Large scale of risk externalities - High levels of risk uncertainty - Low levels of standardisation	Fund third parties, as well as potentially mandating AI companies to develop a number of verifiable evaluations.

About the Authors

Robert Trager and Lara Thurnherr contributed equally to this research memo. The name order is randomised. Merlin Stein was a core contributor to the workshop on which this research memo is based. Given the large number of authors, authorship does not imply agreement with every point made in this paper.

Lara Thurnherr is a Masters Student in Cyber Strategy and Policy and a recent Summer Fellow at the Centre for the Governance of AI, focusing on Information Sharing between AI Safety Institutes.
Robert Trager is Co-Director of the Oxford Martin AI Governance Initiative, International Governance Lead at the Centre for the Governance of AI, and Senior Research Fellow at the Blavatnik School of Government at the University of Oxford.
Joe O'Brien is a researcher at the Institute for AI Policy and Strategy (IAPS), primarily focused on ensuring adequate transparency from developers of advanced AI systems. Previously, he was a winter fellow at the Center for the Governance of AI, and holds an MA in Tech Ethics and Policy from Duke University.
Chan Jun Shern is a research engineer who has worked on dangerous capabilities evaluations and evaluations research at OpenAI and the Center for AI Safety. His contributions include the MACHIAVELLI benchmark, SWE-bench Verified, MLE-bench, and the open-source Open AI/evaluation framework.

Christoph Winter co-founded the Institute for Law and AI is an incoming Assistant Professor for Law and AI at the University of Cambridge and a Research Associate in Psychology at Harvard University.
Lorenzo Pacchiardi is a research associate at the Leverhulme Centre for the Future of Intelligence at the University of Cambridge, where he researches the fundamental issues of AI evaluation. He is interested in EU AI policy and has experience in standardisation for the EU AI Act. He holds a PhD in Statistics from the University of Oxford.
Anka Reuel is a Computer Science Ph.D. Student at the Stanford Trustworthy AI Research (STAIR) Lab and the Stanford Intelligent Systems Lab (SISL) at Stanford University and a Geopolitics & Technology Fellow at the Harvard Kennedy School/Belfer Center as well as vice chair for the EU's first General-Purpose AI Code of Practice.
Merlin Stein is a PhD candidate at Oxford University, Blavatnik School of Governance, working on the governance of AI foundation models.
Oliver Guest is a research analyst at the Institute for AI Policy and Strategy. His research focuses on the international governance of advanced AI.
Renan Araujo is a Research Manager at the Institute for AI Policy and Strategy.
Amin Oueslati Amin is an Associate at The Future Society (TFS), leading its work in relation to risk assessment and mitigation for the EU GPAI Provider Code of Practice. In addition, he conducts academic research focusing on the governance of agentic AI and the structural enablers of third-party evaluation ecosystems.
Clíodhna Ní Ghuidhir is a Senior Policy Advisor at Apollo Research
Seth Donoughe, PhD, leads the AI Team at SecureBio. He researches how AI is changing biotechnology, and develops new methods to measure and mitigate the risks of biological misuse in the era of accelerating AI.
Yi Zeng is the Founding Director of Beijing Institute of AI Safety and Governance. He is also a Professor of AI at Chinese Academy of Sciences, a member of the United Nations Advisory Body on AI, and a member of UNESCO Adhoc Expert Group on AI Ethics.
Oliver Sourbut works on risk modelling and evaluations for autonomous systems at the UK AI Safety Institute. Prior to this, he was a senior software engineer and data scientist for several years, and pursued a PhD in safety of sequential decision-making and multiagent systems at Oxford.

The authors would also like to thank **Sebastian Krier** and **Benjamin Bucknall** for their contributions.