AIxBio Technical Working Group on Evaluation Practice

OVERVIEW

The AIxBio Technical Working Group on Evaluation Practice develops methods and reporting practices for comparing biological capability evaluations across organizations and jurisdictions. It seeks to identify best practices for interpreting findings to inform decisions about safeguards, access conditions, monitoring, deployment, or release. These efforts build on existing evaluation taxonomies by distinguishing what is evaluated from what the findings can establish. An evaluation may assess an underlying model, a configured system, a set of safeguards, or how AI access changes human performance. Depending on its design, the findings may support different conclusions about model or system capability, safeguard effectiveness, or human uplift. None of these evaluations directly measure biological risk.

NTI | bio and Concordia AI convene the Working Group through the AIxBio Global Forum, with technical input from SecureBio. Participants are affiliated with organizations that develop or conduct biological capability evaluations of AI models and systems in China, the European Union, the United Kingdom, and the United States. Evaluation methods are developing in parallel across these jurisdictions, with limited technical exchange among many evaluators. Differences in methods and interpretation can make findings attributed to the same model appear inconsistent, weakening the evidence base on which organizations and governments rely to assess risk and guide policy. By bringing together these evaluators, the Working Group creates a standing channel for technical exchange across principal jurisdictions where much of this work is underway.

Participation in Working Group meetings is by invitation, and the membership list is not published. The Chatham House Rule[1] governs attribution, allowing participants to compare published evaluation methods and discuss their documented limitations candidly without associating remarks with individuals or their organizations. Under the information-sharing rules described below, participants will examine published approaches through high-level, non-operational discussion and review draft work products[2] developed by the Working Group. These exchanges allow participants to compare evaluation practices across organizations and contribute to voluntary methodological guidance and reporting conventions that may inform evaluation practice across jurisdictions.

The initial scope covers general-purpose AI (GPAI) models, biological AI models (BAIMs), and interactions between the two, with a focus on capabilities relevant to deliberate biological misuse. The Working Group will not conduct or certify evaluations of individual models, audit or endorse organizational safety frameworks, or make deployment or release decisions on behalf of participating organizations.

The Working Group will not request or distribute proprietary, confidential, or operationally sensitive information, and participants must not disclose such information in meetings or materials.

[1] The Chatham House Rule governs attribution only; it does not authorize participants to disclose information they are not otherwise permitted to share.

[2] Work products will be derived solely from lawfully public information and high-level, non-operational discussion. Before circulation, they will be reviewed for compliance with the Working Group's information-handling rules.

WORK PROGRAM

The Working Group pursues four related lines of work.

1. Comparability of evaluation results. An evaluation result describes the performance of a specified AI system under a specified evaluation procedure. The system configuration includes the underlying model and any system prompts, inference settings, scaffolding, tools, and safeguards used in the evaluation. The procedure specifies the tasks, evaluator access, resources available to elicit performance, how performance is scored, and the baseline against which it is compared. Changes to either the system configuration or the procedure can alter measured performance without a new model release. For example, setting a model's reasoning effort to "high" instead of "max" can affect its evaluation score, but the setting is not always reported.

Reporting can show whether results differ because evaluators tested different systems or used different procedures; it cannot establish that two methods measure the same construct. Two findings may therefore support a comparison of task performance without supporting equivalent conclusions about underlying capability, safeguard effectiveness, human uplift, or biological risk.

⇒ Using published evaluations, the Working Group will identify which aspects of system configuration and evaluation procedure must be held constant, and which differences must be reported and accounted for, to support a valid comparison. It will define the reporting fields needed to support independent reproduction or validation of a result.[3] Where the published evidence supports comparison only at the level of task performance—or does not support direct comparison—the group will state that limitation and identify the additional comparative study needed before the findings are treated as comparable.

2. Capabilities relevant to biological risk. Inferring a change in biological risk requires a threat model that specifies the threat actor, deployment environment, and pathway to harm, together with evidence that the measured capability would remove or weaken a bottleneck in that pathway or increase the severity of harm. Higher scores on knowledge benchmarks may provide little evidence of increased risk if access to materials, hands-on expertise, or laboratory capacity remains limiting. By contrast, evidence that a system can troubleshoot failures or use scientific tools may support a stronger risk inference if those capabilities allow an actor to complete a step they could not otherwise perform.

⇒ The Working Group will map evaluation targets to existing, publicly documented threat models and risk pathways, compare published taxonomies and terminology, and identify terms or distinctions that lack a shared definition. At a non-operational level, the work will examine scientific reasoning and information retrieval, experimental troubleshooting[4], and the use of scientific tools and specialized biological models at different stages along these pathways. It will distinguish cases in which AI lowers barriers for an actor with less relevant expertise from those in which it expands what an already capable actor can accomplish.

[3] The Working Group will not request or exchange technical data or technology, access credentials, or non-public system configurations, benchmark content, evaluation data, source code, model weights, or implementation artifacts.
[4] The Working Group will not develop or exchange biological or model-specific troubleshooting guidance.

3. Validation and method development. Published evaluations vary in whether their tasks measure the capabilities they are intended to assess, whether their scores are reliable across repeated runs, raters, and comparable task forms, and whether their benchmarks remain discriminating as systems improve.

⇒ The Working Group will assess the validity and reliability of published evaluation methods. It will examine the construct validity of proxy tasks; score reliability across repeated runs, raters, or comparable task forms; the effect of elicitation, scaffolding, and other evaluation conditions on measured performance; benchmark contamination and saturation; and how capability-evaluation results relate to findings from uplift studies. It will identify capabilities that current evaluations omit or measure inadequately and set priorities for new or revised methods. Within the information-sharing boundaries, the group may outline non-operational designs for evaluation studies that organizations can implement independently; it will not request or exchange non-public benchmark content, evaluation data, or proprietary results.[5]

4. Decision relevance and evidentiary limits. Capability evaluations, safeguard evaluations, and uplift studies may use overlapping study designs, but they measure different objects and support different conclusions.

- **Capability evaluations** measure performance on specified tasks to estimate a capability of a model or configured system. Study designs may include benchmark evaluations or red-team exercises, including tasks conducted in agentic environments. The findings can support comparison across model generations only when the tested systems and procedures are comparable and the tasks remain valid and unsaturated. Low performance can help rule out a capability only when the test was sensitive to that capability and the system was adequately elicited; otherwise, it may reflect the evaluation rather than the system.
- **Safeguard evaluations** test whether specified controls prevent or limit harmful assistance under defined system, access, and evaluation conditions. They provide evidence about the effectiveness of those controls in the tested configuration, not about whether the underlying model lacks the capability the safeguards are intended to control.
- **Uplift studies** estimate how AI access changes human performance by comparing outcomes under defined treatment and control conditions. Their interpretation depends on who the participants are, what tasks they perform, what system access and other resources they receive, and whether the study has enough statistical power to detect an effect of practical significance.

⇒ The Working Group will specify the additional evidence or expert review needed to use each type of finding in a safeguard, access, monitoring, deployment, or release decision. This may include evidence connecting a capability result to a pathway to misuse, showing that a safeguard result applies to the intended deployment, or supporting generalization of an uplift result beyond the participants and conditions studied.

[5] The Working Group will not design or coordinate biological experiments or provide operational technical assistance for biological production, AI model development or modification, or the circumvention of safeguards.

INITIAL PRIORITIES

Initial products are expected to include:

- **A shared research agenda** that prioritizes research questions and measurement needs. This could include comparative validation of published methods, studies of how capability-evaluation results relate to uplift-study findings, and methods for capabilities that current evaluations omit or measure inadequately. The agenda may also identify organizations or research communities well placed to lead particular work and areas in which participants have expressed an interest in contributing.
- **A taxonomy of AIxBio evaluations** that builds on existing bio-specific classifications of evaluation methods and domains and general guidance on evaluation design, implementation, and reporting. It will organize evaluations by their objective and object tested; distinguish what is evaluated from what the findings can establish; provide common terms; and help determine when results are comparable.
- **A shared evaluation-reporting template** that identifies common fields for the evaluation objective; model or system tested; evaluation procedure; results and findings; assumptions; uncertainty; limitations; and intended decision use.
- **An evaluation-interpretation exercise** in which participants apply the same hypothetical or published finding to decisions about safeguards, deployment, and release under stated assumptions.[6] The exercise will distinguish whether differences among responses arise from interpretation of the evidence, the threat model and deployment assumptions applied, or an organization's thresholds and decision rules. The discussion will be held under the Chatham House Rule and de-identified in any written synthesis; no proprietary or non-public evaluation data will be used.

INFORMATION HANDLING

Participants agree to two limits on meeting materials and discussion: first, any information presented or discussed must have been lawfully published or otherwise lawfully made publicly available; second, any analysis of evaluation methods must remain high-level and non-operational. Participants remain responsible for complying with applicable law and their respective organizations' policies.

Participants must not share exact prompts; jailbreak or safeguard-bypass methods; model-specific vulnerabilities or attack methods; non-public technical artifacts or implementation details; biological protocols or troubleshooting guidance; or information that could materially increase biological or cyber misuse risk.

Organizations and individuals working on biological capability evaluations can contact Helia Samani at samani@nti.org to discuss participation or share related work. NTI will arrange a brief call to learn more about the prospective participant's current work, assess fit with the Working Group's scope, and discuss whether participation would be appropriate under the terms described above. NTI will not request proprietary or operationally sensitive information when assessing fit.

[6] De-identification or anonymization will not be used as a basis for sharing proprietary or otherwise restricted information.