APOLLO RESEARCH

Subject: Interim Report: Governing AI for Humanity, UN High-Level Advisory Body for AI
Date: 30th of March 2024.

About Apollo Research

Apollo Research is a non-profit AI evaluations research organisation specialising in evaluations for dangerous capabilities. Our current focus is on evaluating the capability for AI systems to evade human control, for example through deceiving either the user or its designer, and the prerequisites to this capability such as situational awareness. We conceptualise the capability of deception as a horizontal layer to other risks and capabilities an AI system may have, amplifying and obfuscating them. This makes our work sector agnostic and applicable across use cases.

A part of our research agenda, we also undertake mechanistic interpretability research with the goal of having a comprehensive understanding of what is driving AI systems' behaviours, capabilities and propensities, and are working on developing 'white-box'[1] evaluations.

Our work was selected to be showcased at the UK's AI Safety Summit hosted in Bletchley Park, November 2023; we are a partner to the UK AI Safety Institute, and a member of the US AI Safety Institute Consortium.

Apollo Research is committed to enabling internationally safe and beneficial AI innovation. We believe that good *international governance frameworks* can: raise the safety and security of AI development and deployment processes; increase consumer trust and raise uptake of beneficial AI applications; prevent the strategic exploitation of countries and jurisdictions with fewer resource to mandate and enforce ethical AI requirements by less scrupulous AI companies; and, help businesses plan and execute their market-access strategies.

[1] White-box evaluations are assessments of an AI model's characteristics that use information from the model's internals, such as activations, weights, or gradients. White-box evaluations are more thorough than black-box (behaviour-only) evaluations (Casper et al., 2024).

About AI System Evaluations

In our submission, we focus on contributing our unique expertise as an independent AI evaluator[2] to the revision of the Interim Report: Governing AI for Humanity. In particular, we focus on the role the UN could play as regards scientific consensus building for robust evaluation regimes, supporting the flourishing of an international AI evaluations ecosystem, and the interplay between governance efforts and evaluation regimes.

Evaluations increasingly underpin a variety of international governance efforts. They are reflected in coordinated efforts such as the Bletchley Declaration or the Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems; they form part of supra-national standard setting processes; and, inform interventions centred on increasing the safety and security of AI system development and deployment in governments' regulatory and policy frameworks, such as in the EU AI Act, the US White House Executive Order 14110 (EO) on AI, via the United Kingdom's AI Safety Institute, the Japanese AI Safety Institute, or the Singaporean AI Verify Foundation. Concurrently, leading AI companies are including evaluations in their responsible scaling policies, informing decisions around model development, deployment and safe scaling[3].

In consideration of the core role that evaluations play in international governance frameworks and their impact on the work of the UN Advisory Body for AI, we briefly detail *what evaluations are* below.

Apollo Research characterises evaluations as "the systematic measurement of properties in AI systems". We consider red-teaming[4] and benchmarking[5] to be distinct sub-components of evaluations. In short, evaluations examine:

- *what an AI system can do (i.e. capabilities)*: for example, the capability to solve a specific coding problem;

[2] Evaluations of frontier AI systems are most commonly undertaken either in dedicated teams within AI companies, or by external parties that can independently verify the safety and security of an AI system. The latter are often specialised to evaluate a model for a particular risk profile, such as CBRN capabilities.

[3] For an example, see Anthropic's RSP or this deployment guidance spearheaded by the Partnership on AI.

[4] Red-teaming is a type of evaluation that actively searches for specific capabilities while interacting with the specific AI system.

[5] Benchmarking is a type of evaluation that aims to identify the likelihood of an AI system behaving in a specific way on a certain range of inputs, typically to understand the likelihood of a behaviour occurring under real-use conditions.

- *the likelihood of the capability presenting across different scenarios or settings (i.e. propensities)*: for example, the tendency to be power-seeking, and;
- *the degree to which an AI system has a propensity to do things that are aligned with human intentions, or not (i.e. alignment)*: for example, how consistently an AI system completes a task as intended by the AI developers' training of or instructions given to it.

Together, a diverse range of evaluations and tests throughout an AI system's life cycle, capturing a multitude of threat models and capabilities, can contribute to a resilient 'defence in depth approach', strengthening existing and informing future governance mechanisms for AI systems.

Executive Summary

We commend the authors of the Interim Report: Governing AI for Humanity (henceforth: Interim Report) on the granular reflections pertaining to institutional functions. In this submission, we focus on leveraging the UN's unique role and expand on several specific institutional functions outlined in the Interim Report.

First, **we recommend the setting up of a UN AI Observatory**. We envision this body as particularly suitable to:

1. Lead on the advancement of the state of science, especially the 'science of evaluations'[6];
2. Serve as an amplifier and coordinator for a global network of evaluation environments, such as AI Safety Institutes and AI offices[7], and;
3. Lead on a cross-jurisdictional effort to track AI harms and incidents.

In more detail, we recommend that a UN AI Observatory would be uniquely well placed within international governance discourse and efforts to:

1. Advance the state of science, especially the 'Science of Evaluations'. In order to accurately identify and mitigate risks prior to the deployment of an AI system, we need adequate and robust evaluation and testing regimes. The field of evaluations is nascent and would benefit from an international body guiding its advancement.

[6] For more details, read Apollo Research's opinion piece on the need for a science of evals. We see the development of this field as a collective endeavour across evaluators, academia, international governments, as well as AI companies.
[7] For example, the EU AI Office, the United Kingdom's AI Safety Institute, or the Japanese AI Safety Institute.

Recommendation 1. A UN AI Observatory would be well placed to act as a hub for pertinent research exchange and support in kind[8], connecting researchers with international efforts and funding opportunities. As part of this, it could host yearly international research conferences on the science of evaluations. This could provide a ‘safe harbour’ for researchers and governments to exchange on the state-of-the-art, identifying priority research gaps and thereby expediting the advancement of the field.

2. Serve as an amplifier and coordinator for a global network of evaluation regimes. Efforts to establish evaluation regimes would benefit from designating an international coordinator to enhance cohesion and counteract fragmentation, diminishing the exploitation of potential cross-jurisdictional divergences in regimes and / or loopholes. This approach fosters coordination while allowing individual jurisdictions to customise governance mechanisms underpinned by evaluations to their specific needs and challenges.

Recommendation 2. A UN AI Observatory would be well placed to serve as an international interface between national evaluation environments, such as AI Safety Institutes. To speed up the rate at which novel safety measures can be identified and implemented across jurisdictions, we recommend that the UN AI Observatory host a protected communication channel between national bodies and collate and share relevant information among them. This enables national testing environments that nevertheless remain cohesive at the international level.

3. Lead a cross-jurisdictional effort to monitor and track AI harms, near-misses and incidents. AI progress can pose international risks, regardless of where the AI system was initially developed or deployed. As such, it is prudent to supplement and enhance non-governmental and national efforts to monitor, track and identify AI harms, near-misses and incidents under the remit of an international body. In turn, these data points can inform national and global governance regimes, highlighting blind spots and areas where swift and coordinated action is needed.

Recommendation 3. A UN AI Observatory would be well placed to aid the amplification of existing monitoring and tracking of AI harms, near-misses and incidents, as well as conduct supplementary monitoring and tracking efforts internationally. In turn, this can be leveraged to inform internationally relevant threat assessment work; to feed into appropriate governance interventions such as cross-border licensing regimes or international treaties; and to enable rapid coordinated responses to catastrophic vulnerabilities or critical incidents. Notably, the UN AI Observatory may be well placed to share monitoring best practices and encourage their adoption internationally.

[8] For example, connecting researchers to compute made available freely for relevant research projects.

Our Recommendations

"*A global governance framework is needed for this rapidly developing suite of technologies and its use by various actors, be they the developers or users of the technology. AI presents distinctly global challenges and opportunities that the UN is uniquely positioned to address, turning a patchwork of evolving initiatives into a coherent, interoperable whole, grounded in universal values agreed by its member states, adaptable across contexts.*" p.6, Interim Report

AI development is rapidly progressing and corresponding governance efforts across the world are starting to take concrete shape – from standardisation, regulation, to institution building. In light of this, a promising addition to the existing ecosystem could be an international body, taking the role of an overarching coordinator for: scientific research efforts; nascent evaluation regimes; and mitigation measures enhancing safety and security. We are cognisant that duplications within the ecosystem can have counterproductive effects and are confident that the light-touch remit we sketch in this submission is a suitable and worthwhile complement to existing bodies.

Below, we outline our core recommendation, to establish a UN AI Observatory, followed by three institutional functions we see as particularly apt for such a body, and beneficial to the broader international ecosystem.

We recommend the setting up of a UN AI Observatory.

Historically, observatories have been established to exchange on, measure and survey natural occurrences, for example, astronomical, geophysical or meteorological events. In this spirit and for the field of AI, we envision the UN AI Observatory as a body staffed by technical and non-technical personnel with a range of subject matter and AI-relevant expertise, to coordinate relevant scientific coordination and enquiry; support exchange on evaluation and testing regimes; and, conduct monitoring and tracking efforts for AI harms, near-misses and incidents.

We note that over the past years, multiple efforts have been made to establish discrete AI Observatories, such as, for example, from the OECD, the EU, Québec, or Italy. These observatories contribute relevant repositories and data sets to the wider ecosystem, ranging from collating international AI strategies, societal sentiments, to providing relevant tools and metrics and hosting expert groups. In order to harness existing efforts and avoid duplication, a UN AI Observatory should collaborate with existing platforms such as these, as well as international bodies with relevant subject expertise, where appropriate.

Notwithstanding that, the institutional functions of *the UN AI Observatory detailed in this submission are distinct from existing functions in any of the aforementioned*, as well as from the remits of any other existing bodies, to the best of our knowledge. As such, a UN AI Observatory executing on the functions we recommend would present a novel – and importantly, value-adding – complement to the ecosystem.

1. Advance the state of science, especially the 'Science of Evaluations' [9].

Fundamental to the development of robust evaluations, and, therefore, a functional governance ecosystem based on evaluations, is a healthy field of scientific research. Yet, this field, 'science of evaluations'[10], is only just starting to take shape. An overreliance on current methods, without significant efforts to develop and fund research to advance the science of evaluations can lead to unexpected and harmful downstream effects, such as the deployment of unsafe AI systems or a mis-classification of AI system risks based on unreliable evaluation outcomes.

A UN AI Observatory acting as a hub for research exchange, coordination between international research efforts and liaising between available international research support, such as through e.g. compute access, and researchers, could significantly expedite the advancement of the field. Conversely, a flourishing science of evaluations will support the development of increasingly more adequate governance interventions, for example, by feeding into risk classifications. The UN AI Observatory could convene on relevant topics such as:

- **Accuracy**; how to ensure evaluations accurately measure the intended property, and not a proxy measure, as well as quantifying confidence in the evaluation's accuracy, and;
- **Consistency**; how to ascertain statistical confidence in the repeatability of a type of an evaluation's results (e.g. prompt engineering, fine-tuning); and ensuring biases in evaluations are measured and managed.

In addition, we suggest that the UN AI Observatory supports a 'system / life cycle approach' within the science of evaluations. In order to achieve an overarching and comprehensive approach towards safety and security of AI systems via evaluations, the field ought to take the full lifecycle of an AI system into consideration. This can include a range of relevant audits, such as training design or governance audits, but for the purpose of this submission we focus on

[9] This recommendation corresponds to '*Institutional Function 1: Assess regularly the future directions and implications of AI*' and '*Institutional Function 5: Promote international collaboration on talent development, access to compute infrastructure, building of diverse high quality datasets and AI-enabled public goods for the SDGs*'.

[10] For more details, read Apollo Research's opinion piece on the need for a science of evals. We see the development of this field as a collective endeavour across evaluators, academia, international governments, as well as AI companies.

relevant actions post-deployment. While most evaluations to date focus on pre-deployment testing for AI systems, substantial modifications (or even a cumulation of more simple modifications) and a change in ‘available affordances’[11] can alter the risk profile of an already deployed AI system in unpredictable ways. This may require the AI system to undergo a repetition of previous evaluations and tests, as well as updated evaluations and tests appropriate to the new risk profile. In particular, we suggest further research into:

- **Evaluations throughout an AI system’s life cycle.** The risk profile of an AI system can be changed through both meaningfully updating the original AI system (including a succeeding of smaller minor updates) and by providing the AI system with a novel set of ‘available affordances’. We propose ‘available affordances’ to describe the set of affordances, such as internet access or access to new datasets, that significantly change the environmental resources and opportunities for affecting the world that are available to an AI system. Changes to an AI system’s risk profile such as those outlined, ought to be accompanied by re-evaluations of the AI system’s capabilities, and, potentially, new sets of evaluations, appropriate to the new risk profile of the AI system.

Recommendation 1. A UN AI Observatory would be well placed to act as a hub for pertinent research exchange and support in kind[12], connecting researchers with international efforts and funding opportunities. As part of this, it could host yearly international research conferences on the science of evaluations. This could provide a ‘safe harbour’ for researchers and governments to exchange on the state-of-the-art, identifying priority research gaps and thereby expediting the advancement of the field.

2. Serve as an amplifier and coordinator for a global network of evaluation environments[13].

Evaluations and testing for dangerous capabilities like chemical, biological, radiological and nuclear (CBRN) or loss of control, as well as identification of misuse and malicious use risks can inform safety and mitigation strategies across regions. In turn, these can enable the implementation of governance frameworks sensitive to and tailored to regional challenges.

A UN AI Observatory would therefore be particularly well placed to both collate findings and pertinent updates from a range of evaluation and testing endeavours from pertinent bodies (e.g., AI Safety Institutes, AI offices) and subsequently disseminate them among their international counterparts. This will aid the establishment and maintenance of cohesive international testing

[11] You can read about this topic in more detail in our publication entitled ‘A Causal Framework for AI Auditing and Regulation’, especially section 2.3.1.

[12] For example, connecting researchers to compute made available freely for relevant research projects.

[13] This recommendation corresponds to ‘*Institutional Function 3: Develop and harmonize standards, safety, and risk management frameworks*’.

environments, while enabling national specifications. Below, we list a number of suggestions and benefits corresponding with this institutional function:

- **Enable knowledge transfer in order to act as a value maximizer**. Establish appropriate channels to act as a core interface sharing knowledge and expertise between existing and future AI Safety Institutes and national AI Offices, liaising on evaluation and testing regimes for AI systems. It is unlikely that all countries will have sufficiently robust testing regimes within the coming 5 years. Therefore we expect it to be highly beneficial to:
    - Support the amplification and fair distributions of relevant work done by a small number of global researchers, including to nation states without adequate local talent comparable to the AI challenges they encounter.
- **Counteract global fragmentation**. The provision of a mechanism for sharing and articulating relevant efforts, lessons learnt, and identifying blindspots will quickly become crucial in reducing the likelihood of less scrupulous AI companies or users from exploiting loopholes within and / or between nascent regional evaluation and testing regimes. For example, divergences in what constitutes a 'substantial change' in a deployed model, meriting re-evaluation.
- **Support the establishment of international best practices and norm setting**. Coordinate and amplify discussions and standard setting activities surrounding these institutes, their best practices, and lessons learnt. This could include bringing expertise from technical expert bodies and standards setting institutions, for example, NIST, CEN CENELEC or the ISO and international governments.
- Ambitiously, this could also **support the continuation and implementation of relevant next steps agreed upon during country-led AI Safety Summits**. As a start, the UN AI Observatory could set up a team acting as secretariat supporting future country-led AI Safety Summits.

We envision that the aforementioned can complement an ambitious international effort involving **global horizon scanning** by tracking AI progress and mapping of emerging threats. These can, e.g.:

- Contribute to more robust anticipatory measures;
- Inform future needs for evaluation and testing regimes[14]; and,
- Enable swift emergency responses, if needed.

In Section 3, we discuss in more detail the initial shape this may take at the UN AI Observatory.

[14] In turn, these could be supported by research and funded via mechanisms outlined under Section 1 in this submission.

Recommendation 2. A UN AI Observatory would be well placed to serve as an international interface between national evaluation environments, such as AI Safety Institutes. To speed up the rate at which novel safety measures can be identified and implemented across jurisdictions, we recommend that the UN AI Observatory host a protected communication channel between national bodies and collate and share relevant information among them. This enables national testing environments that nevertheless remain cohesive at the international level.

3. Lead on a cross-jurisdictional effort to monitor and track AI harms and incidents[15].

AI can lead to harm and incidents across borders, not 'just' in the country where the AI system originated from or was primarily deployed in. It is therefore prudent to establish an oversight framework at international level, such as an 'incident database'. We note that several efforts[16] exist to monitor, track and identify AI harms and incidents. Taking an international perspective, a UN AI Observatory would be well placed to complement these. Data collection on incidents, harms, mitigation measures, as well as evaluation results can contribute to robust oversight frameworks, enabling evidence-based policy measures proportionate to proven benefits and risks posed by AI. Beyond its benefit in informing national and international governance regimes, a global AI harms and incident database can support the identification of areas of concern where swift and coordinated action is needed. In short, we see this function sitting across national governance regimes. Below, we sketch this institutional function:

- **Global network for data collection on harms and incidents.** This would take the form of a UN AI Observatory database detailing near misses, harms and incidents: (i) reported by companies directly; (ii) shared by national databases. Where national databases are unavailable or unlikely to be a policy priority, the UN AI Observatory would act as a temporary stand-in, collating data directly from companies and sharing relevant geographical information received with the respective government for further action. This can be supplemented by further mechanisms enabling the capturing of a wide range of harms and incidents, such as:
    - Mechanisms for the machine learning community to report harms, including their analysis. Taking inspiration from the cybersecurity community, this could be done through creation of 'bounties', or establishing routes to enable credible AI researchers to share vulnerabilities or hazards they have detected.
    - Subject to consent and buy in from nation states, establish international whistleblower pipelines. The UN AI Observatory may want to establish secure

[15] This recommendation corresponds to '*Institutional Function 6: Monitor risks, report incidents, coordinate emergency response*'.

[16] See for example Partnership on AI's AI Incident Database or the reporting database led by the OECD.

channels for whistleblowers, allowing relevant information captured to benefit all nations, whilst protecting individuals' anonymity.

- **Lead on and coordinate threat assessment and interventions.** Access to and insight into pertinent datasets would enable the UN AI Observatory to identify arising threats within the landscape and instances that necessitate timely coordinated action at an international level. For example, taking an AI system that has caused significant harm in one region quickly off the market in others. Over time, trends and relevant patterns could emerge, allowing for more refined foresight and interventions, such as developing clearer 'red-lines' on an international scale.

We envision that efforts such as the one recommended in this section, complemented by the proposal in section 2 can eventually contribute to the development of an international Treaty on AI. An effort such as this, could build on and complement recent developments in that direction such as the UN AI General Assembly AI Resolution and the Council of Europe's Convention on Artificial Intelligence, Human Rights, Democracy, and the Rule of Law.

Recommendation 3[17]. A UN AI Observatory would be well placed to aid the amplification of existing monitoring and tracking of AI harms, near-misses and incidents, as well as conduct supplementary monitoring and tracking efforts at an international level. In turn, this can be leveraged to inform internationally relevant threat assessment work; to feed into appropriate governance interventions such as cross-border licensing regimes or international treaties; and to enable rapid coordinated responses to catastrophic vulnerabilities or critical incidents.

Conclusion

We commend the UN Advisory Body for AI on the breadth of considerations outlined in its Interim Report and the well thought through draft schema for institutional structures supporting these considerations. As a next step, we recommend the Advisory Body for AI focus on the implementation of a suitable institutional infrastructure, complementing existing efforts and alleviating blindspots. In our response, *we propose that a UN AI Observatory may be best suited for such an endeavour*. Our response is informed by our expertise in the field of evaluations and its implications on international governance efforts. As such, we homed in how a UN AI Observatory's institutional functions could: advance the state of science; convene the development of global evaluation and testing regimes; and, act as an alert system for border transgressing risks, all the while empowering jurisdictional independence and adjustment. We thank the UN Advisory Body for AI for the opportunity to provide feedback.

[17] This recommendation corresponds to '*Institutional Function 6: Monitor risks, report incidents, coordinate emergency response*'.