Input to UN Interim Report on Governing AI for Humanity

– April 2024 –

This document was written by Bradford Saad, with assistance from Andreas Mogensen and Jeff Sebo. Jakob Lohmar provided valuable research assistance. The document benefited from discussion with or feedback from Frankie Andersen-Wood, Adam Bales, Ondrej Bajgar, Thomas Houlden, Jojo Lee, Toby Ord, Teruji Thomas, Elliott Thornley and Eva Vivalt.

Input to UN Interim Report on Governing AI for Humanity

[Questions from input form, responses]
After reviewing the Interim Report, please provide your feedback on the following sections:
(If you have no input for a specific question below, you may leave it blank)

Opportunities and Enablers
(Maximum 3,000 characters)

Recommendation concerning openness:
The Interim Report appears to adopt a broadly favorable stance toward openness in AI (Box 1, #19, Figure 1, #67, #68, Institution Function #5). In line with #17, we recommend adopting a more nuanced and critical outlook on openness in AI throughout the report, an outlook that is more explicit about associated risks. (See Seger et al. (2023) and Harris et al. (2024)—all references are linked below.)

- For example, rather than emphasizing the benefits of openness by discussing it in the noted parts, the report might instead include something like the following in the 'Challenges to be addressed' section:

> "Openness in AI carries the potential for benefits and harms. While openness offers a mechanism for broadly distributing benefits of AI, it is one such mechanism among many. Some forms of openness—such as model weight sharing—harbor the potential to irreversibly lead to the widespread availability of harmful AI capabilities. This could happen even with an initially safe model: after its weights are published, safeguards could be removed; fine-tuning could elicit dangerous capabilities that were previously latent; or such capabilities could be added by modifying or augmenting the model. An AI governance framework should be sensitive to these tradeoffs, bearing in mind that selective openness (e.g. openness with respect to safe and beneficial services but not model weights) may provide a way to capture many of the associated benefits while mitigating risks."

Risks and Challenges
(Maximum 3,000 characters)
Recommendation concerning existential risk:
The Interim Report briefly alludes to the risk of AI causing an existential catastrophe for humanity (#29). However, the Interim Report could be read as inviting doubt about whether this threat is to be taken seriously: the threat is referenced indirectly (as the content of a concern) rather than directly, in a qualified manner (as a possible threat), and with reservation (per the remark on debate about whether to assess such threats). Elsewhere, the Interim Report devotes little direct attention to existential risk from AI, and in some places (e.g. #2, Box 3, #24, Institutional Function 6) conspicuously omits explicit reference to it. Further, the Interim Report alludes to the precautionary principle (#31) but does not explicitly acknowledge the principle's application to existential risk from AI.

However, it is a mainstream position among leading AI labs, international leaders, and relevant experts that advanced AI might cause civilizational collapse or human extinction and that global governance

measures may be needed soon to mitigate this risk. We suggest that serious consideration and extended discussion of existential risk from AI (understood as risk of AI causing human extinction or a similarly bad and irreversible outcome) should be added to the revised report and that it recommend a precautionary approach to existential risk from AI.

For example, the qualified mention of existential risk could be removed and a numbered item could be added afterward (circa #31) that says (something like):

> "Many domain experts believe that AI will pose an existential threat to humanity in the coming decades (for many examples, see Bengio et al. (2023), and Harris et al. (2024)). While there is uncertainty and expert disagreement about the probability of an existential catastrophe from AI, a precautionary approach recommends seeking effective measures for mitigating this risk, even in the face of scientific uncertainty. The situation here parallels other large-scale threats such as future pandemics, climate catastrophes, and nuclear war. Despite attendant uncertainties, each of these threats merits serious consideration in governance, as they threaten humanity with large-scale catastrophes that would be difficult to reverse. With appropriate governance mechanisms, each of these threats can be monitored, evaluated, and mitigated. As with these other large-scale risks, existential risk from AI may warrant a more cautious approach than is warranted for AI technologies whose potential negative consequences would be less severe and easier to remedy."

Short of according a more prominent role to existential risk from AI in the report, we suggest acknowledging (for example, just after #31) that different levels of caution within AI governance may be required for different risks, with high levels of caution being appropriate for risks of outcomes that would be extremely severe and difficult to remedy.

Guiding Principles to guide the formation of new global governance institutions for AI (Maximum 3,000 characters)

Recommendation concerning the precautionary principle:
While the report does a commendable job of highlighting potential benefits of AI, we believe the report should provide clearer and additional acknowledgement of downside risks of this technology (while recognising scientific uncertainty around these risks), expanding on ideas in #29 and #31. This suggestion could be implemented by including the precautionary principle as an additional guiding principle. Borrowing wording from the UN General Assembly's Rio Declaration on Environment and Development, the principle might be phrased as follows: "Where there are threats of serious or irreversible damage, scientific uncertainty concerning risks from AI shall not be used as a reason for postponing cost-effective preventive measures." As a point on this guiding principle, it could be noted that it has application to risks of AI causing an existential catastrophe, large-scale harms to a wide range of vulnerable populations (about which more below), and risks from open sourcing model weights.

If the report does not incorporate the precautionary principle as a guiding principle, it could instead note, in connection with Institution Function 3 (e.g. just after #64), that: "The UN could also play a critical role in helping build consensus about what precautionary measures to take against the most severe AI-involving risks."

Recommendation concerning Guiding Principle 1.
As written, Guiding Principle 1 says that AI should be governed inclusively, by and for the benefit of all. However, since not all moral stakeholders are capable of participating in governance, this formulation of the principle can be read as using a restricted use of 'all' that excludes some moral stakeholders. The category of moral stakeholders that may not be capable of participating in governance may include some humans (infants, some cognitively impaired humans, and future humans), some non-human animals, and some AI systems that might exist in the future and might have capabilities that confer moral standing. To prevent this exclusionary reading, we recommend changing the principle to say "AI should be governed inclusively and for the benefit of all."

We also recommend adding a point on the principle (e.g. just after #47) along the lines of: "AI governance should also be for the benefit of all moral stakeholders who may be unable to participate in the governance process. Such stakeholders may include some (present and future) humans, some non-human animals, and some AI systems that might exist in the future and possess capabilities that confer moral standing."

***Institutional Functions* that an international governance regime for AI should carry out (Maximum 3,000 characters)**

Recommendation concerning Institution Function 5: We suggest that this function should not include the promotion of (responsible) sharing of open-source models, as including that is: (1) questionable (per open questions about the safety of open models acknowledged in #17) and (2) premature (as whether promoting open source-models should be an element of institutional function is an issue to be addressed downstream of further investigation of associated risks and benefits).

Recommendation concerning Institution Function 6: In line with the potential for rapidly emerging threats from AI noted in #70 and the need to be proactive acknowledged in #32, we suggest that Institution Function 6 should include not only the coordination of emergency responses but also advanced preparation for emergency responses and contingency planning.

- The report could recommend AI crisis simulation exercises in which representatives from different States and relevant organizations play out different scenarios: for example, a scenario in which an open source model is discovered to pose a biosecurity risk, in which a system deviates from its intended behavior and instigates cyberattacks or copies itself to other data centres, or in which it becomes known that models exhibit various markers of moral standing, resulting in public pressure to provide legislative guidance within a short time frame.
- As a safeguard against dangerous technology-race dynamics, contingency planning could seek to secure conditional commitments for how States will respond if certain capability or compute thresholds are exceeded. Such preparations would provide an appropriate context in which to explore unusual governance mechanisms—such as coordinated pauses on development, mutually agreed upon limits on deployment, strict international oversight of supply chains and safety testing—that might be required, given unprecedented aspects of AI technology. For relevant discussion, see Harris et al. (2024) and President Biden's Executive Order on AI (2023).

Other comments on the International Governance of AI section (aside from Principles and Functions, covered in above questions)
(Maximum 3,000 characters)

Recommendation to address AI systems with moral standing: In line with the Interim Report's aim to promote a universal and inclusive approach to AI governance, we believe that the revised report should explicitly take into account future AI systems that could - like humans and other animals - have moral standing, due to their (non-negligible chance of) possessing mental capacities for consciousness, sentience, and/or agency. As noted above, we recommend an approach to AI governance that is for the benefit of all moral stakeholders, including humans and other animals (Singer & Tse, 2023). However, we focus here on recommendations concerning potentially morally significant AI systems, since this topic is increasingly important yet still unknown by most policymakers.

While there are profound theoretical issues concerning which types of AI systems, if any, can have such mental capacities, many domain experts regard AI systems with such capacities as a live possibility. Compare: in a recent survey of professional philosophers (Chalmers & Bourget, 2023), only 9.69% of philosophers of mind rejected consciousness in future AI systems; and in a recent survey of consciousness scientists, only 3.03% denied that present and future "machines (e.g. robots)" could have consciousness (Francken et al. 2022). Moreover, many domain experts who are uncertain whether AI systems will exhibit such capacities nonetheless think AI systems will at least be *good enough candidates* for exhibiting those capacities to be owed moral and legal consideration. (See Butlin et al. (2023), Chalmers (2023), Schwitzgebel & Garza (2015), Sebo & Long (2023), Shulman & Bostrom (2021).) Current evidence on this score calls for pro-actively exercising humility and caution, not dismissing or ignoring the relevance of potentially morally significant AI systems to AI governance. As a first step in this direction, the report could note (e.g. in discussing Guiding Principle 1 just below #46) that "AI governance should include evaluations of advanced AI systems for markers of consciousness, sentience, agency, and other morally significant capacities. And if and when there is a realistic chance that advanced AI systems have moral standing, AI governance should consider how governance decisions would affect their welfare."

Any other feedback on the Interim Report
(Maximum 3,000 characters)
Recommendations concerning the implications of AI systems with moral standing.

- The report could note in the 'Challenges to be addressed' section that "AI governance must navigate, and perhaps advance ongoing efforts to reduce, uncertainty about which AI systems will have moral standing, as mistakes on this score could engender mistreatment of AI systems with moral standing and disagreements about which AI systems have moral standing could pose obstacles to coordinating norms concerning their treatment. (See Schwitzgebel & Garza (2015).)"
- The report could note in the 'Risks of AI' section that "There is a real possibility that we will create AI systems with moral standing and then - in line with economic incentives, and perhaps unknowingly - inflict suffering upon them or violate their rights by treating them as mere tools. This could quickly happen on a large scale, resulting in a moral catastrophe. AI governance should mitigate this catastrophic risk."

- The report could note in the 'Risks of AI' section that "The prospect of AI systems with moral standing complicates the evaluation and mitigation of AI-involving risks. For example, there could be moral costs associated with deleting, restricting, or reprogramming AI systems that generate cyberattacks or political misinformation. Similarly, ensuring that autonomous AI systems do not harm humans may require invasive monitoring and control that would, if applied to an AI with moral standing, violate rights listed in the Universal Declaration of Human Rights. AI governance should address such complications and explore options for avoiding them."
- The report could, in discussing Institution Function 7, note that "The creation of AI systems with moral standing could threaten compliance with international norms, the legitimacy of institutions, and global stability. For example, existing institutions and norms may be ill-suited to addressing the question of when, if ever, AI systems that exhibit different markers of moral standing should be not only protected against possible forms of harm but also potentially recognized as legal persons or political citizens. AI governance should help humanity navigate the danger of creating AI systems with moral standing that would either be mistreated or have destabilizing effects on existing norms and institutions. It cannot be ruled out that this would require international restrictions on which types of AI systems may be created."

Among the next steps mentioned in the Interim Report are deep dives on certain topics. We suggest that the list be expanded to include a deep dive on existential risk from AI and a deep dive on AI systems with moral standing.

We are grateful to the Advisory Body for considering our input. We are open to dialogue with Body members about our suggestions.

The views expressed herein are broadly agreed upon by signatories. They do not necessarily reflect the views of their respective institutions.

Signatories:

Lucius Caviola, Senior Research Fellow, Global Priorities Institute, University of Oxford
Sofia Fogel, Coordinator, Mind, Ethics, and Policy Program, New York University
Riley Harris, Research Assistant, Global Priorities Institute, University of Oxford
Thomas Houlden, Predoctoral Research Fellow, Global Priorities Institute, University of Oxford
Jojo Lee, Predoctoral Research Fellow, Global Priorities Institute, University of Oxford
Joshua Lewis, Assistant Professor, New York University
Jakob Lohmar, Research Assistant, Global Priorities Institute, University of Oxford
Andreas L. Mogensen, Senior Research Fellow, Global Priorities Institute, University of Oxford
Toby Ord, Senior Research Fellow, University of Oxford
Bradford Saad, Senior Research Fellow, Global Priorities Institute, University of Oxford
Johanna Salu, Predoctoral Research Fellow, Global Priorities Institute, University of Oxford
Jeff Sebo, Director of the Mind, Ethics, and Policy Program, New York University
Toni Sims, Researcher, Mind, Ethics, and Policy Program, New York University
Teruji Thomas, Senior Research Fellow, Global Priorities Institute, University of Oxford
Elliott Thornley, Postdoctoral Research Fellow, Global Priorities Institute, University of Oxford
Hayden Wilkinson, Postdoctoral Research Fellow, Global Priorities Institute, University of Oxford
Timothy Luke Williamson, Postdoctoral Research Fellow, Global Priorities Institute, University of Oxford

References

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Harari, Y. N., ... & Mindermann, S. (2023). Managing AI risks in an era of rapid progress. arXiv preprint arXiv:2310.17688.

Bourget, David & Chalmers, David J. (2023). Philosophers on Philosophy: The 2020 PhilPapers Survey. *Philosophers' Imprint* 23 (1).

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., ... & VanRullen, R. (2023). Consciousness in artificial intelligence: insights from the science of consciousness. arXiv preprint arXiv:2308.08708.

Chalmers, D. (2023) Could a large language model be conscious?. Philarchive.

Francken, J. C., Beerendonk, L., Molenaar, D., Fahrenfort, J. J., Kiverstein, J. D., Seth, A. K., & Van Gaal, S. (2022). An academic survey on theoretical foundations, common assumptions and the current state of consciousness science. Neuroscience of consciousness, 2022(1), niac011.

Harris, E., Harris, J., & Beall, M. (2024) Defense in Depth: An Action Plan to Increase the Safety and Security of Advanced AI. Report commissioned by US State Department.

Schwitzgebel, Eric & Garza, Mara (2015). A Defense of the Rights of Artificial Intelligences. Midwest Studies in Philosophy 39 (1):98-119.

Sebo, J., & Long, R. (2023). Moral consideration for AI systems by 2030. AI and Ethics, 1-16.

Seger, E., Dreksler, N., Moulange, R., Dardaman, E., Schuett, J., Wei, K., ... & Gupta, A. (2023). Open-sourcing highly capable foundation models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives. Centre for the Governance of AI.

The White House (2023) Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.

Shulman, C., & Bostrom, N. (2021). Sharing the world with digital minds. In S. Clarke, H.
Zohny, & J. Savulescu (Eds.), *Rethinking Moral Status*. Oxford University Press.

Singer, P., & Tse, Y. F. (2023). AI ethics: the case for including animals. AI and Ethics, 3(2), 539-551.