The Manhattan Declaration on Inclusive Global Scientific Understanding of Artificial Intelligence

In the context of the High-Level meeting "*Towards a Common Understanding of AI Capabilities, Opportunities, and Risks: Forging the Path for a Positive Future for All,* " convened at United Nations (UN) headquarters during the 79th session of the UN General Assembly (UNGA), we, the undersigned scientists and researchers in artificial intelligence and technology policy, put forth this declaration.

We recognize the urgent need for a shared global understanding of AI capabilities, opportunities, and risks as highlighted by recent developments, including the release of the recommendations of the UN Secretary-General's High-Level Advisory Body on Artificial Intelligence (HLAB-AI) and the release of the *International Scientific Report on the Safety of Advanced AI (Interim).*

Developments in AI, particularly in foundation models, have demonstrated both the beneficial potential and significant risks in the development and use of these technologies. As we approach the development of more powerful systems, it is crucial that we come together as a global scientific community to anticipate the challenges ahead and to support a safe and beneficial use of AI.

We declare the need for:

1. **Global scientific cooperation**: We call for enhanced international collaboration on AI research, particularly on issues of AI safety, ethics, and societal impact. No single country or organization can address these challenges alone.
2. **Assessment of capabilities and mitigation of AI risks**: We recognize the pressing need to assess the capabilities and associated risks, and address the challenges posed by AI systems, especially increasingly capable ones. We commit to prioritizing research on AI alignment with human values, beneficial impacts and robustness, as well as the mitigation of both ongoing harms and anticipated risks.
3. **Fostering AI as a global public good**: We reaffirm our commitment to developing AI systems that are beneficial to humanity and acknowledge their pivotal role in attaining the global Sustainable Development Goals, such as improved health and education. We emphasize that AI systems' whole life cycle, including design, development, and deployment, must be aligned with core principles, safeguarding human rights, privacy, fairness, and dignity for all.
4. **Inclusive participation**: We emphasize the importance of including diverse perspectives from researchers worldwide, regardless of their backgrounds, geographic locations, or institutional affiliations. The impacts of AI are global, and our approach to its development and governance must be fair and equally inclusive.

5. **Transparent research and risk assessment**: We commit to promoting open science and transparent research practices in AI, particularly for work that has significant implications for global AI governance and safety.
6. **Interdisciplinary approach**: We recognize that seizing the opportunities and addressing the challenges posed by AI will require insights from various fields, including computer science, information and hardware security, ethics, economics, cognitive science, neuroscience, social science, cultural studies, mathematical sciences, and more. We commit to interdisciplinary and transdisciplinary collaboration.
7. **Responsible development, deployment and use**: We advocate for a measured approach to AI design, development, deployment, and use that prioritizes beneficial uses (such as the SDGs) and safety, ethical considerations, and societal benefit over rapid advancement at any cost.
8. **Support for governance initiatives**: We back efforts by the UN and other national, regional, and international organizations to develop evidence-based governance frameworks for AI, and that, among other things, foster interoperability and minimize fragmentation, recognizing that good governance can serve as a key enabler for innovation in the public interest. We commit to providing scientific expertise to inform these initiatives.
9. **Public engagement**: We recognize the importance of, and will take actions on, engaging with policymakers and the public to build a shared understanding of AI capabilities, benefits, limitations, and potential impacts, empowering society to make informed choices.
10. **Long- and near-term perspectives**: We commit to considering the long-term implications and impacts on future generations of AI development, including potential major global risks and transformative benefits, in our research and recommendations, while not discounting the current day and near-term risks and harms.

We urge AI scientists and technology-policy researchers worldwide and across sectors to join us in this commitment to responsible AI development and international collaboration.

We invite policymakers and member states to actively engage with the global scientific community in addressing the critical challenges posed by AI.

As AI scientists and technology-policy researchers, we advocate for a truly inclusive, global approach to understanding AI’s capabilities, opportunities, and risks. This is essential for shaping effective global governance of AI technologies. Together, we can ensure that the development of advanced AI systems benefits all of humanity.

Signatories

- **Yoshua Bengio (Declaration’s co-sponsor)**, Full Professor at Université de Montréal, Scientific Director of Mila - Quebec AI Institute, Canada CIFAR AI Chair
- **Alondra Nelson (Declaration’s co-sponsor)**, Harold F. Linder Professor and Science, Technology, and Social Values Lab at the Institute for Advanced Study, Member, UN High-Level Advisory Body on AI
- **Benjamin Prud’homme**, Vice President, Policy, Safety and Global Affairs at Mila - Quebec AI Institute
- **B. Ravindran**, Professor at the Indian Institute of Technology Madras and the head of the Centre for Responsible AI (CeRAI) and the Wadhwani School of Data Science and AI (WSAI) at IIT Madras
- **Yi Zeng**, Professor at Chinese Academy of Sciences, Director of Beijing Institute of AI Safety and Governance, Director of Center for Long-term AI
- **Carme Artigas,** Co-Chair, UN High-Level Advisory Body on AI and Senior Fellow Harvard Belfer Center
- **Ran Balicer**, Deputy-DG and Chief Innovation Officer at Clalit Health Services, Israel, Public Health Professor at Ben-Gurion University, Israel Society for Quality in Healthcare Chair, UN High-Level Advisory Body on AI member
- **Peter Gluckman**, President of the International Science Council and Director; Koi Tū; the Centre for Informed Futures, University of Auckland, New Zealand
- **Timo Harakka**, Member of Parliament of Finland, Vice Chair of the Committee for the Future, Member of AI Finland Advisory Board and IQM Quantum Council
- **Jaan Tallinn**, Cofounder of the Future of Life Institute
- **Jian Wang**, Founder of Alibaba Cloud, Alibaba Group
- **Mariano-Florentino (Tino) Cuellar**, President, Carnegie Endowment for International Peace
- **Amal El Fallah Seghrouchni**, Executive President of AI Movement-UM6P and UNESCO, Full Professor at Sorbonne University - Paris
- **Brian Tse**, Founder and CEO, Concordia AI
- **Jung-Woo Ha**, Head of Future AI Center at NAVER, Head of AI Innovation at NAVER Cloud, Co-Chair of Citizen’s Coalition for Scientific Society
- **Seydina Moussa Ndiaye**, Senior Lecturer and FORCE-N Program Director at Cheikh Hamidou Kane Digital University, President of Senegalese Association for Artificial Intelligence, Member, UN High-Level Advisory Body on AI
- **Dan Hendrycks**, Executive Director of the Center for AI Safety
- **Rumman Chowdhury,** CEO and co-Founder, Humane Intelligence
- **Akiko Murakami**, Board Member of The Association for Natural Language Processing in Japan
- **James Manyika**, Co-Chair, UN High-Level Advisory Body on AI, SVP of Google-Alphabet, President for Research, Technology & Society
- **Francesca Rossi**, IBM AI Ethics Global Leader, AAAI President, Co-chair of the OECD Expert Group on AI Futures, Co-chair of the GPAI Working Group on Responsible AI, Board member of the Partnership on AI