La Déclaration de Manhattan sur la compréhension scientifique mondiale et inclusive de l'intelligence artificielle

Dans le contexte de la rencontre à haut niveau, « Towards a Common Understanding of AI Capabilities, Opportunities, and Risks: Forging the Path for a Positive Future for All » (*Vers une compréhension commune des capacités, des opportunités et des risques de l'IA : ouvrir la voie à un avenir positif pour tous et toutes*), organisée au siège de l'Organisation des Nations unies (ONU) lors de la 79e session de l'Assemblée générale des Nations unies (AGNU), nous, les scientifiques et chercheurs soussignés en intelligence artificielle et politique technologique, formulons la présente déclaration.

Nous reconnaissons le besoin urgent d'une compréhension mondiale partagée des capacités, des opportunités et des risques de l'IA, tels que mis en lumière par les récents développements, dont la diffusion des recommandations de l'Organe consultatif de haut niveau du secrétaire général de l'ONU sur l'intelligence artificielle et la parution du *Rapport scientifique international sur la sécurité de l'intelligence artificielle avancée* (*intérimaire*).

Les développements en intelligence artificielle, en particulier dans les modèles de fondation, ont démontré à la fois le potentiel bénéfique et les risques significatifs dans le développement et l'utilisation de ces technologies. Alors que le développement de systèmes plus performants se profile, il est crucial de s'unir en tant que communauté scientifique mondiale pour anticiper les défis et soutenir un usage sûr et bénéfique de l'IA.

Nous déclarons le besoin de :

1. **Coopération scientifique mondiale** : Nous appelons à une collaboration internationale accrue sur la recherche en IA, en particulier sur les enjeux concernant la sécurité, l'éthique et l'impact sociétal de l'IA. Aucun pays ou organisation ne peut relever seul ces défis.

2. **Évaluation des capacités et mitigation des risques en IA** : Nous reconnaissons le besoin pressant d'évaluer les capacités et les risques associés, et de relever les défis posés par les systèmes d'IA, en particulier ceux qui sont de plus en plus performants. Nous nous engageons à prioriser une recherche en IA alignée sur les valeurs humaines, les impacts bénéfiques et la robustesse, ainsi que la mitigation des effets néfastes actuels et les risques anticipés.

3. **Favoriser le développement de l'IA en tant que bien public mondial** : Nous réaffirmons notre engagement à développer des systèmes d'IA qui sont bénéfiques à l'humanité et nous reconnaissons leur rôle central dans l'atteinte des Objectifs de développement durable comme l'amélioration de la santé et

l'éducation. Nous soulignons que tout le cycle de vie des systèmes d'IA, incluant la conception, le développement et le déploiement, doit être aligné avec des principes fondamentaux protégeant les droits de la personne, la vie privée, l'équité et la dignité de tous et de toutes.

4. **Participation inclusive** : Nous insistons sur l'importance d'inclure des perspectives diversifiées, provenant de chercheurs et de chercheuses du monde entier, indépendamment de leurs origines, de leurs emplacements géographiques ou de leurs affiliations institutionnelles. Les impacts de l'IA sont mondiaux et notre approche de son développement et de sa gouvernance doit être juste et tout aussi inclusive.

5. **Recherche transparente et évaluation des risques** : Nous nous engageons à promouvoir la science ouverte et les pratiques de recherche transparentes en IA, particulièrement pour tout travail qui a des implications significatives sur la gouvernance mondiale de l'IA et la sécurité.

6. **Approche interdisciplinaire** : Nous reconnaissons que saisir les opportunités et faire face aux défis posés par l'IA demandera la contribution de plusieurs disciplines, incluant l'informatique, la sécurité matérielle et la sécurité de l'information, l'éthique, l'économie, les sciences cognitives, les neurosciences, les sciences sociales, les études culturelles, les sciences mathématiques, et d'autres. Nous nous engageons à collaborer de manière interdisciplinaire et transdisciplinaire.

7. **Développement, déploiement et usage responsable** : Nous défendons une approche mesurée de la conception, du développement, du déploiement et de l'usage de l'IA qui priorise les usages bénéfiques (tels que les ODD), les considérations éthiques et de sécurité, et le bénéfice sociétal plutôt que l'avancement rapide à tout prix.

8. **Soutien des initiatives de gouvernance** : Nous soutenons les efforts de l'ONU et d'autres organisations nationales, régionales et internationales dans le développement de cadres de gouvernance pour l'IA basés sur des données probantes et qui, entre autres, promeuvent l'interopérabilité et minimisent la fragmentation, reconnaissant qu'une bonne gouvernance peut servir de catalyseur essentiel à l'innovation dans l'intérêt public. Nous nous engageons à fournir l'expertise scientifique pour éclairer ces initiatives.

9. **Engagement public** : Nous reconnaissons l'importance d'engager le dialogue avec les décideurs politiques et le public afin de parvenir à une compréhension commune des capacités, des bénéfices, des limitations et des impacts potentiels de l'IA, offrant ainsi à la société les moyens de faire des choix éclairés. Nous prendrons des mesures en ce sens.

10. **Perspectives à court et à long terme** : Nous nous engageons à considérer les implications à long terme et les impacts sur les générations futures du développement de l'IA, incluant les risques mondiaux majeurs potentiels et les avantages transformateurs, dans nos recherches et nos recommandations, tout en ne négligeant pas les risques et les effets néfastes actuels et à court terme.

Nous exhortons les scientifiques en IA et les chercheurs en politique technologique partout dans le monde et dans tous les secteurs à se joindre à cet engagement pour le développement d'une IA responsable et d'une collaboration internationale.

Nous invitons les décideurs politiques et les États membres à s'engager activement avec la communauté scientifique mondiale afin de faire face aux grands défis soulevés par l'IA.

En tant que scientifiques en IA et chercheurs en politique technologique, nous plaidons pour une approche véritablement inclusive et mondiale dans la compréhension des capacités, des opportunités et des risques liés à l'IA. Cela est essentiel pour façonner une gouvernance mondiale efficace des technologies de l'IA. Ensemble, nous pouvons nous assurer que le développement de systèmes avancés d'IA bénéficie à toute l'humanité.

Signataires

- **Yoshua Bengio (co-instigateur de la Déclaration)**, professeur titulaire (Université de Montréal), directeur scientifique (Mila – Institut québécois d'intelligence artificielle), titulaire d'une Chaire en IA Canada-CIFAR
- **Alondra Nelson (co-instigatrice de la Déclaration)**, professeure Harold F. Linder du laboratoire Science, Technology, and Social Values (Institute for Advanced Study), membre (Organe consultatif de haut niveau de l'ONU sur l'IA)
- **Benjamin Prud'homme**, vice-président, Politiques publiques, sécurité et affaires mondiales (Mila – Institut québécois d'intelligence artificielle)
- **B. Ravindran**, professeur (Indian Institute of Technology Madras) et chef (Centre for Responsible AI et Wadhwani School of Data Science and AI, IIT Madras)
- **Yi Zeng**, professeur (Chinese Academy of Sciences), directeur (Beijing Institute of AI Safety and Governance et Center for Long-term AI)
- **Carme Artigas,** coprésidente (Organe consultatif de haut niveau de l'ONU sur l'IA) et *Senior Fellow* (Harvard Belfer Center)
- **Ran Balicer**, directeur général adjoint et chef de l'innovation (Clalit Health Services, Israël), professeur de santé publique (Ben-Gurion University), titulaire de la chaire en soins de santé (Israel Society for Quality in Healthcare), membre de l'Organe consultatif de haut niveau de l'ONU sur l'IA
- **Peter Gluckman**, président (International Science Council) et directeur (Koi Tū; the Centre for Informed Futures, University of Auckland, Nouvelle-Zélande)
- **Timo Harakka**, membre du Parlement finlandais, vice-président (Committee for the Future), membre (AI Finland Advisory Board et IQM Quantum Council)
- **Jaan Tallinn**, cofondateur (Future of Life Institute)

- **Jian Wang**, fondateur (Alibaba Cloud, Alibaba Group)
- **Mariano-Florentino (Tino) Cuellar**, président (Carnegie Endowment for International Peace)
- **Amal El Fallah Seghrouchni**, présidente exécutive (AI Movement-UM6P), membre (UNESCO) et professeure titulaire (Sorbonne Université, Paris)
- **Brian Tse**, fondateur et PDG (Concordia AI)
- **Jung-Woo Ha**, chef (Future AI Center à NAVER), chef de l'innovation en IA (NAVER Cloud), coprésident (Citizen's Coalition for Scientific Society)
- **Seydina Moussa Ndiaye**, maître de conférence et directeur du programme FORCE-N (Cheikh Hamidou Kane Digital University), président de l'Association sénégalaise pour l'intelligence artificielle, membre (Organe consultatif de haut niveau de l'ONU sur l'IA)
- **Dan Hendrycks**, directeur exécutif (Center for AI Safety)
- **Rumman Chowdhury**, PDG et cofondateur (Humane Intelligence)
- **Akiko Murakami**, membre du conseil d'administration (Association for Natural Language Processing, Japon)
- **James Manyika**, coprésident (Organe consultatif de haut niveau de l'ONU sur l'IA), vice-président senior et président Recherche, technologie et société (Google-Alphabet)
- **Francesca Rossi**, leader mondiale (IBM AI Ethics), présidente (AAAI), coprésidente (OECD Expert Group on AI Futures), coprésidente (Groupe de travail sur l'IA responsable, Partenariat mondial sur l'intelligence artificielle), membre du conseil d'administration (Partnership on AI)