# L'IA Responsable
# Rapport du Groupe de travail

Novembre 2021 – Sommet du PMIA Paris

GPAI | THE GLOBAL PARTNERSHIP ON ARTIFICIAL INTELLIGENCE

*Le présent rapport a été élaboré par les experts du groupe de travail sur l'IA Responsable du Partenariat Mondial en Intelligence Artificielle (PMIA). Le rapport reflète les opinions personnelles des experts du PMIA et ne reflète pas nécessairement le point de vue des organisations des experts, du PMIA, de l'OCDE ou de leurs membres respectifs.*

# Mot de bienvenue des coprésidents

**Yoshua Bengio,**
Fondateur et directeur scientifique
de Mila

**Raja Chatila**
Directeur
Laboratoire SMART sur les interactions
homme-machine
Université de la Sorbonne

C'est avec grand plaisir que nous vous présentons ce rapport sur notre mandat et notre mission, qui consistent à « favoriser et contribuer au développement, à l'utilisation et à la gouvernance responsables des systèmes d'IA centrés sur l'homme, en cohérence avec les objectifs de développement durable des Nations Unies ».

Bien que le Groupe de Travail que nous avons le privilège de superviser en tant que coprésidents soit à proprement parler composé d'experts dans le domaine de l'IA, notre mandat est en fait un appel à l'action pour un partenariat plus large, regroupant d'une part, des experts en IA convaincus par notre mission, et d'autre part, des agents des gouvernements, des entreprises, de la société civile et du grand public ayant une vision similaire, ceci dans le but de comprendre et de faire progresser les technologies d'IA s'alignant avec les priorités et les valeurs de nos sociétés.

Ce partenariat, impliquant tous les acteurs de la société, est l'une des choses qui nous a le plus enthousiasmés lorsque nous avons rejoint le PMIA. Rien de moins n'est requis pour l'ensemble des défis sur lesquels nous nous sommes concentrés en 2021 et c'est le moyen d'accomplir notre mission actuelle.

Nous sommes fiers de vous présenter nos progrès sur deux défis urgents qui ont été nos priorités de travail en 2021. Pour l'un comme pour l'autre, une réponse de la société tout entière est nécessaire :

- Le premier projet est de proposer une feuille de route pragmatique sur la façon dont l'IA peut être développée, utilisée et gouvernée de manière responsable afin de participer à la lutte contre le changement climatique, qui est une priorité commune des gouvernements, comme l'atteste l'Accord de Paris de 2015.
- Le second est de chercher à mieux comprendre les relations entre les utilisateurs de réseaux sociaux et les contenus préjudiciables sur Internet en développant des méthodes pour que les parties prenantes externes (gouvernements et groupes citoyens) puissent collaborer avec les entreprises du secteur et étudier les effets des systèmes de recommandation. Ce projet a débuté en Nouvelle-Zélande en tant qu'étude de cas susceptible d'être ajustée à l'avenir. Il s'appuie sur les objectifs communs définis par les gouvernements et les entreprises ayant répondu à l'Appel de Christchurch.

Ces deux projets répondent à des priorités urgentes reconnues par les membres du PMIA. Ainsi, nous espérons créer une dynamique autour de ces premiers résultats présentés au Sommet 2021, et considérerons les deux rapports comme une base de collaboration qui permettra aux membres du PMIA de mettre en œuvre des actions pratiques en 2022.

L'ambition de notre Groupe de Travail ne se limite toutefois pas à cela. Bien que nous soyons enthousiastes à l'idée de poursuivre ces deux questions en 2022, le Groupe de Travail a également identifié un ensemble de questions plus large qui bénéficieraient du partenariat que représente le PMIA. Établir des priorités est une tâche ardue, mais nous nous efforcerons d'élargir la vision du

Groupe de Travail.

En conclusion, nous tenons à remercier tous les membres du Groupe de Travail pour leur dévouement, leur implication, leur créativité et leur travail tout au long de l'année passée. Nous avons hâte de voir ce qu'ils accompliront en 2022.

# Aperçu du groupe de travail

Ce Groupe de Travail rassemble 35 experts originaires de 20 pays différents (plus 11 observateurs) autour d'un même mandat : favoriser et contribuer au développement, à l'utilisation et à la gouvernance responsables des systèmes d'IA centrés sur l'homme, en cohérence avec les objectifs de développement durable des Nations unies.

Ce mandat est étroitement lié à la mission globale du PMIA et le Groupe de Travail est ravi d'avoir pu débuter des collaborations avec les deux autres Groupes de Travail soutenus par le Centre d'expertise de Montréal (le « CEIMIA »). Plus précisément, nous avons particulièrement apprécié l'expertise que le Groupe de Travail sur la Gouvernance des Données a pu apporter sur les éléments afférant aux projets relatifs aux données. Nous avons hâte de poursuivre cette collaboration sur les projets pilotes sur les fiducies de données en climatologie dans le cadre des travaux sur une stratégie d'IA responsable pour l'environnement. De même, nous avons beaucoup apprécié notre collaboration avec le sous-groupe consacré à l'IA et la réponse à la pandémie, qui travaille sur l'IA en matière de découverte de médicaments dans le domaine public.

# Membres du Groupe de travail du PMIA sur l'IA responsable

## Membres du Groupe de travail

**Yoshua Bengio** (coprésident) – Mila (Canada)
**Raja Chatila** (coprésident) – Université de la Sorbonne (France)
**Emile Aarts** – Tilburg University (Pays-Bas)
**Carolina Aguerre** – Center for Technology and Society (Argentine / UNESCO)
**Cesar Alberto Penz** – Federal Institute of Education, Science and Technology of Santa Catarina (Brésil)
**Genevieve Bell** – Australian National University (Australie)
**Ivan Bratko** – University of Ljubljana (Slovénie)
**Joanna Bryson** – Hertie School (Allemagne)
**Partha Pratim Chakrabarti** – Indian Institute of Technology Kharagpur (Inde)
**Jack Clark** – OpenAI (États-Unis)
**Virginia Dignum** – Umeå University (Suède / EU)
**Dyan Gibbens** – Trumbull Unmanned (États-Unis)
**Kate Hannah** – Te Pūnaha Matatini, University of Auckland (Nouvelle-Zélande)
**Toshiya Jitsuzumi** – Chuo University (Japon)
**Bogumił Kamiński** – Warsaw School of Economics (Pologne)
**Alistair Knott** – University of Otago (Nouvelle-Zélande)
**Pushmeet Kohli** – DeepMind (Royaume-Uni)
**Marta Kwiatkowska** – Oxford University (Royaume-Uni)
**Christian Lemaître Léon** – Metropolitan Autonomous University (Mexique)
**Miguel Luengo-Oroz** – UN Global Pulse (Espagne)
**Vincent C. Müller** – Technical University of Eindhoven (UE)
**Wanda Muñoz** – SEHLAC Mexico (Mexique)

**Alice H. Oh** – KAIST School of Computing (Corée du Sud)
**Luka Omladič** – Institute of Applied Ethics (Slovénie)
**Julie Owono** – Internet Sans Frontières (UNESCO)
**Dino Pedreschi** – University of Pisa (Italie)
**V K Rajah** – Advisory Council on the Ethical Use of Artificial Intelligence and Data (Singapour)
**Marley Rebuzzi Vellasco** – Tecgraf Institute of Technical-Scientific Software Development of PUC-Rio University (Brésil)
**Catherine Régis** – Université de Montréal (Canada)
**Francesca Rossi** – IBM Research (Italie)
**David Sadek** – Groupe Thales (France)
**Rajeev Sangal** – International Institute of Information Technology Hyderabad (Inde)
**Matthias Spielkamp** – Algorithm Watch (Allemagne)
**Osamu Sudo** – Chuo University (Japon)
**Joaquín Quiñonero** – Facebook (Espagne)

## Observateurs

**Ricardo Baeza-Yates** – Universitat Pompeu Fabra & Northeastern University
**Amir Banifatemi** – AI Commons
**Vilas Dhar** – The Patrick J. McGovern Foundation
**Marc-Antoine Dilhac** – ALGORA Lab
**Mehmet Haklidir** – Informatics and Information Security Research Centre
**Nicolas Miailhe** – The Future Society
**Karine Perset** – OCDE
**Golestan Radwan** – Gouvernement égyptien
**Sasha Rubel** – Section Innovation et transformation numérique, Secteur de la Communication et de l'information, UNESCO
**Stuart Russell** – UC Berkeley
**Cédric Wachholz** – Section Innovation et transformation numérique, Secteur de la Communication et de l'information, UNESCO

# Rapport d'activité

Lors du Sommet de 2020, le Groupe de Travail s'était engagé à axer ses efforts sur le développement d'environnements favorables aux technologies de l'IA en vue d'atteindre les objectifs de développement durable (ODD) des Nations Unies ainsi que d'autres objectifs clés.

Il a ainsi décidé de créer cinq comités internes :

1. **Le comité sur la découverte de médicaments et la science ouverte** (en lien avec l'ODD 3 : Bonne santé et bien-être) ;
2. **Le comité sur les changements climatiques et la préservation de la biodiversité** (ODD 13 : Lutte contre les changements climatiques) ;
3. **Le comité sur l'IA et l'éducation** (ODD 4 : Éducation de qualité) ;
4. **Le comité sur la gouvernance et la transparence des réseaux sociaux** (ODD 16 : Paix, justice et institutions efficaces) ;
5. **Un comité transversal sur les enjeux et les moyens de gouvernance** (qui pourrait travailler sur les mécanismes de certification, d'évaluation et d'audit utilisés pour évaluer les systèmes d'IA).

À la suite d'un processus de réflexion et d'engagement avec le Comité de pilotage et le Conseil du PMIA, le Groupe de Travail a choisi deux projets prioritaires en 2021, un troisième (**découverte de médicaments et science ouverte**) étant réalisé en collaboration avec le sous-groupe consacré à l'IA et la réponse à la pandémie :

1. **Une stratégie d'IA responsable pour l'environnement :** la sélection de ce projet reconnaît que le combat pour préserver la biodiversité et lutter contre le changement climatique représente un des défis les plus urgents auxquels l'humanité est confrontée. Tous les pays membres du PMIA ont fait de cet enjeu une priorité et pris des engagements forts, notamment via l'Accord de Paris signé en 2015. Ce projet vise à élaborer une stratégie mondiale pour l'adoption de l'IA responsable, dans le but de lutter contre le changement climatique et de préserver la biodiversité. Le comité de projet (dont la liste complète des membres figure à l'Annexe 1) est codirigé par Raja Chatila et Nico Miailhe. Il a travaillé en collaboration avec le *Centre for AI and Climate* et avec l'initiative *Climate Change AI*.

2. **IA responsable pour la gouvernance des réseaux sociaux :** la sélection de ce projet reflète un consensus croissant sur la nécessité pour les gouvernements de revoir l'efficacité des réglementations actuelles concernant l'influence des réseaux sociaux sur la dynamique du discours public, afin que ces processus soient entrepris de manière démocratique et systématique, plutôt que réservés aux entreprises privées. Ce projet vient répondre aux inquiétudes grandissantes sur le niveau de mauvais usage, qui peut être préjudiciable, servir la désinformation, promouvoir l'extrémisme et la violence, et favoriser de nombreuses formes de harcèlement et d'abus. Il a pour but d'identifier un ensemble de techniques et de méthodes démocratiques que les gouvernements pourraient adopter pour poser en toute sécurité un ensemble de questions convenues sur les effets des systèmes de recommandation des réseaux sociaux et pour en mesurer les effets. Le comité de ce projet du PMIA (dont la liste complète des membres figure à l'Annexe 1) est codirigé par Alistair Knott, Dino Pedreschi et Kate Hannah, en collaboration avec les universités d'Otago et d'Auckland. Il s'appuie sur l'Appel de Christchurch (un engagement des gouvernements et des entreprises du secteur de la technologie à éliminer les contenus terroristes et extrémistes violents en ligne) ; la Nouvelle-Zélande constituant la première étude de cas de ce projet.

## 1 - Une stratégie d'IA responsable pour l'environnement

En vue du Sommet de 2021, le Groupe de travail a collaboré avec l'initiative *Climate Change AI* et avec le *Centre for AI and Climate* afin de publier une feuille de route concrète destinée à guider les décideurs chargés de développer les stratégies de lutte contre le changement climatique.

L'intelligence artificielle offre des opportunités majeures pour accélérer la lutte contre le changement climatique. Ses applications permettent, par exemple, de prévoir la production d'énergie solaire, d'optimiser les systèmes de chauffage et de climatisation des bâtiments, de repérer la déforestation sur les images satellites, ou encore d'analyser les rapports financiers des entreprises afin d'extraire des informations pertinentes en matière d'environnement.[1] Toutefois, l'IA est une technologie d'application générale qui a de multiples usages possibles dans la société. Cela signifie qu'elle a déjà pu servir à entraver la lutte contre le changement climatique, que ce soit par ses effets immédiats ou par ses répercussions systémiques plus larges.[2]

Notre feuille de route donne **des recommandations concrètes sur la façon dont les gouvernements peuvent encourager l'utilisation responsable de l'IA dans le contexte de la lutte contre le changement climatique.** Ces recommandations sont le fruit d'une consultation avec un vaste ensemble de parties prenantes et peuvent être regroupées en trois catégories principales : a) encourager l'utilisation responsable de l'IA en vue d'atténuer le changement climatique et de s'y adapter, b) réduire les impacts négatifs de l'IA lorsque celle-ci est utilisée de façon contraire aux objectifs environnementaux, et c) développer des implémentations et des évaluation pertinentes ainsi que des capacités de gouvernance pour un vaste ensemble d'entités.

Ces trois catégories principales sont détaillées ci-dessous :

## 1 - a - Encourager les applications de l'IA contribuant à l'atténuation du changement climatique et à l'adaptation

Considérant que la société doit s'engager à court terme dans la lutte contre le changement climatique, il est essentiel que les solutions climatiques responsables puissent être déployées et adaptées rapidement à différents secteurs clés. Toutefois, beaucoup de ces solutions ne dépassent pas le stade de la recherche ou n'atteignent jamais le niveau de maturité technologique nécessaire. Et même après leur déploiement initial, elles continuent à rencontrer des difficultés de passage à l'échelle. Nous proposons donc que les gouvernements prennent l'initiative d'encourager l'utilisation de l'IA dans la lutte contre le changement climatique par :

- La favorisation du développement responsable et de l'accès aux **données et aux infrastructures numériques** susceptibles d'encourager la mise au point et l'adoption d'applications de l'IA pour le climat (ex. : données pertinentes, environnements de simulation, bancs de test, bibliothèques de modèles et infrastructures de calcul) ;
- Le **financement de la recherche et de l'innovation** de manière ciblée pour permettre la réalisation de travaux interdisciplinaires et intersectoriels au carrefour de l'IA et du changement climatique axés sur les impacts de ce dernier ;
- L'encouragement au **déploiement et à l'intégration des systèmes** d'applications de l'IA dédiées au climat via la définition et l'évaluation de politiques ciblées, la conception de marchés et de modèles économiques, y compris au sein des secteurs hautement réglementés tels que l'énergie, le transport, l'agriculture et l'industrie lourde.

## 1 - b - Réduire les impacts négatifs de l'IA sur le climat

**Toutes les applications de l'IA ont des répercussions sur le climat.** De ce fait, aligner l'IA avec les stratégies de lutte contre le changement climatique implique non seulement de soutenir les applications bénéfiques de l'IA, mais également de définir la place globale de l'IA pour que les applications commerciales courantes soient plus en phase avec les contraintes environnementales. L'IA augmente notamment les émissions de gaz à effet de serre, essentiellement de trois façons : a) *via* son utilisation pour des applications contribuant immédiatement à la production d'émissions de gaz à effet de serre ; b) *via* des répercussions à l'échelle du système, notamment en augmentant la demande ou par un effet de verrouillage associé aux applications d'IA ; et c) *via* l'empreinte carbone

[1] Tackling Climate Change with Machine Learning, Rolnick et al. (2019).
[2] AI and Climate Change: How they're connected, and what we can do about it, Dobbe et Whittaker (2019).

associée au cycle de vie des logiciels et du matériel informatique.[3] Les gouvernements peuvent travailler à limiter les impacts négatifs de l'IA en **prenant ses répercussions environnementales en compte dans les réglementations, stratégies, mécanismes de financement et programmes d'achat en matière d'IA**.

## 1 - c - Développer les capacités de mise en œuvre, d'évaluation et de gouvernance

À travers l'ensemble des recommandations susmentionnées transparait, de manière transversale, le besoin de développer les capacités des institutions en matière de mise en œuvre, d'évaluation et de gouvernance responsables de l'IA dans le contexte du changement climatique. Ces capacités doivent s'appuyer sur un vaste éventail d'organisations, y compris des entités gouvernementales internationales, nationales et locales, ainsi que sur des organismes privés et de la société civile dans les secteurs liés à l'environnement (par exemple l'énergie, le transport, l'industrie lourde ou l'agriculture). Nous proposons que les gouvernements prennent les mesures suivantes pour encourager le développement des capacités des institutions pertinentes en :

- Intégrant des **principes d'IA responsable** dans la conception des initiatives et des structures de gouvernance (par exemple les principes recommandés dans cette feuille de route), ce qui implique de favoriser l'inclusion de participants issus de la société civile, des gouvernements locaux, des pays de l'hémisphère sud et de groupes marginalisés ;
- Favorisant **l'évaluation de l'impact de l'IA sur le climat** en recueillant des données sur les répercussions des émissions liées à l'IA et en définissant des cadres normatifs pour les mesures et l'établissement des rapports.
- Développant les **capacités de mise en œuvre, d'évaluation et de gouvernance** grâce à la littérature, aux compétences, aux talents, aux normes, aux outils et aux meilleures pratiques.

### Une feuille de route axée sur l'action

Alors que l'utilisation de l'IA se répand très rapidement dans la société, il devient impératif que les gouvernements se montrent proactifs et contribuent à façonner ces développements en tenant compte de la lutte contre le réchauffement climatique. Outre la participation de la société civile, des secteurs académique et privé, les mesures significatives en lien avec ces initiatives requerront, au sein de chaque pays, une **collaboration entre plusieurs branches ou secteurs du gouvernement** (par exemple les agences spécialisées dans l'IA et le numérique, celles spécialisées dans la lutte contre le changement climatique ou les secteurs liés à l'environnement, les organismes de normalisation et de régulation, les gouvernements locaux). **Des collaborations multilatérales ou internationales** (ex. : création de consortiums polyvalents ou renforcement des capacités des organisations internationales existantes) peuvent également éviter les doublons inutiles, faciliter le partage des connaissances et renforcer l'ensemble des efforts. Nous espérons que les recommandations et la liste des goulets d'étranglement existants, qui figurent dans la feuille de route, serviront de tremplin à ces initiatives.

Nous profitons de cette occasion pour remercier l'initiative *Climate Change AI* et le *Centre for AI and Climate* pour leur excellent travail et leur participation à l'élaboration de la feuille de route. Nous saluons le dévouement et le talent de leur équipe : David Rolnick, Priya Donti, Lynn Kaack et Peter Clutton-Brock. Nous avons hâte de poursuivre sur cette lancée en 2022.

[3] Artificial Intelligence and Climate Change: Opportunities, considerations, and policy levers to align AI with climate change goals, Kaack et al. (2020).

## 2 - Une IA responsable en gouvernance des médias sociaux

Le rapport du comité du projet, publié à l'occasion du sommet de 2021, présente ses conclusions et recommandations.

Le projet s'intéresse à deux questions connexes :

1. Comment définir le concept de « contenu préjudiciable » sur les réseaux sociaux. Nous passons en revue les définitions existantes de « contenus préjudiciables » qu'utilisent les entreprises du secteur et, plus largement, les communautés universitaire et politique. Mais, nous avançons également une proposition particulière : les communautés d'un pays donné devraient pouvoir s'emparer de la définition de « contenus préjudiciables », avec une attention accrue pour les communautés les plus durement impactées. Sur les réseaux sociaux, les discours haineux ciblent certaines communautés en particulier : dans le cadre de notre projet, nous donc avons testé une méthode pour inviter ces groupes à partager leurs expériences afin de faire émerger des définitions véritablement significatives pour le terme de « contenu préjudiciable ». Ces travaux se concentrent sur un seul pays (la Nouvelle-Zélande) en tant qu'étude de cas, mais la méthode que nous testons actuellement est conçue pour s'adapter aux autres pays. L'un de nos objectifs est de souligner l'importance des variations régionales dans les définitions de « contenus préjudiciables » et de proposer un possible modèle de gouvernance régionale des plateformes Internet à cet égard.

2. La seconde question porte sur les systèmes d'IA qui diffusent des contenus sur les réseaux sociaux, c'est-à-dire les algorithmes de recommandation. Ces algorithmes obtiennent des informations sur les utilisateurs individuels à partir de leurs actions sur ces plateformes et se servent de ces connaissances pour diffuser des contenus personnalisés dans leurs flux. L'algorithme de recommandation choisit des éléments pour un utilisateur donné en fonction des informations dont il dispose sur son comportement sur la plateforme. Toutefois, ces choix influencent directement le comportement de l'utilisateur, qui découle largement des éléments apparaissant dans son flux. Les informations qu'un système de recommandation obtient sur l'utilisateur dépendent donc en partie de son apprentissage initial. Les théoriciens de l'IA l'ont démontré : il est possible que cette interdépendance du système de recommandation pousse les utilisateurs vers des niches de contenus arbitraires, c'est ce qu'on appelle l'effet « bulle de filtres ». Nous passons en revue les modèles théoriques qui ont mis en évidence cet effet. Nous examinons également les preuves selon lesquelles les contenus préjudiciables sur les réseaux sociaux ont des répercussions dans le monde, ainsi que celles portant sur les biais cognitifs qui poussent les utilisateurs vers divers types de contenus préjudiciables. Considérées dans leur globalité, ces études indiquent que le premier sujet d'inquiétude, selon lequel les algorithmes de recommandation peuvent pousser les utilisateurs des réseaux sociaux vers des contenus préjudiciables, est légitime. C'est l'objet de notre second projet.

Le débat scientifique est vif autour des effets des algorithmes de recommandation sur les utilisateurs des plateformes et sur la manière de mesurer ces effets. Là encore, nous passons en revue la littérature existante. À ce jour, la quasi-totalité des études ont été menées en dehors des plateformes de réseaux sociaux, à l'aide de données accessibles publiquement, obtenues soit grâce à des expériences sur les utilisateurs et les interfaces des réseaux sociaux, soit *via* des API fournies par les entreprises pour mettre en évidence certains aspects de leur fonctionnement. Les constats de ces études sont très mitigés : certaines études concluent que les systèmes de recommandation ont des effets préjudiciables importants, d'autres qu'ils n'en ont aucun, et d'autres encore qu'ils ont des effets mineurs. Nous estimons que cette diversité des résultats s'explique en grande partie par le fait que les techniques disponibles pour étudier les systèmes de recommandation en dehors des plateformes sont imparfaites et souffrent d'une série de problèmes méthodologiques. Plus particulièrement, aucune des méthodes existantes ne permet de tester correctement les hypothèses causales concernant les effets des systèmes de recommandation sur les utilisateurs. Ces méthodes externes ne fournissent tout simplement pas d'informations suffisantes sur les effets des systèmes de recommandation sur les utilisateurs. En particulier, les gouvernements qui envisagent des options de réglementation pour les plateformes de réseaux sociaux ne sont pas suffisamment informés pour le moment et doivent d'abord en apprendre plus.

Pour tester une hypothèse causale portant sur les effets d'un système de recommandation donné, il est nécessaire de réaliser des expériences qui manipulent le système, en testant différentes versions sur divers groupes d'utilisateurs et en recherchant les différences de comportement sur ces groupes. Élément important : il s'agit de la méthode employée par les entreprises du secteur pour développer et optimiser leurs propres systèmes. Dans leurs études, ces entreprises cherchent en premier lieu à mesurer l'engagement des utilisateurs. Sur la base de ce second projet, nous recommandons donc aux gouvernements de collaborer avec les entreprises du secteur des réseaux sociaux pour réaliser des études employant leurs méthodes et examiner ainsi les effets des systèmes de recommandation sur la relation entre les utilisateurs avec les contenus préjudiciables. Cela leur permettra d'accéder à une bien meilleure compréhension de ces effets et d'obtenir des informations essentielles pour développer leurs politiques en toute connaissance de cause. En outre, cela constituera également une nouvelle mesure de transparence pour les réseaux sociaux. En effet, il est important de souligner que la transparence est davantage liée aux effets des algorithmes plutôt qu'à leur conception interne ou aux données qu'ils exploitent : la propriété intellectuelle de l'entreprise, de même que les données à caractère personnel des utilisateurs de la plateforme, sont protégées.

Encore une fois, notre projet se concentre sur la Nouvelle-Zélande, le pays faisant l'objet de l'étude de cas. Dans le cadre de ce projet du PMIA, le gouvernement néozélandais a invité une entreprise du secteur des réseaux sociaux à collaborer avec nous afin d'intégrer des indicateurs relatifs aux attitudes des utilisateurs néozélandais vis-à-vis des contenus préjudiciables dans les méthodes d'optimisation de ses algorithmes de recommandation. Notre étude de cas est focalisée sur la Nouvelle-Zélande, toutefois, l'exercice proposé pourrait être mis en place par n'importe quel gouvernement pour tester les effets des algorithmes de recommandation de n'importe quelle entreprise. Notons par ailleurs que l'exercice proposé n'affecte en rien l'expérience des utilisateurs : il vise simplement à obtenir de nouvelles informations à partir des méthodes déjà utilisées par les entreprises pour tester leurs algorithmes de recommandation, à éclairer l'élaboration des politiques, et à fournir un moyen de mesurer la transparence. En outre, cet exercice permet de mesurer la gouvernance régionale des plateformes de réseaux sociaux.

L'exercice que nous proposons consiste à collaborer avec une entreprise du secteur des réseaux sociaux afin d'étudier les effets de son système de recommandation sur les attitudes des utilisateurs envers les contenus préjudiciables. La définition de « contenus préjudiciables » est donc une fois de plus mise en cause. À des fins pratiques, nous proposons de nous concentrer sur la catégorie « contenus terroristes et extrémistes violents » (TVEC pour *Terrorist and Violent Extremist Content*), qui fait déjà l'objet de collaborations productives entre les entreprises du secteur des technologies et les gouvernements du monde entier. Par l'intermédiaire du Forum mondial de l'Internet contre le terrorisme (GIFCT pour *Global Internet Forum to Counter Terrorism*), les entreprises collaborent pour créer et utiliser une base de données commune pour les contenus TVEC. Cette collaboration bénéficie du soutien des entreprises et des pays (y compris tous les pays membres du PMIA) qui ont répondu à l'Appel de Christchurch visant à éliminer les contenus TVEC en ligne. Coïncidence, l'un des thèmes de l'axe de travail de l'Appel de Christchurch pour cette année est d'explorer « le parcours de l'utilisateur [vers les contenus TVEC] et le rôle que cela joue dans le processus plus large de radicalisation ». Les participants à l'Appel, parmi lesquels figurent toutes les plus grandes entreprises de technologie, se sont déjà engagés à « concevoir un processus multipartite visant à définir quelles méthodes peuvent être utilisées en toute sécurité et quelles informations sont nécessaires (dans le respect du secret industriel) pour permettre aux parties prenantes de mieux comprendre les résultats des processus algorithmiques et leur potentiel d'amplification de la TVEC ». Les premiers résultats sont attendus entre novembre 2021 et mai 2022. Cet engagement en faveur de la collaboration et ce calendrier fournissent un contexte idéal pour l'exercice que nous envisageons.

L'exercice que nous proposons pose évidemment de nombreuses questions politiques et juridiques. La troisième partie de notre projet, dirigée par des avocats spécialisés en IA et dans la gouvernance des réseaux sociaux (Tom Barraclough et Curtis Barnes), porte sur ces questions.

# Regarder vers l'Avenir

Il a été convenu avec le comité de pilotage du PMIA qu'en 2022, le Groupe de Travail poursuivra ses deux projets actuels afin de créer une dynamique et d'atteindre son potentiel pour produire un effet concret et significatif sur les membres du PMIA.

Comme nous l'avons indiqué, le Groupe de Travail a également des intérêts plus larges représentés par les comités formés à la suite du sommet de 2020. Tout en continuant à mobiliser autour de ces deux projets, le Groupe de Travail entend développer une vision plus large de son mandat en 2022.

Le projet pour une **stratégie d'IA responsable pour l'environnement** tâchera d'accélérer la mise en œuvre de la feuille de route par les décideurs politiques, les investisseurs et la communauté des développeurs en :

- *Élargissant le champ d'application* de la feuille de route et du catalogue de cas d'utilisation à la **préservation de la biodiversité** ;
- *Développant* **le catalogue de cas d'utilisation pour en faire un référentiel vivant**, en collaboration avec le monde universitaire, les organisations internationales concernées, le secteur privé et les acteurs de la société civile, dans le but de mieux comprendre le potentiel de l'IA responsable dans la lutte contre le changement climatique ;
- *Impliquant la communauté* par **l'inscription de la feuille de route dans l'agenda des principales organisations intergouvernementales** (PNUE, GIEC, COP), des états membres du PMIA, et des communautés des pays du sud (**en développant si nécessaire un programme d'engagement adapté aux besoins de chacun**) ; cela inclus l'organisation d'une série limitée d'ateliers de consultation pour promouvoir la feuille de route lors d'évènements clés (notamment la COP 27 sur le changement climatique en novembre 2022 et la COP 16 sur la biodiversité en avril 2022) ;
- *Guidant* la communauté par le développement des éléments suivants :
  - **des plans stratégiques de mise en œuvre** de la feuille de route identifiant les opportunités les plus pertinentes et efficaces pour la coopération internationale en matière de recherche et développement, de déploiement et de passage à l'échelle ;
  - **des cadres/instruments pour la planification et l'évaluation des impacts**, tels que l'indice *Global Climate & AI* ou des **critères de référence techniques** permettant d'établir une base comparative internationale pour mesurer la performance par rapport à la feuille de route. Ces outils pourraient guider les communautés d'investisseurs, de décideurs et de développeurs vers les opportunités les plus prometteuses en termes d'impact ;
- *Dirigeant la communauté* en **pilotant de nouveaux mécanismes**, à commencer par la collaboration sur les **fiducies de données climatologiques** avec le Groupe de Travail sur la gouvernance des données. Le comité cherchera à créer des partenariats avec des organisations (comme par exemple des fondations) intéressées par les défis identifiés grâce aux critères de référence techniques de l'indice *Global Climate & AI*.

**Pour le Sommet de 2022, le Comité produira une Feuille de route 2022** offrant une perspective plus précise grâce à cette activité.

S'agissant de l'**IA responsable en gouvernance des médias sociaux**, le Comité a proposé deux dimensions : 1) poursuivre l'étude de cas néo-zélandaise en élargissant les méthodes employées, et 2) appliquer à d'autres pays, les méthodes développées en 2021 pour la Nouvelle-Zélande.

La poursuite de l'étude de cas néo-zélandaise, *réalisera* l'exercice d'enquête décrit dans le projet actuel et *étendra* les méthodes de consultation de la communauté développées dans ce cadre pour les doter d'un rôle plus fonctionnel au sein des entreprises du secteur des réseaux sociaux. Nous souhaitons atteindre cet objectif grâce à une collaboration entre ce secteur et le gouvernement, dans l'esprit de la mission multipartite du PMIA.

L’application des méthodes développées pour l’étude de cas de la Nouvelle-Zélande devrait ensuite servir de modèle pour les autres gouvernements en leur montrant comment entamer le dialogue avec les entreprises du secteur des réseaux sociaux afin de s'enquérir des effets de leurs systèmes de recommandation sur les citoyens de ce pays. L’extension de ce projet à d’autres pays fera l’objet d’une consultation avec les pays membres du PMIA.

# Annexe 1

## Comité sur les changements climatiques

### Coprésidents

**Raja Chatila** – Université de la Sorbonne
**Nicolas Mailhe** – The Future Society

### Membres

**Yoshua Bengio** – Mila, Institut québécois d'intelligence artificielle
**Marta Kwiatkowska** – Oxford University
**Christian Lemaître Léon** – Metropolitan Autonomous University
**Virginia Dignum** – Université d'Umeå
**David Sadek** (Observer) – Groupe Thales
**Karine Perset** (Observer) – OCDE

### Experts invités

**Przemyslaw Biecek** – Warsaw University of Technology
**Alan Paic** – OCDE
**Cyrus Hodes** – AI Initiative
**Bertrand Monthubert** – Occitanie Data
**Allan Feitosa** – Eldorado Research Institute
**Andrew Zolli** – Planet Labs
**Claire Melamed** – Global Partnership for Sustainable Development Data
**Claire Monteleoni** – University of Colorado Boulder
**David Jensen** – UN Environment
**Bistra Dilkina** – University of Southern California
**Florence Rabier** – European Centre for Medium-Range Weather Forecasts (ECMWF)
**Florian Pappenberger** – European Centre for Medium-Range Weather Forecasts (ECMWF)
**Carla P. Gomes** – Cornell University
**Iarla Kilbane-Dawe** – Office for Artificial Intelligence, gouvernement britannique
**Jay Ashton-Butler** – Office for Artificial Intelligence, gouvernement britannique
**Janez Potočnik** – SYSTEMIQ
**Neil David Lawrence** – University of Cambridge
**Aglaé Jézéquel** – Laboratoire de Météorologie Dynamique - Institut Pierre-Simon Laplace (IPSL)
**André Loesekrug-Pietri** – Joint European Disruptive Initiative (J.E.D.I)
**Eric Badiqué** – Commission européenne

## Comité sur la gouvernance et la transparence des médias sociaux

### Coprésidents

**Kate Hannah** – University of Auckland
**Alistair Knott** – University of Otago
**Dino Pedreschi** – University of Pisa

### Membres

**Yoshua Bengio** – Mila, Institut québécois d'intelligence artificielle
**Raja Chatila** – Université de la Sorbonne

**Carolina Aguerre** – Center for Technology and Society (CETyS)
**Ivan Bratko** – University of Ljubljana
**Joanna Bryson** – Hertie School
**Dyan Gibbens** – Trumbull Unmanned
**Toshiya Jitsuzumi** – Chuo University
**Marta Kwiatkowska** – Oxford University
**Osamu Sudo** – Chuo University
**Przemyslaw Biecek** – Warsaw University of Technology
**Jack Clark** – Anthropic
**Christian Lemaître Léon** – Metropolitan Autonomous University
**Amir Banifatemi** – AI Commons

## Observateurs

**Marc-Antoine Dilhac** – ALGORA Lab
**Alan Paic** – OCDE
**Karine Perset** – OCDE
**Stuart Russell** – University of California, Berkeley

## Experts invités

**Nicolas Miailhe** – The Future Society
**Yeong Zee Kin** – Infocomm Media Development Authority
**Sebastian Hallensleben** – VDE (Verband der Elektrotechnik Elektronik Informationstechnik e.V.)
**Matija Damjan** – University of Ljubljana
**Anderson Soares** – Federal University of Goias
**Jaco Du Toit** – UNESCO
**Alejandro Pisanty Baruch** – National Autonomous University
**Colin Gavaghan** – University of Otago
**David Eyers** – University of Otago
**Andrew Trotman** - University of Otago
**Tapabrata Chakraborti** – University of Oxford
**Sanjana Hattotuwa** – University of Auckland
**Curtis Barnes** – Brainbox Institute, Auckland
**Tom Barraclough** – Brainbox Institute, Auckland