Responsible AI
Working Group Report

November 2021 - GPAI Paris Summit

GPAI / THE GLOBAL PARTNERSHIP
ON ARTIFICIAL INTELLIGENCE

Co-Chairs' Welcome

Yoshua Bengio,
Founder and Scientific Director
Mila

Raja Chatila
Director
SMART Laboratory of Human-Machine
Interaction
Sorbonne University

We are delighted to report on our mandate and mission to "foster and contribute to the responsible development, use and governance of human-centred AI systems, in congruence with the UN Sustainable Development Goals".

Although the Working Group itself is made up of experts in the field of AI that we are privileged to lead as Co-Chairs, the mandate is really a *call to action* for a wider partnership that brings together AI experts that believe in this mission with like-minded governments, industry, civil society, and the wider public at large to understand and advance AI technologies that align with the priorities and values of our societies.

That 'whole of society' partnership is one of the things that excited us most about first joining GPAI, and nothing less is required for the set of challenges we have focused on in 2021, and our ongoing mission.

We are proud to present our progress on two pressing challenges that have been the primary focus of our work in 2021. Both demonstrate the necessity of a 'whole of society' response:

- The first is an actionable roadmap on how AI can be responsibly developed, used and governed to take on the fight against climate change - a common priority amongst governments as recognised by the 2015 Paris Agreement
- The second focuses on developing understanding of social media users' relationships with harmful online content, by aiming to develop methods allowing external stakeholders (governments and citizens' groups) to work collaboratively with social media companies, and study the effects of recommender systems. The project has been initiated with New Zealand as a case study that can be scaled in future, and builds upon the shared objectives agreed by governments and industry in the Christchurch Call.

Both projects address urgent priorities recognised by GPAI members: we therefore hope to build momentum behind these first outputs presented at Summit 2021, and consider both reports as a platform for collaboration on practical action amongst GPAI's membership in 2022.

Our ambition as a Working Group does not stop here either. Whilst we are excited to pursue these two issues in 2022, the Working Group did identify a broader set of issues that would benefit from the partnership that GPAI represents. Whilst prioritisation is hard, we will work on setting out the broader vision of the Working Group in 2022.

In closing, we would like to thank all the Working Group for their dedication, commitment, creativity and hard work over the past year. We look forward to what the Working Group will achieve in 2022.

Yoshua Bengio   Raja Chatila

Working Group Overview

The Working Group brings together 35 experts from 20 countries (plus 11 Observers) around a shared mandate: to foster and contribute to the responsible development, use and governance of human-centred AI systems, in congruence with the UN Sustainable Development Goals.

This mandate relates closely to GPAI's overall mission, and the Working Group has been pleased to initiate collaborations with the two other Working Groups supported by the Montréal Centre of Expertise (the "CEIMIA). More specifically, the Working Group has appreciated the expertise offered by members of the Data Governance Working Group on the elements of its projects that relate to data, and is looking forward to further collaboration on climate focused data trust pilots in alignment with the work of the Responsible AI Strategy for the Environment. It has also appreciated the collaboration with its Subgroup, AI and Pandemic Response, on AI for Public Domain Drug Discovery.

Membership of GPAI's Data Governance Working Group

Working Group members

Yoshua Bengio (Co-Chair) – Mila (Canada)

Raja Chatila (Co-Chair) – Sorbonne University (France)

Emile Aarts – Tilburg University (Netherlands)

Carolina Aguerre – Center for Technology and Society (Argentina / UNESCO)

Cesar Alberto Penz – Federal Institute of Education, Science and Technology of Santa Catarina (Brazil)

Genevieve Bell – Australian National University (Australia)

Ivan Bratko – University of Ljubljana (Slovenia)

Joanna Bryson – Hertie School (Germany)

Partha Pratim Chakrabarti – Indian Institute of Technology Kharagpur (India)

Jack Clark – OpenAI (USA)

Virginia Dignum – Umeå University (Sweden / EU)

Dyan Gibbens – Trumbull Unmanned (USA)

Kate Hannah – Te Pūnaha Matatini, University of Auckland (New Zealand)

Toshiya Jitsuzumi – Chuo University (Japan)

Bogumił Kamiński – Warsaw School of Economics (Poland)

Alistair Knott – University of Otago (New Zealand)

Pushmeet Kohli – DeepMind (UK)

Marta Kwiatkowska – Oxford University (EU)

Christian Lemaître Léon – Metropolitan Autonomous University (Mexico)

Miguel Luengo-Oroz – UN Global Pulse (Spain)

Vincent C. Müller – Technical University of Eindhoven (EU)

Wanda Muñoz – SEHLAC Mexico (Mexico)

Alice H. Oh – KAIST School of Computing (South Korea)

Luka Omladič – Institute of Applied Ethics (Slovenia)

Julie Owono – Internet Sans Frontières (UNESCO)

Dino Pedreschi – University of Pisa (Italy)

V K Rajah – Advisory Council on the Ethical Use of Artificial Intelligence and Data (Singapore)

Marley Rebuzzi Vellasco – Tecgraf Institute of Technical-Scientific Software Development of PUC-Rio University (Brazil)

Catherine Régis – Université de Montréal (Canada)

Francesca Rossi – IBM Research (Italy)

David Sadek – Thales Group (France)

Rajeev Sangal – International Institute of Information Technology Hyderabad (India)
Matthias Spielkamp – Algorithm Watch (Germany)
Osamu Sudo – Chuo University (Japan)
Joaquín Quiñonero – Facebook (Spain)

Observers

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 – OECD

Golestan Radwan – Government of Egypt

Sasha Rubel – Digital Innovation and Transformation Section, Communication and Information Sector, UNESCO

Stuart Russell – UC Berkeley

Cédric Wachholz – Digital Innovation and Transformation Section, Communication and Information Sector at the UNESCO

Progress Report

At Summit 2020, the Working Group committed to focus on developing enabling environments for AI technologies to achieve the UN Sustainable Development Goals and other key objectives.

It decided to create five internal committees:
1. **The Committee on Drug Discovery & Open Science** (linked to SDG 3: Good health and wellbeing)
2. **The Committee on Climate Change** and Biodiversity Preservation (SDG 13: Climate action)
3. **The Committee on AI & Education** (SDG 4: Quality education)
4. **The Committee on Governance & Transparency of Social Media** (SDG 16: Peace, Justice, and Strong Institutions)
5. **A Transversal Committee on Issues and Means of Governance** (that could work on the certification, assessment and audit mechanisms used to evaluate AI systems)

Following a process of ideation and engagement with GPAI's Steering Committee and Council, the Working Group prioritised two projects in 2021, with a third **(Drug Discovery and Open Science)** being taken forward in collaboration with the AI and Pandemic Response Subgroup:

1. **A Responsible AI Strategy for the Environment:** the selection of this project recognises that the combined fight against climate change and preservation of biodiversity represents one of the most pressing challenges humanity is facing. All GPAI Member countries have put this at the top of their agenda and have taken strong commitments, especially through the Paris Agreement signed in 2015. As a response, this project aims to develop a global responsible AI adoption strategy for climate action and biodiversity preservation. The project committee (fully listed under Annex 1) is co-led by Raja Chatila and Nico Miailhe, and has collaborated with the Centre for AI and Climate Change and Climate Change AI.

2. **Responsible AI for Social Media Governance:** The selection of this project reflects a growing consensus that governments should review the effectiveness of current regulations on the influence of social media platforms on the dynamics of public discourse, so these processes are undertaken democratically and systematically, rather than solely by private companies. It

responds to growing concerns about the level of misuse which can be harmful and serve to propagate disinformation, extremism, violence and many forms of harassment and abuse. The aim of the project is therefore to identify a set of technical and democratic methods that governments could adopt to safely ask a set of agreed questions and measurements about the effects of social media recommender systems. The GPAI project committee (fully listed under Annex 1) is co-led by Alistair Knott, Dino Pedreschi, and Kate Hannah, in collaboration with the Universities of Otago and Auckland. It builds upon the Christchurch Call (a commitment by Governments and tech companies to eliminate terrorist and violent extremist content online), with New Zealand as the first case study for the project.

1 - A Responsible AI Strategy for the Environment

For Summit 2021, the Working Group collaborated with Climate Change AI and the Centre for AI and Climate to publish an action-oriented roadmap to guide policymakers developing climate action strategies.

Artificial intelligence presents significant opportunities to accelerate climate action via applications such as forecasting solar power production, optimizing building heating and cooling systems, pinpointing deforestation from satellite imagery, and analyzing corporate financial disclosures for climate-relevant information.[1] At the same time, AI is a general-purpose technology with many applications across society, which means it has also been applied in ways that impede climate action both through immediate effects and broader systemic effects.[2]

In the roadmap, we provide **actionable recommendations as to how governments can support the responsible use of AI in the context of climate change.** These recommendations were obtained via consultation with a broad set of stakeholders, and span three primary categories: (a) supporting the responsible use of AI for climate change mitigation and adaptation, (b) reducing the negative impacts of AI where it may be used in ways that are incompatible with climate goals, and (c) building relevant implementation, evaluation, and governance capabilities for and among a wide range of entities.

These three primary categories are summarised as below:

1 - a - Supporting AI applications in climate change mitigation and adaptation

Given the short time scales on which society must address climate change, it will be critical that responsible climate solutions be rapidly deployed and scaled across key sectors. However, many such solutions often get stuck in research or early stages of technological readiness, and even after initial deployment, often face difficulties scaling. We propose that governments can take leadership in supporting the use of AI to address climate change by:

- Fostering the responsible development of and access to **data and digital infrastructure** — e.g., relevant data, simulation environments, testbeds, model libraries, and computational hardware — that can support the development and adoption of AI-for-climate applications.

- Targeting **research and innovation funding** to enable interdisciplinary and cross-sectoral work at the intersection of AI and climate change that is guided by climate impact.

- Supporting **deployment and systems integration** of AI-for-climate applications via targeted policy design and evaluation, market design, and business models, including within highly-regulated sectors such as energy, transportation, agriculture, and heavy industry.

---

[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 and Whittaker (2019).

1 - b - Reducing AI's negative impacts on the climate

Every application of AI affects the climate, which means aligning AI with climate change strategies involves not only facilitating beneficial applications of AI, but also shaping the space of AI overall so that business-as-usual applications are more climate-aligned. Notably, there are three principal ways in which AI can increase greenhouse gas emissions: (a) via its use for applications with immediate negative impacts on emissions, (b) via system-level impacts such as induced demand or lock-in effects associated with AI applications, and (c) via the carbon footprint associated with the life-cycle impacts of the associated software and hardware.[3] Governments can work to reduce the negative impacts of AI by **incorporating climate impact considerations into AI regulation, strategies, funding mechanisms, and procurement programs.**

1 - c - Building implementation, evaluation, and governance capabilities

Cutting across the previous recommendations is the need to build institutional capabilities aimed at the responsible implementation, evaluation, and governance of AI in the context of climate change. Such capabilities must be built across a wide range of organizations, including governmental entities at the international, national, and local levels, as well as private and civil society organizations in climate-relevant sectors (e.g., energy, transport, heavy industry, or agriculture). We propose that governments can support the development of relevant institutional capabilities by:

- Embedding **responsible AI principles** into the design of initiatives and governance structures (e.g., those recommended in this roadmap), which includes fostering inclusion of participants from civil society, local governments, the Global South, and marginalized groups.
- Fostering **climate-cognizant impact assessment of AI** via collection of data on AI's emissions impacts, and by establishing standard measurement and reporting frameworks.
- Building **capacity for implementation, evaluation, and governance** in the form of literacy, skills and talent, standards, tools, and best practices.

A roadmap for action

As the use of AI grows rapidly across society, it is imperative that governments be proactive in helping shape these developments with climate action in mind. Within individual countries, meaningful action on these initiatives will require **collaborations among multiple branches or arms of government** — e.g., agencies focused on AI or digitalization, agencies focused on climate change or climate-relevant sectors, standards bodies, regulators, and local governments — in addition to participation from civil society, academia, and the private sector. **Multilateral or international collaborations** — e.g., via the development of cross-functional consortia or capacity building within existing international organizations — can also prevent unnecessary duplication, facilitate knowledge sharing, and strengthen overall efforts. We hope the recommendations and enumeration of existing bottlenecks contained in the roadmap will provide a launching point for these initiatives.

We would like to take this opportunity to thank the Climate Change AI and Centre for AI and Climate on their excellent work and collaboration in developing the Roadmap, and would like to commend the dedication and talent of their team: David Rolnick, Priya Donti, Lynn Kaack, and Peter Clutton-Brock. We look forward to building on this milestone in 2022.

2 - Responsible AI for Social Media Governance

The project committee's report, published for Summit 2021, presents their emerging findings and recommendations.

---

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

The project considers two related questions:

1. How to define the concept of 'harmful' social media content. We review existing definitions of harmful content that are used within social media companies and in wider academic and policy communities. But we also advance a particular proposal: that communities in a given country should take the lead in formulating definitions of harmful content—and that the loudest voice in these discussions should be given to the communities that suffer the most harm. Hate speech on social media is particularly directed at certain communities: in our project we have trialled a method for engaging with these groups about their lived experiences, to surface meaningful definitions of harmful content. This work focuses on a single country (New Zealand) as a case study, but the method we are trialling is designed to be extendable to other countries. One of our aims is to highlight the importance of regional variations in definitions of harmful content, and to propose a possible model of regional governance for Internet platforms in relation to harmful content.

2. The second question focuses on the AI systems that disseminate content on social media platforms, namely recommender algorithms. These algorithms learn about individual platform users, from their actions on the platform, and exploit this learning to deliver personalised content to user feeds. A recommender algorithm's choice of items for a given user is influenced by what it knows about this user's behaviour on the platform. But it also directly influences users' behaviour, which is in large part driven by the items arriving in their feeds. What a recommender system learns about a user thus depends in part on its earlier learning. As AI theorists have shown, this dependence makes it possible for recommender systems to push users towards arbitrary pockets of Internet content: the so-called 'filter bubble' effect. We review the theoretical models that demonstrate this effect. We also review evidence that harmful content on social media has harmful effects in the world, and evidence for a range of cognitive biases that push social media users towards harmful content of several kinds. Taken together, these studies establish a prima facie cause for concern that recommender algorithms may lead social media users towards harmful content. Our second project investigates this particular concern.

There is an active scientific debate about what effects recommender algorithms have on platform users, and on how to measure these effects. Again we review the existing literature on these effects. Nearly all studies to date have been conducted externally to social media platforms, using publicly available data, obtained either through experiments on social media users and interfaces, or through APIs provided by companies to surface certain aspects of their operation. The findings of these studies are very mixed: some studies find significant harmful effects of recommender systems, some find none, and some find only small effects. We argue an important reason for this diversity is that the techniques available for studying recommender systems externally to platforms are all flawed, suffering from a range of methodological problems. In particular, none of the existing methods test properly causal hypotheses about the effects of recommender systems on users. These external methods simply don't provide good enough information about the effects of recommender systems on users. In particular, governments considering regulatory options for social media platforms don't yet know enough—and need to know more.

To test a causal hypothesis about the effects of a given recommender system, it is necessary to conduct experiments that manipulate the system, trying out different versions on different groups of users, and looking for differences in the behaviours of different groups. Crucially, this is the method social media companies use themselves, to develop and optimise their own systems. Companies are primarily interested in measuring user engagement in their studies. The recommendation of our second project is that governments should engage with social media companies to conduct studies using these same methods to examine the effects of recommender systems on users' relationship with harmful content. This will enable them to gain a much better understanding of these effects, and provide vital information to inform subsequent policy development. It will also provide a useful new measure of transparency for social media companies. Importantly, the transparency relates to effects of companies' algorithms, rather than to their internal design, or the data they run on: company IP and the personal data of platform users are protected.

Again, our project focuses on New Zealand as a case study country. As part of the GPAI project, the New Zealand government has invited one social media company to work with us, to incorporate metrics measuring users' attitudes towards harmful content within their existing methods for optimising recommender algorithms for New Zealand users. Again, while our case study focuses on New Zealand, the proposed exercise is one that could be initiated by any government, to test the effects of any company's recommender algorithm. Note that the proposed exercise won't have any impact on user experiences: its aim is simply to learn something new from the methods companies already use for trialling their recommender algorithms, to inform policy development, and to provide a measure of transparency. Note again that the exercise provides a measure of regional governance of social media platforms.

Our proposed exercise involves collaborating with a social media company to study the effect of its recommender system on users' attitudes towards harmful content. The definition of 'harmful content' is once again at issue. Our practical proposal is to focus on the category of 'Terrorist and Violent Extremist Content' (TVEC), that is already the focus of productive collaboration between tech companies and governments around the world. Through the Global Internet Forum to Counter Terrorism (GIFCT), companies are collaborating in the creation and use of a shared database of TVEC material—a collaboration supported by companies and countries (including all the GPAI countries) participating in the Christchurch Call to eliminate TVEC online. Coincidentally, one of the topics for this year's Christchurch Call workstream is to explore the 'user journey [towards TVEC], and the role this may play in the broader radicalisation process'. Call participants, including all the major tech companies, have already committed to 'design a multi-stakeholder process to establish what methods can safely be used and what information is needed—without compromising trade secrets—to allow stakeholders to better understand the outcomes of algorithmic processes and their potential to amplify TVEC', surfacing results between November 2021 and May 2022. This commitment to collaboration, and this timeframe, provide ideal context for the exercise we have in mind.

There are of course many legal and policy issues to address in the exercise we propose. The third part of our project, conducted by lawyers specialising in AI and social media governance (Tom Barraclough and Curtis Barnes), surveys these issues.

Forward look

For 2022, the Working Group has agreed with GPAI's Steering Committee to continue its two current projects in order to build momentum and realise their potential to have a significant practical impact with GPAI's members.

As has been noted, the Working Group also has broader interests represented by the Committees that were introduced following the 2020 Summit. Whilst building momentum on these two projects, the Working Group aims to develop a broader vision of the Working Group's mandate in 2022.

In the case of the project on a **Responsible AI Strategy for the Environment**, the project will aim to accelerate action on the roadmap with policy makers, investors and developer communities by:

- *Expanding the scope* of the roadmap and the booklet of use-cases to **biodiversity preservation**
- *Growing* **the booklet of uses-cases into a living repository** in collaboration with academia, relevant international organizations, private sector, and civil society actors with the goal of deepening the understanding of responsible AI for climate action potential
- *Engaging the community* by **anchoring the roadmap in key IGOs' agenda** (UNEP, IPCC, COP), in GPAI Member State agendas, and Global South communities - **developing a more bespoke engagement programme if required.** This includes the organization of a limited series of consultation workshops to promote roadmap in key venues (including COP-27 on Climate in November 2022 and COP-16 on Biodiversity in April 2022)

- *Guiding* the community by developing:
  - **strategic implementation plans** for the roadmap identifying the most appropriate and efficient opportunities for international cooperation on Research & Development, deployment, and scaling.
  - **impact planning and assessment frameworks/instruments** such a Global Climate & AI Index, **technical benchmarks** to provide an international comparative baseline for performance against the roadmap. This could help guide investors, policymakers and developer communities in the most impactful opportunities to pursue.
- *Leading the community* by **piloting new mechanisms** - starting with the collaboration on **climate-focused data trusts** with the Data Governance Working Group. The Committee will seek to partner with organisations (e.g foundations) interested in the 'challenge space' identified by the Global Climate & AI Index's technical benchmarks.

By Summit 2022, the Committee will produce Roadmap 2022, offering a more granular perspective enabled by this activity.

For **Responsible AI for Social Media Governance**, the Committee has proposed two components: (1) to continue the New Zealand case study with an extension to the methods it employs, and (2) apply the methods developed for New Zealand in 2021 to other countries.

The continuation of the New Zealand case study will *implement* the fact-finding exercise specified in the current project and *extend* the community consultation methods developed in the current project, to allow them to play a more functional role in Social Media companies. We aim to achieve this through a collaboration between industry and government, in the spirit of GPAI's multi-stakeholder mission.

The application of the methods developed for the New Zealand case study should then serve as a model for how other governments can engage with other Social Media companies, to ask about the effects of its recommender systems on that country's citizens. We will be consulting with GPAI countries on the extension of the project into other countries.

Annex 1

Committee on Climate Change

Co-Chairs

Raja Chatila – Sorbonne University
Nicolas Mailhe – The Future Society

Members

Yoshua Bengio – Mila, Quebec Artificial Intelligence Institute
Marta Kwiatkowska – Oxford University
Christian Lemaître Léon – Metropolitan Autonomous University
Virginia Dignum – Umeå University
David Sadek (Observer) – Thales Group
Karine Perset (Observer) – OECD

Invited specialists

Przemyslaw Biecek – Warsaw University of Technology
Alan Paic – OECD
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, UK Government
Jay Ashton-Butler – Office for Artificial Intelligence, UK Government
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 – European Commission

Committee on Governance & Transparency of Social Media

Co-Chairs

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

Members

Yoshua Bengio – Mila - Quebec Artificial Intelligence Institute
Raja Chatila – Sorbonne University

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

Observers

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

Invited Specialists

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