Responsible Al Working Group Report

November 2022 - GPAI Tokyo Summit



This report was developed by the Global Partnership on Artificial Intelligence's Working Group on Responsible AI. Its contents reflect the opinions of the GPAI Experts involved and do not necessarily represent the official views of GPAI Members. GPAI is a separate entity from the OECD and accordingly, the opinions expressed and arguments employed therein do not reflect the views of the OECD or its Members.

CONTENT

Co-Chair's Welcome	1
Introducing RAI	2
RAI's Members	3
Mandate of RAI	5
Working Group Timeline	6
Progress Report	7
A Responsible AI Strategy for the Environment (RAISE)	8
Responsible AI for Social Media Governance	11
Al for public-domain drug discovery	13
Pandemic Resilience	15
Forward Look	17

Co-Chair's Welcome



Catherine Régis Professor, Faculty of Law, University of Montréal Canada Research Chair, Co-Director H-POD Researcher at Mila



Raja Chatila

Director of the SMAR Laboratory on 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".

Our Working Group unites experts coming from the technical world (e.g. machine learning, information technologies) and experts coming from the social and human sciences sector and fields like law, ethics, philosophy, international affairs, international development, economics, communications, management, and political science. This year we were pleased to welcome the experts of the sub-group on AI and Pandemic Response, whose experts have now been part of our working group since March 2022.

We put at the heart of our work the GPAI Council top three priorities :

- (1) The fight against climate change and biodiversity preservation
- (2) Health and life sciences, including combatting the Covid-19 pandemic and preparing the world for future pandemics
- (3) Impact of AI on human rights, with a focus on gender equality and inclusiveness, factoring AI's potential to impact those with disabilities

Our Working Group strongly supports the overall mission of GPAI and we believe that members and experts can move dialogue towards responsible and ethical use of AI. We believe that like-minded governments, industry, civil society, and the wider public can all together advance AI technologies by aligning with the priorities and values of our societies. We stove to bring 'the whole of society' at pace for the challenges we are currently facing.

We are proud to present our progress that closely align with GPAI Council top priorities and we aim at pursuing our work in 2023 towards those pressing challenges that have been the primary focus of our work and demonstrate the necessity of a 'whole of society' response:

(1) Responsible AI Strategy for the Environment (RAISE) presents how AI can be responsibly developed, used and governed to take on the fight against climate change including biodiversity preservation - one of GPAI top priorities as well as a common priority amongst governments that was recognised by the 2015 Paris Agreement

(2) Responsible AI for Social Media Governance which focuses on improving understanding of social media users' relationships with harmful online content, by developing methods allowing external stakeholders (governments and citizens' groups) to work collaboratively with social media companies, and study the effects of recommender systems - the most ubiquitous AI systems in use today.

(3) Al for public-domain drug discovery that responds to international demand from governments to find ways to accelerate drug discovery more broadly, by providing a set of recommendations for GPAI member countries and the international community in general to create an enabling environment for open AI research towards the development of new drugs or the repurposing of existing ones to address public health challenges.



(4) Pandemic Resilience (before named AI-powered immediate response to pandemics) directly supports impactful and practical AI initiatives to help in the fight against pandemics.

Those four projects address urgent priorities recognised by GPAI members: we therefore hope that the outputs we are presenting at Summit 2022 will support reflections on how we can overcome challenges and guide the responsible development, use and adoption of AI.

Our ambition as a Working Group does not stop here either. Whilst we are excited to pursue these issues in 2022, the Working Group did identify three new projects - presented in our 2023 Work Plan - 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 2023.

Introducing RAI

The Working Group on the responsible development and use of AI (RAI for short) brings together 64 experts from 25 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 the Responsible AI Strategy for the Environment. It has also appreciated the collaboration its experts coming from the AI and Pandemic Response subgroup, on the Pandemic Resilience project which aims at providing AI driven solutions in response to pandemics.

40% of RAI's members are women, a number which we'll work to increase in the future.

Most members (65%) come from the science sector, 25% are from the civil society and 10% are from the industry. A better balance should be achieved in the coming months and years as we believe that the collaboration of *all* stakeholders will be necessary to ensure AI is produced and used in a responsible manner.

RAI also represents an interesting diversity of countries, although more countries and international organizations should be represented in the short and medium term, especially countries and entities from the Global South.

Members were based in 25 countries, that is Argentina, Australia (2), Belgium (3), Brazil (1), Canada (2 people), Denmark, France (3), Germany (3), Greece, India (5), Italy (2), Ireland, Israel, Japan (3), Korea, Mexico (3), the Netherlands (3), New Zealand (6), Poland (2), Singapore (2), Slovenia (4), Spain (2), Sweden, the United Kingdom (6), and the USA (5).

These members have either been designated by the members of GPAI or through the self-nomination process. It's worth mentioning that irrespective of the nomination, all members act with full independence inside RAI.

Finally, 11 additional specialists take part in RAI's activities as observers. One of them is a representative of the OECD, a strategic partner of GPAI, and another one a representative of UNESCO.

Next page presents RAI's experts and observers



RAI's Members

Experts of GPAI's Responsible AI Working Group

Catherine Régis (Co-Chair) – University of Montréal (Canada) Raja Chatila (Co-Chair) – Sorbonne University (France) Aditya Mohan - National Standards Authority of Ireland (Ireland) Alice H. Oh – KAIST School of Computing (South Korea) Alistair Knott - Victoria University of Wellington (New Zealand) Amrutur Bharadwaj - Indian Institute of Science (India) Anurag Agrawal - Council of Scientific and Industrial Research (India) Arunima Sarkar – World Economic Forum (India) Blaž Zupan – University of Ljubljana (Slovenia) Bogumił Kamiński – Warsaw School of Economics (Poland) Carolina Aguerre – Center for Technology and Society / Catholic University of Uruguay (Argentina) Cesar Alberto Penz – Federal Institute of Education, Science and Technology of Santa Catarina (Brazil) Christian Lemaître Léon – Metropolitan Autonomous University (Mexico) Daniele Pucci – Istituto Italiano di Tecnologia Research Labs Genova (Italy) David Sadek - Thales Group (France) Dino Pedreschi – University of Pisa (Italy) Dirk Brockmann – Humboldt University of Berlin (Germany) Dyan Gibbens - Trumbull Unmanned (United States) Emile Aarts - Tilburg University (Netherlands) Enrico Coiera – Macquarie University (Australia) Francesca Rossi – IBM Research (United States) Gaël Varoquaux – INRIA (France) Genevieve Bell – Australian National University (Australia) Hiroaki Kitano – Sony Computer Science Laboratories Inc (Japan) Howie Choset - Carnegie Mellon University (United States) Inese Podgaiska – Association of Nordic Engineers (Denmark) Ivan Bratko – University of Ljubljana (Slovenia) Jack Clark – Anthropic (USA) Joanna Bryson – Hertie School (Germany) Joaquín Quiñonero – LinkedIn (Spain) Juliana Sakai - Transparência Brasil (Brazil) Julie Owono – Internet Sans Frontières (UNESCO) Kate Hannah – The Disinformation project (New Zealand) Kim McGrail – University of British Columbia (Canada) Konstantinos Votis - CERTH / ITI (Greece) Leong Tze Yun – National University of Singapore (Singapore) Luka Omladič – University of Ljubljana (Slovenia) Margarita Sordo-Sanchez – Brigham and Women's Hospital at Harvard Medical School (Mexico) Mario Poljak – University of Ljubljana (Slovenia) Marta Kwiatkowska - Oxford University (EU) Marzyeh Ghassemi – University of Toronto (Canada) Matthias Spielkamp – AlgorithmWatch (Germany) Michael Justin O'Sullivan - University of Auckland (New Zealand) Michael Plank – University of Canterbury (New Zealand) Miguel Luengo-Oroz – UN (Spain) Nozha Boujemaa – IKEA (Netherlands) Osamu Sudo - Chuo University (Japan) Partha Pratim Chakrabarti – Indian Institute of Technology Kharagpur (India) Paul Suetens – KU Leuven (Belgium / European Union nominee) Peter-Paul Verbeek – University of Twente (Denmark) Przemyslaw Biecek – Warsaw University of Technology (Poland) Pushmeet Kohli – DeepMind (UK) Rachel Dunscombe – Imperial College London (UK) Rajeev Sangal – International Institute of Information Technology Hyderabad (India) Rob Heyman – Brussel University (Belgium) Seán Ó hÉigeartaigh – Centre for the Future of Intelligence, University of Cambridge (UK) Toshiya Jitsuzumi – Chuo University (Japan) V K Rajah – Advisory Council on the Ethical Use of Artificial Intelligence and Data (Singapore)



Venkataraman Sundareswaran – World Economic Forum (India) Vincent C. Müller – Technical University of Eindhoven (EU) Virginia Dignum – Umeå University (Sweden / France) Wanda Muñoz – SEHLAC Mexico (Mexico) Yoshua Bengio – Mila (Canada) Yuval Roitman – Israeli Ministry of Justice (Israel)

Observers

Amir Banifatemi – Al and Data Commons Celine Caira – OECD Clara Neppel – IEEE Cyrus Hodes – Blockchain Web Services Karine Perset – OECD Marc-Antoine Dilhac – ALGORA Lab Nicolas Miailhe – The Future Society Prateek Sibal – UNESCO Ricardo Baeza-Yates – Universitat Pompeu Fabra & Northeastern University Stuart Russell – UC Berkeley Tom Lenaerts – Université Libre of Bruxelles Vilas Dhar – The Patrick J. McGovern Foundation

Mandate of RAI

As mentioned in the foreword, RAI's work is grounded in a vision of AI that is human-centered, fair, equitable, inclusive and respectful of human rights and democracy, and that aims at contributing positively to the public good.

RAI's mandate aligns closely with that vision and GPAI's overall mission, that is, RAI strives to foster and contribute to the responsible development, use and governance of human-centered AI systems, in congruence with the UN Sustainable Development Goals.

It is worth noting that RAI, as all other GPAI Working Groups, does not operate in silo within GPAI. Indeed, RAI collaborates with other working groups whenever indicated. For instance, RAI will interface with the Data Governance Working Group when their respective projects share common dimensions.

Finally, it should be mentioned that the *ad hoc* AI and Pandemic Response Subgroup, that was created in July 2020, to support the responsible development and use of AI-enabled solutions to COVID-19 and other future pandemics, was merged to the RAI last february 2022. The projects this group was working on were also transferred under the steer of the RAI.



Working Group Timeline

JANUARY

First Working Group meeting (13th): welcoming new self-nominated experts and project updates.

FEBRUARY

Second meeting of the Working Group (15th) – workshop discussion for the work plan 2023 around the following themes : human rights & AI, governance and global health.

MARCH

Third meeting of the Working Group (10th) – round table discussion on the ideation process for next work plan.

Experts developing project proposals for the RAI 2023 work plan based on the Council top priorities.

APRIL

Fourth meeting of the Working Group (6^{th}) – project updates and presentation of the proposals to be included in the work plan 2023.

Submission of the RAI 2023 Work Plan (29th)

JUNE

Fifth and sixth meeting of the Working Group (1^{ts} and 29^{tht}) – project updates and presentation of the proposals to be included in the work plan 2023.

JULY

Working Group members to meet one-on-one with the Montreal Center of Expertise Coordinator during the summer period. Those meetings allowed us to discuss the work field of each expert and also gather useful comments on the projects, GPAI structure and what ideas could be explored to deliver more impact.

AUGUST

Seventh meeting of the Working Group (31^{stt}) - special presentation on technocoloniality followed by project updates and timeline up to the plenary meeting at the Tokyo Summit.

SEPTEMBER

Eight meeting of the Working Group (26^{tht}) - project updates followed by two presentations by experts on their work to share expertise and knowledge on responsible AI use, development and governance.

OCTOBER

Ninth meeting of the Working Group (26th) – finalisation of the session that will be presented at the plenary session at the Summit followed by a welcoming presentation by two experts of the working group.

NOVEMBER

Tenth meeting of the Working prior to the Summit (15th) Presentation of finalized outputs and open workshop on next projects at the Summit (21st & 22nd)

Progress Report

Building on the preliminary findings presented at the 2021 Summit, the working group observed a pattern in the life cycle of the project's development. During the first year (2021), the issues are defined and framed. In the second year (2022), there's a deeper dive into the topics and extensions are identified. The third year (2023) enables us to work concretely on partnerships and concrete implementations to make the project more practical.

Following the presentation of the working group outputs at the Summit 2021, the working group has continued to build momentum and scale up to deliver further impact on its selected projects.

- (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 made 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 (steering committees fully listed under Annex 1) is co-led by Raja Chatila, Nicolas Miailhe and Lee Tiedrich, and in addition to its partnership with the OECD has collaborated with the Centre for AI and Climate, Oxford University, the WWF, KPMG, and Saint Joseph University.
- (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 and Dino Pedreschi. The project dovetails with ongoing work in the Christchurch Call and the Global Internet Forum to Counter Terrorism, which both bring together Governments, tech companies and citizens' groups with the aim of eliminating terrorist and violent extremist content online.
- (3) Al for public-domain drug discovery: this project responds to international demand from governments to find ways to accelerate drug discovery more broadly, by providing a set of recommendations for GPAI member countries and the international community in general to create an enabling environment for open AI research towards the development of new drugs or the repurposing of existing ones to address public health challenges. The Project Committee (fully listed under Annex 1) is co-led by Alice Oh and Yoshua Bengio, with research support from Mila (Allison Cohen and Elliot Layne).
- (4) Pandemic Resilience : this project reflects the ongoing global emergency presented by COVID-19 and the urgency with which the subgroup was established. Its overarching goal is to directly support impactful and practical AI initiatives to help in the fight against the COVID-19 pandemic. It has two outputs: (1) an update and upgrade of the catalog of practical initiatives that the subgroup commissioned in 2020, transforming it into a living repository, and (2) an evaluation of initiatives to identify impactful and scalable initiatives that could benefit from partnership with GPAI. The insights from these activities will help establish research/technology for fighting against future epidemics/pandemics. The Project Committee (fully listed under Annex 1) is co-led by Michael Sullivan and Margarita Sordo.

A Responsible AI Strategy for the Environment (RAISE)

The world's leading environmental scientists agree that humanity is rapidly approaching and exceeding planetary boundaries, which is why GPAI countries have put climate action and biodiversity preservation at the top of their agenda. As a general-purpose technology, AI can be harnessed responsibly to accelerate positive environmental action.

In 2022, the Working Group and Project RAISE (Responsible AI Strategy for the Environment) began to implement some of the recommendations of its AI & climate roadmap through initiating work towards assessing the environmental impacts of AI compute and preparing AI readiness booklets for key industries to achieve net-zero. The scope was also expanded to start building a roadmap on AI and biodiversity preservation. The Working Group and its experts have worked in partnership with the OECD on climate action. On biodiversity preservation, the group has been supported in its work by the Centre for AI and Climate, Oxford University, the WWF, KPMG, and Saint Joseph University.

These efforts built upon the important foundational work completed by the Working Group in 2021, which was based on a collaboration with Climate Change AI and the Centre for AI and Climate that lead to the publication of an action-oriented set of recommendations to governments to guide policy makers developing climate action strategies¹.

Project (1) The AI footprint: measuring the environmental impacts of AI compute and applications, report drafted in consultation with RAI experts and the OECD.AI Expert Group on AI Compute and Climate.

The green and digital "twin transitions" offer the promise of leveraging digital technologies and artificial intelligence for a sustainable future. At the same time, the computational needs of AI systems are growing, raising sustainability concerns. The joint report² offers a framework and a number of indicators and measurement tools that can help pave the way towards quantifying the direct environmental impacts from AI compute, as well the indirect environmental impacts from AI applications.

The proposed framework distinguishes between direct and indirect positive and negative environmental impacts. Direct impacts stem from the AI compute resources lifecycle (i.e. the production, transport, operations and end-of-life stages). Analysis indicates that direct impacts are most often negative and stem from resource consumption, such as the use of water, energy and its associated greenhouse gas (GHG) emissions, and other raw materials. Indirect impacts result from AI applications and can be either positive, such as smart grid technology or digital twin simulations, or negative, such as unsustainable changes in consumption patterns.

Policy makers must ensure that AI is part of the solution to meet global sustainability targets. A starting point is to address five measurement gaps with policy implications that the report identifies:

- 1. Measurement standards for sustainable AI are needed.
- 2. Data collection on the environmental impacts of AI compute and applications should be expanded.
- 3. Al-specific measurements are difficult to separate from general-purpose compute.
- 4. Environmental impacts beyond operational energy use and GHG emissions should be considered.
- 5. Efforts are needed to improve environmental transparency and equity everywhere.

By creating and tracking AI-specific measures of compute, sharing best practices, and supporting new and innovative AI applications for fighting climate change, countries can ensure that AI is trained and deployed in the most sustainable way possible, while minimizing negative impacts, for the good of the planet.

² OECD (Forthcoming), The AI footprint: measuring the environmental impacts of AI compute and applications, OECD Publishing, Paris



¹ See the 2021 report <u>here</u>

Project (2) AI readiness guidance booklets for net zero in four key sectors: energy, transport, agriculture, and foundation industries.

Al optimisation of electricity systems will be critical to deliver a low-cost transition to net zero. The need for highly efficient balancing markets to rapidly enable more variable renewable generation will require the use of Al to ensure that every energy asset is fully optimized to the market. Al is already being used in a wide range of use-cases in the electricity system, and its importance will only grow as more applications emerge. However, a major roadblock is that many electricity companies are not yet ready to apply Al in their operations.

To support companies in assessing their current level of AI readiness and to map out areas for further investment we provide an AI for Electricity Readiness Assessment. This self-assessment highlights five key themes that electricity companies should advance to become AI ready:

- 1. Opportunity identification: The first step is to identify where AI can be used to support net zero challenges. We recommend companies build comprehensive inventories of AI opportunities to make informed decisions about how to use AI to support decarbonisation.
- 2. Human capacity: Effective management of AI requires a deep and broad understanding of the technology. Executives should receive training in understanding, procuring and managing AI solutions. For in house solutions, data scientists recruited to the sector should be given training to support a thorough understanding of the electricity sector. For external providers companies should ensure their contracting and procurement processes allow ongoing iteration and learning.
- 3. Data: The quality of AI solutions is highly dependent on the quantity and quality of the data used. To effectively apply AI companies will need to invest in AI-grade data collection, management and sharing systems. For any specific AI project companies will need to identify the data required for the project, review existing internal data, identify data capture needs, and invest in data cleaning and labeling to ensure data is AI-ready.
- 4. Digital Infrastructure: Powerful AI algorithms require enhanced compute and data storage infrastructure. An important choice is between hosting infrastructure through on-premises hardware or in the cloud. Companies should assess and procure the best solution for compute and data storage for initial pilot projects, and subsequently assess and procure the best solution for compute and data storage for wider organizational AI roll out.
- 5. Al Governance: Al is a powerful tool with many benefits but also some risks. We recommend electricity companies establish principles and processes to manage Al-specific risks.

The AI readiness guidance for the energy sector has been prioritized for publication as electricity companies worldwide are urgently looking for ways to conserve energy and ensure energy security. The three booklets for transport, agriculture, and foundation industries are set for subsequent publication in the months following the summit, which allows for continuous engagement and visibility for Project RAISE and GPAI.

Project (3) Biodiversity and AI: Opportunities and Recommendations for Action

The draft report on AI and biodiversity is set to be disseminated for feedback and consultation ahead of COP15 in Montreal in December. With the relationship between AI and biodiversity preservation being much less established and understood than for climate action, a series of emerging findings are summarized below.

Biodiversity is being lost rapidly creating an increasing risk of ecosystem collapse, which would undermine the critical life-supporting functions that biodiversity provides humanity. There is an urgent need for action to address this crisis. This will require unprecedented levels of coordinated action and a move away from business-as-usual approaches to biodiversity conservation.

Al is one tool that we can use to support enhanced action on biodiversity loss, however it won't solve the problem and should only be applied where it can add real value which outweighs risks and environmental cost. The report seeks to help build a better understanding of where Al can add value to biodiversity conservation, and proposes how we can advance the responsible application of Al.

At present the majority of AI-for-biodiversity applications are focused on enhancing current approaches to conservation, albeit at a greater scale than previously conducted. The most common existing applications involve AI supporting biodiversity monitoring by helping classify species and landscapes



captured through camera traps and satellite images, and to some extent, the monitoring of the drivers of biodiversity, for example by monitoring fishing trawlers or illegal timber logging.

These are important applications and AI can increase the scale, speed and accuracy of many of them. However, to create a step change in biodiversity conservation, these approaches need to be augmented with novel, innovative ways of shifting the politics of biodiversity conservation to place it firmly on the agenda of policy makers. This will require new ways of engaging and empowering citizens and communities to manage their land and to advocate for biodiversity; increasing transparency around the immediate political and financial drivers of biodiversity policy development; improved prioritization on biodiversity strategy and policy; and driving transparency and optimisation across key commodity value chains and financial asset classes. Though it cannot solve these problems, AI has the potential to play a supporting role in these efforts.

Partnerships and Dissemination

The results of these projects are expected to be presented at the GPAI 2022 Summit in Tokyo, as well as at the UNFCCC COP27 in Egypt, and the UN Biodiversity Conference COP15 in Montreal. RAISE also engaged with various partners to anchor the recommendations in international agendas, including UNEP's CODES (Coalition for Digital Environmental Sustainability) network, the IEEE Planet Positive 2030 initiative, the Research Coordination Network for the Digital Economy and Environment, and is also envisioning the possibility of an "exploratory" joint workshop with IPCC and IPBES, to be initiated in early 2023. The committee, which includes experts from different Working Groups, works closely with the Data Governance WG for instance to identify climate and biodiversity use cases for the data trust and PETs projects.

Outlook for 2023

In 2023, RAISE will continue to deepen its work, operationalising its AI adoption strategy for climate action and biodiversity preservation and implementing the opportunities identified with a target audience that includes GPAI member countries, international organizations and investors. This includes a "Responsible AI for the Environment Framework" and AI impact assessment, organizing workshop series with GPAI members in the new sherpa format model to drive deeper engagement and meaningful impact, and developing a structured cooperation program with IPCC and IPBES. RAISE will also seek to prototype and scale AI solutions for climate and biodiversity action, through implementation partnerships and by securing external funding. Several GPAI member countries have already expressed high interest in supporting and co-championing such efforts.



Responsible AI for Social Media Governance

Social media platforms rely on several kinds of AI technology for their operation. Much of the appeal of these platforms comes from their ability to deliver content that is *tailored* to individual users. This ability is provided in large part by AI systems called **recommender systems**: these systems are the focus of our project.

Recommender systems curate the 'content feeds' of platform users, using machine learning techniques to tailor each user's feed to the kinds of item they have engaged with in the past. They essentially function as a personalised newspaper editor for each user, choosing which items to present, and which to withhold. They rank amongst the most pervasive and influential AI systems in the world today.

The starting point for our project is a concern that recommender systems may lead users in the direction of harmful content of various kinds. This concern is at origin a technical one, relating to the AI methods through which recommender systems learn. But it is also a social and political one, because the effects of recommender systems on platform users could potentially have a significant influence on currents of political opinion.

At present, there is very little public information about the effects of recommender systems on platform users: we know very little about how information is disseminated to users on social media platforms. It is vital that governments, and the public, have more information about how recommender systems steer content to platform users, particularly in domains of harmful content.

In the first phase of our project which culminated in the <u>2021 report</u>, we reviewed possible methods for studying the effects of recommender systems on user platform behaviour. We concluded the best methods available for studying these effects are the methods that companies use themselves. These are methods that are only available internally to companies. We proposed *transparency mechanisms*, in which these company-internal methods are used to address questions in the public interest, about possible harmful effects of recommender systems.

We focussed on the domain of Terrorist and Violent Extremist Content (TVEC), because this type of content is already the focus of discussion in several ongoing initiatives involving companies, including the Global International Forum to Counter Terrorism (GIFCT) and the Christchurch Call to Eliminate TVEC Online. Our proposal was for a form of **fact-finding study**, that we argued would surface relevant information about recommender system effects in this area, without compromising the rights of platform users, or the intellectual property of companies. We presented and argued for this proposed fact-finding study at last year's GPAI Summit.

Over the past year, our project has pursued the practical goal of piloting our proposed fact-finding study in one or more social media companies. This has involved discussions with several companies, often mediated by governments; and participation in several international initiatives relating to TVEC, in particular the Christchurch Call and the Global Internet Forum to Counter Terrorism (GIFCT). At the recent Christchurch Call Summit, a scheme for running a pilot project of the kind we advocate was announced: the initiative involves two governments (the US and New Zealand) and two tech companies (Twitter and Microsoft), and centres on the trialling of 'privacy-enhancing technologies' developed by a third organisation, OpenMined. In our report for this year, we summarise the discussions that led to this initiative, in the context of other ongoing discussions around transparency mechanisms for recommender systems.

We are very much looking forward to participating in the scheme initiated by the US and New Zealand, and engaging with Twitter, Microsoft and OpenMined. But discussions about recommender system transparency will certainly continue beyond this initiative as well. Our report also offers some recommendations about how these ongoing discussions can be made more efficient.

• First, we suggest that discussions should have more involvement from *company engineers*. At present, companies are represented primarily through legal and policy teams. But the questions



under discussion concern technical mechanisms operating within companies: the engineers who design and use these mechanisms could make valuable contributions, under suitable non-disclosure arrangements.

- Second, while privacy-enhancing technologies are potentially very valuable, we foresee the need for an *ongoing discussion* about these technologies between companies and external stakeholders. As companies' technologies develop, and new questions arise, the functionality of privacy-enhancing technologies may have to be expanded, or adjusted.
- Third, we don't see privacy-enhancing technologies as a substitute for discussions with company engineers. We think discussions with engineers should begin immediately, while privacy-enhancing technologies are being developed; and we foresee a role for discussions with engineers even after these technologies are first put in place.
- Fourth, we suggest transparency discussions about recommender systems should focus on the specification and implementation of *pilot projects* that can be trialed in particular companies. Discussions with company engineers can define a set of possible pilot projects. The proposed pilots can then provide a concrete focus for discussions with company lawyers, to ensure protection of user rights and company intellectual property. These discussions with individual companies can inform broader discussions about general transparency processes that apply across companies.
- Fifth, we foresee an ongoing role for the piloting of recommender system transparency mechanisms, as social media platforms continue to change and develop. We suggest that discussions about pilots could be coordinated by an independent regulatory body, sitting between companies and governments, and informed by a new public science of recommender system effects

Al for public-domain drug discovery

In 2022, the Working Group explored various means of creating, promoting, and incentivizing AI-uptake in the drug development process. The Working Group has been supported in this work by Quebec's AI Institute Mila and McGill University.

These efforts built upon the important foundational work completed by the Working Group in 2021, which included an assessment of the current drug discovery ecosystem, its challenges, and opportunities for AI innovation in its space³.

The Drug Discovery Project Committee's efforts specifically focused on addressing existing barriers in the following three target areas:

- (1) Research and development in fields of drug discovery that are valuable to society, and necessary to public health, are not being sufficiently addressed by stakeholders in industry.
- (2) Uptake of AI throughout the entire drug discovery and development pipeline.
- (3) A shift in culture towards open-data among stakeholders in academia and industry when undertaking research on drug discovery and development.

Stakeholders from both public and private sectors (which included industry, international health organisations, and venture capitalist firms) were engaged to provide feedback on the proposed recommendations of the roadmap as means for progressing the above-listed focus areas. These recommendations are summarised below:

- 1. Governments need to invest in multi-disciplinary academic research in the field of AI-driven drug discovery, and especially fund research into applications of AI in public health concerns where there is currently insufficient commercial interest and investment.
- 2. Governments should incentivize AI-capacity building inside of the drug discovery and development ecosystem.
- Governments should set up novel innovation procurement programs for stimulating and incentivizing the efforts of biotech, pharma, healthcare or public research organisations, to go from academic prototypes (software, biotech methodologies, candidate drugs) to industrialstrength development pipelines and optimised drugs.
- 4. Governments should set up financial incentives when the drugs being developed are sufficiently promising to address significant public health issues and when usual commercial incentives are not sufficient to motivate the pharmaceutical industry to fund the clinical trials themselves.
- 5. Governments should internationally coordinate (1,2,3,4) above to favour
 - a. research collaborations, knowledge sharing and transfer of know-how across countries and in particular from richer countries to LMICs;
 - **b.** more uniform innovation policies (across countries) in their procurement and incentive mechanisms in (3) and (4) to make it easier for companies that are involved to comply with similar legal and operational frameworks across countries;
 - c. access to the resulting technologies and drugs at low prices in LMICs;
 - d. joint funding on efforts with international scope (see (6)).
- 6. Either create a permanent international non-profit organisation or leverage an existing one which will be responsible over the longer term to coordinate (as recommended in (5)) across countries, as well as manage internationally funded projects (from discovery to fabrication to deployment),

³ See the full 2021 report <u>here</u>.



and make sure to fund this organisation to allow it to reach the desired goals.

A summary of the feedback received is provided below:

Public Sector

- There exists global interest in tackling severe public health threats multiple government bodies have listed Antimicrobial Resistance (AMR) and bolstering drug R&D as top priorities
- There is a general agreement among G7 countries that public health priorities should be collaboratively managed at the global level
- There is an introductory level of knowledge within governments on AI's potential in the domains of predictive analytics, surveillance, and drug development

Private Sector

- Although open data and open science themes are preferred, existing infrastructure and assets are insufficient time lags and inappropriate data sets for the application at hand generally leads to the need to create assets in-house
- Creating biotech companies and other small and medium enterprises in the drug discovery industry is challenging due to the large, up-front capital expenditure that is required; rendering government investment in this domain vital
- The traditional business model of Venture Capital firms is not aligned with the required time horizons of the drug R&D lifecycle, however AI may present efficiencies which would help to align the timelines of these processes
- Greater amounts of innovation can be derived from the demand-side, rather than the supply-side of the market; meaning, if the drug discovery ecosystem is predicated on consumer demand rather than available supply, there will be greater innovation in the domain
- International health agencies often act conservatively without having the freedom and flexibility to innovate; thus, navigating the bureaucracy of the institution itself is critical using strategic points of intervention
- Coordination is challenging data governance agreements are required since populations may have differing approaches to how they wish their data to be governed, protected, and compensated.
- Innovation is disproportionate when comparing the Global North and the Global South, with the Global South often contributing to datasets in an extractive process that does not result in a sharing of benefits

The work completed this year will be summarised in a report⁴ in advance of the 2022 Summit, where the Project Committee looks forward to discussing the findings and possible next steps with GPAI experts and members.

⁴ See the 2022 summary report <u>here</u>.



Pandemic Resilience

In 2022, the Working Group explored possible collaborations between candidate AI initiatives focused on the role of non-pharmaceutical interventions (NPIs) in pandemic management.

The Working Group has been supported in this work by the COVID-19 International Modelling Consortium (CoMo Consortium), Covid Modelling Aotearoa (CMA), University of Auckland, Carnegie Mellon University (CMU) Africa, and Cognizant Technology Solutions.

This collaboration effort built upon foundational work completed by the Working Group throughout 2020 and 2021, which included cataloging and assessing a representative set of promising initiatives relevant for the COVID-19 pandemic and beyond⁵, summarising those initiatives with intrinsic scalability and high potential to mitigate the current and future pandemics⁶, preparing an expansion to the initial set of summaries⁷, and compiling an accompanying living repository⁸ (which is being transformed into a live, online repository for Summit 2022).

Amongst the discovered initiatives, many had a focus on non-pharmaceutical interventions (NPIs). The non-intrusive nature and the compatibility of these initiatives provided an opportunity to bring them together in the creation of a tool that could inform the decision making for present and future pandemics.

With this goal in mind, the Project Committee began to explore possible collaborations with candidate initiatives focusing on the following objectives.

- Compare how different combinations of border control strategies, home isolation, and testing at varying levels of vaccine coverage affect the risk of an infected traveler causing a community outbreak of COVID-19;
- Determine the effectiveness of non-pharmaceutical interventions on COVID-19
- aid policymakers' and health officials' decision-making across various topics related to the COVID-19 pandemic;
- Help policymakers simulate pandemic spread dynamics for different government control measures;
- Illustrate how the COVID-19 pandemic could develop under different national guidelines throughout the pandemic;
- Provide a country-level risk modeling framework intended to assist the government and individuals in making informed decisions;
- Automatically determine the most effective non-pharmaceutical intervention (NPI) strategies to contain the spread of COVID-19; and
- Model the spread of COVID-19 based on the prevalence of mask-wearing in a population.

A workshop was held, inviting modelling groups from various initiatives to present their work, and then to discuss opportunities for possible collaboration, with the following outcomes:

• Bring together models of non-pharmaceutical interventions (NPIs) from various initiatives throughout the world;

⁸ See the living repository file here



⁵ See the full 2020 report <u>here</u>

⁶ See the full 2021 report here

⁷ See the full 2022 report here

- Embed these models within an automated calibration framework that will tune the model parameters to best fit available data;
- Extract the effects of individual and combined NPIs to provide a decision model for deployment of NPIs;
- Add consideration of economic effects to these models; and then
- Test the process on data from sub-Saharan countries to help inform their strategy for fighting against Covid-19 and keeping their economies stable.

Each modelling team provided a summary of their model and how it functions, including highlighting underlying mechanisms, inputs and outputs, and any special considerations. Datasets used were discussed, in addition to an exploration of shared/common parameters to help contextualise possible calibration frameworks.

Following the initial discovery workshop, five modelling groups agreed to participate in the codevelopment aspects of the project. A second workshop, to explore how this collaboration can proceed, was held at the time of writing - the Project Committee looks forward to discussing its outcomes and learnings from this process and next steps at Summit 2022.

Forward Look

The Working Group has agreed to work on projects that align closely with GPAI trop 3 priorities :

- (1) The fight against climate change
- (2) Health and life sciences, including combatting the Covid-19 pandemic and preparing the world for future pandemics
- (3) Impact of AI on human rights, including gender equality and inclusiveness, factoring AI's potential to impact those with disabilities

For 2023, the Working Group aims at continuing two of its current projects (RAISE and Social Media) and starts three new ones that were identified through an ideation process started last January 2022. These new proposals would focus on issues that could benefit the 'whole of society' by building momentum and a significant practical impact with GPAI's members.

The Working Group has proposed the following three new projects for 2023 subject to GPAI Council approval at the Tokyo Summit:

- (1) Towards Real Diversity and Gender Equality in Artificial Intelligence: Evidence-Based Promising Practices and Recommendations
- (2) Sandbox for responsible governance a procurement scenario
- (3) RAI deployment challenge & fund

For **Towards Real Diversity and Gender Equality in Artificial Intelligence: Evidence-Based Promising Practices and Recommendations**, the Working Group aim to contribute ensuring that the AI ecosystem, in particular States in the role as duty bearers, have the tools and the commitment to incorporate effective gender and diversity approaches throughout the AI cycle, and to demonstrate their impact and results with indicators in accordance with international standards. This responds to one of the Council's priorities on the impact of AI on human rights. The project also seeks to be highly practical by demonstrating the potential risks of AI and provide practical solutions to mitigate them. This project will work on the intersectionality approach by considering factors such as race, ethnicity, and disability.

For **Sandbox for responsible governance – a procurement scenario** we propose to develop a prototype "Procurement Sandbox" from within the family of regulatory sandboxes, which includes process steps and a simulation environment, to help address gaps in evidence, policies and capacity. The objective is to define concrete functionality for auditing and/or conformity assessments and provide a simplified specification of how these can be understood and evaluated in a sandbox environment. This contributes and extends current efforts on operationalizing governance agenda, and can lead to concrete procurement guidelines, as well as benchmarking, standards and metrics.

Finally, the Working Group suggested a project that would provide opportunities to deploy and scale responsible AI solutions, encouraging cross-functional collaboration, and helping adopt performance metrics would greatly contribute to tangible outcomes that could help operationalize the Responsible AI framework within GPAI and even beyond. The **RAI deployment challenge & fund** would complement existing and future GPAI projects and would create an opportunity to provide a practical pathway for implementation, testing, and validation to benefit all parties involved in responsible AI developments and dissemination.

As noted earlier in this report, the Working Group will also focus on completing two of its current projects:

(1) For Responsible AI Strategy for the Environment (RAISE), will continue to deepen its work,



developing an AI adoption strategy for climate action and biodiversity preservation and implementing the opportunities identified with a target audience that includes GPAI member countries, international organisations and investors.

(2) For the project *Responsible AI for Social Media Governance*, the Committee has proposed two objectives for 2023: (1) build a collaboration with a social media company to conduct a fact-finding study to investigate the hypothesis that these systems may incrementally move users towards various kinds of harmful content – in particular, towards terrorist and violent extremist content (TVEC), and (2) based on the outcomes of the fact-finding studies, to consider how recommender systems could be modified, to alleviate any negative effects that are found to achieve socially beneficial effects.

We're looking forward to starting 2023 with these upcoming projects in the pipeline. We're hopeful that the next months will be productive and our future research agenda will guide the next steps on opportunities to go further and deeper in advancing research and practice on responsible development, use and governance of AI.



Annex1

RAISE Project advisory group

Co-Leads

Raja Chatila – Sorbonne University Nicolas Miailhe – The Future Society Lee Tiedrich – Duke University

Biodiversity Steering Committee

Dave Thau – WWF Tanya Berger-Wolf – Ohio State University Christian Lemaitre Léon – UAM Sara Beery – Caltech Daniel Barthélemy – CIRAD

AI Readiness Steering Committee

David Jensen – UNEP Arunimar Sakar – World Economic Forum Aidan O'Sullivan – UCL Irene Sturm – Deutsche Bahn Jack Kelly – Open Climate Fix Sarah Hunter – Google X

Expert Groups Consulted

OECD.AI Expert Group on AI Compute and Climate

Karine Perset – OECD Celine Caira – OECD Jack Clark – Anthropic Keith Strier – NVIDIA Jennifer Tyldesley – UK DCM

Project advisory group on Governance & Transparency of Social Media

Co-Leads

Alistair Knott – Victoria University of Wellington Dino Pedreschi – University of Pisa

Members

Raja Chatila – Sorbonne University Ricardo Baeza-Yates – Universitat Pompeu Fabra & Northeastern University Alejandro Pisanty Baruch – National Autonomous University Toshiya Jitsuzumi – Chuo University

Observer Stuart Russell – University of California, Berkeley

Invited Specialists

David Eyers – University of Otago Andrew Trotman - University of Otago Tapabrata Chakraborti – University of Oxford Curtis Barnes – Brainbox Institute, Auckland Tom Barraclough – Brainbox Institute, Auckland Lama Saouma – GPAI's Montreal Center of Expertise (CEIMIA) Virginia Morini – University of Pisa Valentina Pansanella – Scuola Normale Superiore

Project advisory group on AI for public-domain Drug Discovery

Co-Leads

Yoshua Bengio - Founder and Scientific Director, Mila – Quebec Artificial Intelligence Institute & Professor of Computer Science, University of Montreal Alice Oh - Associate Professor of Computer Science, Korea Advanced Institute of Science and Technology (KAIST)

Invited Specialists

Allison Cohen - Applied Al Projects Lead, Al for Humanity - Mila Elliot Layne - PhD Candidate - McGill University, School of Computer Science

Project advisory group on Pandemic Resilience

Co-Leads

Michael Justin O'Sullivan – University of Auckland Margarita Sordo-Sanchez – Brigham and Women's Hospital at Harvard Medical School Paul Suetens – KU Leuven