Responsible AI Working Group Report

December 2023 - New Delhi Summit

GPAI / THE GLOBAL PARTNERSHIP ON ARTIFICIAL INTELLIGENCE

This report was developed by the Global Partnership on Artificial Intelligence's Expert 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.

The GPAI Expert Working Group Responsible AI agreed to declassify this report and make it publicly available.

Co-Chair's Welcome

Catherine Régis
Professor, Faculty of Law,
University of Montréal
Canada Research Chair and
Canada-CIFAR Chair in AI,
Associate Academic Member
at Mila

Raja Chatila
Profesor Emeritus
Robotics and Ethics
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, ensuring diversity and inclusivity to promote a resilient society, in particular, in the interest of vulnerable and marginalised groups**."

Our Expert Working Group considers that ensuring responsible and ethical AI is more than designing systems whose results can be trusted - it is about the way we design them, why we design them, and who is involved in designing them. Responsible AI is not, as some may claim, a way to give AI systems some kind of 'responsibility' for their actions and decisions, and in the process, discharge people, governments and organizations of their responsibility. Rather, it is those that shape the AI tools who should take responsibility and act in accordance with the rule of law and in consideration of an ethical framework - which includes respect for human rights - in such a way that these systems can be trusted by society.

In order to develop and use AI responsibly, we need to work towards technical, societal, institutional, legal methods, and tools that provide concrete support to AI practitioners and deployers, as well as awareness and training to enable the participation of all, to ensure the alignment of AI systems with our societies' principles, values, needs, and priorities, where the human being is at the heart of the decisions and the purposes in the design and use of AI.

The foundational work we carried out for the last three years has been essential to fulfilling GPAI's mission of supporting applied AI projects while providing a mechanism for sharing multidisciplinary analysis, foresight, and cooperation of AI practitioners from academics, AI entrepreneurs, and professionals engaged in AI work across private and public sectors.

We are proud to present in this report our project progress that closely aligns with the current GPAI council priorities :

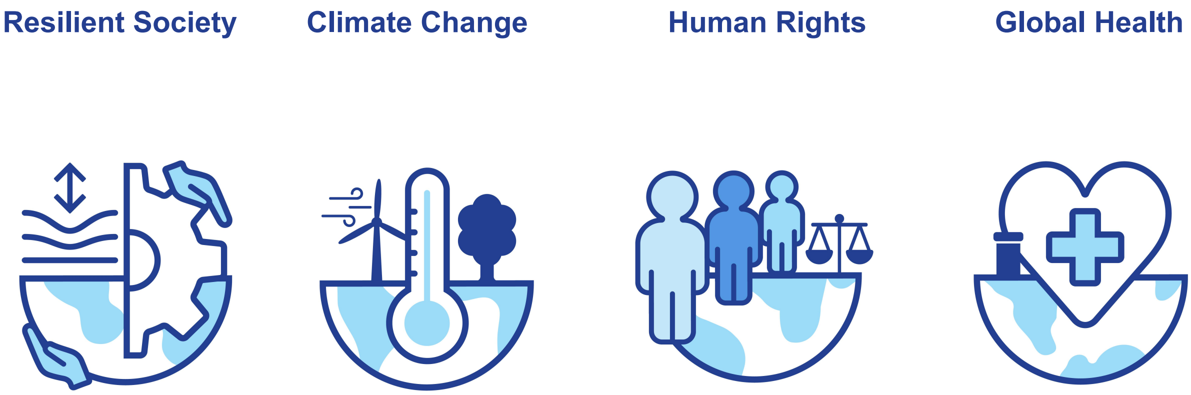
(1) Ensuring a resilient society knowing the challenges ahead and preparing for them;

(2) Continue strengthening our efforts to mitigate the effects of climate change across humanity including promoting the preservation of biodiversity;

(3) Making AI a powerful and effective tool while ensuring that future society is built upon respect of human rights

(4) Supporting the healthcare systems of all nations through AI including addressing the future new pandemics and threats to health that will require international coordination and cooperation.

GPAI Council Priorities 2023



Our ambition as a Working Group does not stop there. Whilst we are excited to pursue our work around these challenges in 2024, our Working Group has identified new initiatives emerging from the first GPAI Innovation Workshop held in-person in Montreal in September 2023. This workshop sought to establish a direct dialogue between Experts and Members to dig into the real, concrete challenges of GPAI Members with regards to the impact and governance of Generative AI. The Innovation Workshop created a space for both GPAI Experts and Members to co-design impactful initiatives which are now integrated into the Work Plan 2024.

In closing, we would like to thank all the Experts of the Working Group for their dedication, commitment, creativity and hard work. It was a real privilege to co-chair this group of brilliant minds who genuinely care about the future of AI development for social good. We look forward to the great projects coming ahead in 2024.

Introducing the Responsible AI Expert Working Group

The Working Group on the responsible development, use and governance of AI (RAI for short) brings together 45 experts, including two Observers, from 26 countries around a shared mandate that is grounded in a vision of AI that is human-centred, fair, equitable, inclusive and respectful of human rights and democracy, and that aims at contributing positively to the public good. RAI aligns closely with GPAI's overall mission, striving to foster and contribute to the responsible development, use and governance of human-centred AI systems, in congruence with the UN Sustainable Development Goals and human rights including ensuring diversity and inclusivity to promote a resilient society, in particular, in the interest of vulnerable and marginalised groups.

It is worth noting that RAI and the other GPAI Expert Working Groups do not operate in silos. Indeed, RAI continuously collaborates in shaping the overall mission of the GPAI through the activities carried forward in 2023 through the Multistakeholder Expert Group which we familiarly call the MEG - reuniting all Expert Working Groups. This year RAI contributed to the Townhall Meeting on Generative AI, the Experts Working Group convenings and the first Innovation Workshop held in Montreal last September 2023. The Working Group is also eager to pursue our collaboration through next year's MEG transversal project on *Safety and*

Assurance of Generative AI (SAFE). RAI is looking forward to pursuing its work with the Data Governance Working Group on the elements of the project that relate to data and is looking forward to our joint collaboration on next year's new project *Repositories of Public Algorithms*. Worth to mention that one of our project co-leads (Lee Tiedrich, project RAISE) is a member of the Innovation and Commercialization WG.

Currently, 31% of RAI's Experts are women, a number which we'll work to increase in the future. Most members (67%) come from the science sector, 20% are from civil society and 13% are from industry. Our group also represents an interesting diversity of countries, although more countries should be represented, especially middle-low-income countries. A better balance should be achieved in the coming months and years as we believe that the collaboration of *all stakeholders* is necessary to ensure responsible governance of AI.

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

Finally, two Observers take part in RAI's activities. One is a representative of the OECD and the second is a representative of UNESCO.

The next page presents the Working Group'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)
Adrian Weller – Centre for Data Ethics and Innovation (United Kingdom)
Alistair Knott – Victoria University of Wellington (New Zealand)
Amir Banifatemi – AI and Data Commons (United States)
Arunima Sarkar – World Economic Forum (India)
Bogumił Kamiński – Warsaw School of Economics (Poland)
Clara Neppel – IEEE (Austria)
Daniele Pucci – Istituto Italiano di Tecnologia, Genova (Italy)
Dino Pedreschi – University of Pisa (Italy)
Dubravko Ćulibrk – Institute for Artificial Intelligence Research and Development of (Serbia)
Emile Aarts – Tilburg University (Netherlands)
Farah Magrabi – Macquarie University, Australian Institute of Health Innovation (Australia)
Francesca Rossi – IBM Research (United States)
Hiroaki Kitano – Sony Computer Science Laboratories Inc (Japan)
Inese Podgaiska – Association of Nordic Engineers (Denmark)
Ivan Bratko – University of Ljubljana (Slovenia)
Ivan Reinaldo Meneghini – Instituto Federal Minas Gerais (Brazil)
Joaquín Quiñonero – LinkedIn (Spain)
Juan David Gutierrez – Universidad de los Andes (Colombia)
Juliana Sakai – Transparência Brasil (Brazil)
Kate Hannah – The Disinformation project (New Zealand)
Konstantinos Votis – CERTH / ITI (Greece)
Mehmet Haklidir – TUBITAK BILGEM (Türkye)
Michael Justin O'Sullivan – University of Auckland (New Zealand)
Miguel Luengo-Oroz – UN (Spain)
Myuhng-Joo Kim – Seoul Women's University (Korea)
Nicolas Miailhe – The Future Society
Osamu Sudo – Chuo University (Japan)
Paola Ricaurte Quijano – Tecnológico de Monterrey (Mexico)
Przemyslaw Biecek – Warsaw University of Technology (Poland)
Rachel Dunscombe – Imperial College London (UK)
Ricardo Baeza-Yates – Universitat Pompeu Fabra & Northeastern University (Spain)
Rob Heyman – Brussel University (Belgium)
Sandro Radovanović – University of Belgrade (Serbia)
Seydina Moussa Ndiaye – FORCE-N Program at the Cheikh Hamidou Kane Digital University (Senegal)
Stuart Russell – UC Berkeley (United States)
Susan Leavy – School of Information and Communication,University College Dublin (Ireland)
Tom Lenaerts – Université Libre de Bruxelles/ FARI (Belgium)
Venkataraman Sundareswaran – World Economic Forum (India)
Vilas Dhar – The Patrick J. McGovern Foundation (United States)
Virginia Dignum – Umeå University (Sweden)
Yuval Roitman – Israeli Ministry of Justice (Israel)

Observers

Celine Caira – OECD
Dafna Feinholz – UNESCO

Working Group Timeline

JANUARY

First Working Group meeting (13th): kick-off meeting to welcome new self-nominated and member-nominated experts. This meeting also presented the objectives for the six approved projects including a first brainstorming session to start off the three new projects coming ahead for the year.

FEBRUARY

Second meeting of the Working Group (13th) – project updates followed by two Experts presentations on their work to share expertise and knowledge on responsible AI use, development and governance.

MARCH

Third meeting of the Working Group (20th) – workshop on delivering impact and defining which key performance indicators could be used to assess the impact of the projects.

APRIL

Fourth meeting of the Working Group (12th) – project updates followed by a plenary discussion to identify the potential risks and benefits of Generative AI including defining the ideal set of responsibilities for providers of these AI solutions.

MAY

Fifth meeting of the Working Group (8th) – plenary discussion to identify and define what key messages the RAI Working Group will bring to the attention of GPAI State Members to responsibly assess the development of Generative AI. This meeting allowed the Experts of the group to address questions raised by the State Members in preparation for the Townhall meeting

➔ GPAI Townhall meeting on Generative AI held live May 15th 2023

JUNE

Sixth meeting of the Working Group (13th) – project updates and plenary discussion of next year's project proposals to be included in the work plan 2024.

JULY

Seventh meeting of the Working Group (5th) - project updates and presentation of the next year's project proposals to be included in the work plan 2024 followed by an agreement of the Working Group through a written process.

SEPTEMBER

Eighth meeting of the Working Group (5th) - special Working Group convening co-organised by India GPAI Summit Host and Council's Incoming Chair to present the RAI Working Group projects and engage through breakout sessions in discussion with the live audience.

OCTOBER

Ninth and tenth meeting of the Working Group (4th and 2nd) – project outputs finalisation by providing guidance to review the reports and preparation for the Summit sessions.

DECEMBER

Presentation of finalized outputs at the New Delhi GPAI Summit (12th to 14th)

Progress Report 2023

Overview of the Responsible AI Expert Working Group Projects

Responsible AI Strategy for the environment (RAISE)

AI can be harnessed to support the fight against climate change including the preservation of biodiversity. Its foundational work, launched at COP26, presents an action-oriented set of recommendations for policymakers to develop climate action strategies. Since then, the project has continued to deepen its work by developing guides and tools on how AI can support biodiversity preservation and help key industries support their net-zero transition. The project also developed an AI footprint to help measuring the environmental impacts of AI compute.

Social Media Governance

This project responds to growing concerns about the misuse of social media platforms, which can be harmful and serve to propagate disinformation, extremism, violence, harassment and abuse. The objective of the project is to analyse and make recommendations on recommender systems, harmful content classifiers and foundation models.

Scaling Responsible AI Solutions

Implementing Responsible AI across public and private organizations is crucial and deploying at scale remains problematic. To address this need, this project provides opportunities to deploy and scale Responsible AI solutions to contribute to the operationalization of the Responsible AI framework. For the first year GPAI experts are providing direct mentorship and workshop development to a group of selected teams developing Responsible AI Solutions. The group extracted the main challenges AI-focused teams face in real world use cases to provide governments and AI-focused teams with recommendations for scaling responsibly.

Towards Real Diversity and Gender Inclusion in AI Ecosystems

AI offers a wide range of possibilities for enhancing the well-being of different groups and contributing to the SDGs. However, AI can also deepen economic, knowledge, gender and cultural divides. AI is generally designed, developed, monitored, and evaluated without systematic gender and diversity approaches, becoming an obstacle to the development and adoption of Responsible AI. The project provides guidance to ensure just, responsible, and inclusive AI throughout its full life cycle to ensure diversity and inclusivity for a resilient society.

Sandbox for Responsible AI Public Procurement

This project proposes to develop a sandbox pilot for a responsible AI public procurement use case, to help address gaps in evidence, policies and capacity. The project is defining concrete functionality for auditing assessments and provide a simplified specification of how these can be understood and evaluated in a sandbox environment. For its first phase, the project is planning to develop at least two use cases scenarios with governments.

Pandemic Resilience

This project reflects the global emergency presented by COVID-19 and the need for collaboration for an efficient and timely response to global threats. Its overarching goal is to directly support impactful and practical AI initiatives to help in the fight against the COVID-19 pandemic. Through the living repository of impactful and scalable initiative developed by this GPAI project in its first phase, in 2023 the group gathered some of the most promising initiatives to develop an AI-calibrated ensemble of pandemic spread prediction models. The outcome demonstrates a high potential for better predictions in face of uncertainty and provides recommendations for decision makers to strengthen their evidence based decision making.

A Responsible AI Strategy for the Environment (RAISE)

Project Co-Leads

Raja Chatila
Professor Emeritus
Sorbonne University

Nicolas Miailhe
Founder & President
The Future Society

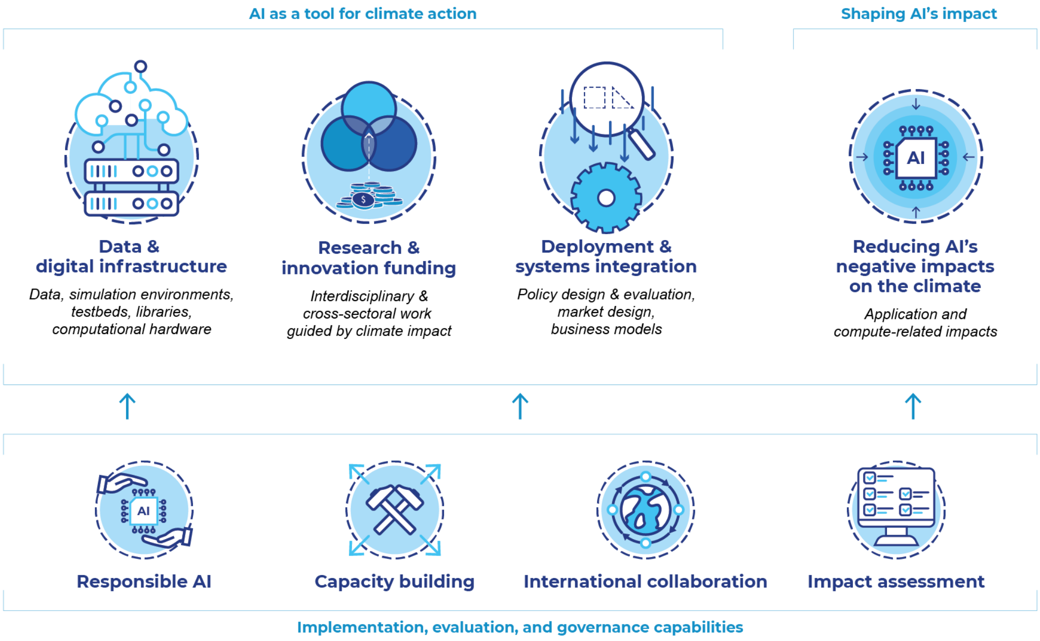
Lee Tiedrich
Visiting Professor
Duke Law School

The world's leading environmental scientists agree that humanity is rapidly approaching and exceeding planetary boundaries. The world population continues to increase and is about to reach 8 billion, with the effects on the environment becoming more and more noticeable. If current activity is maintained, the planet may face a warming in excess of 1.5 °C between 2030 and 2052.

In this context GPAI member countries have put climate action and biodiversity preservation at the top of the agenda. General-purpose AI technologies, particularly when fostering inclusion, may offer novel solutions to help move towards a low-carbon economy as well as to adapt to the impacts of climate change. They can identify opportunities to reduce emissions and can be used to model local climate impact systems that may nevertheless contribute to climate change and resource consumption, with an exponential increase in computing power usage to produce, store and analyse the mass of data necessary for AI systems.

Building from the learning of its 2021 foundational work, an action-oriented set of recommendations to governments to guide policy makers developing climate action strategies, **in 2023 the RAISE project focused its work around one of its key recommendations: fostering international collaboration**. These activities aimed at supporting knowledge sharing between governments, industries, and key stakeholders on the potential of AI to mitigate climate change and support biodiversity preservation.

Figure 1. RAISE key recommendations including international collaboration


AI as a tool for climate action
Shaping AI's impact
Data & digital infrastructure
Data, simulation environments, testbeds, libraries, computational hardware
Research & innovation funding
Interdisciplinary & cross-sectoral work guided by climate impact
Deployment & systems integration
Policy design & evaluation, market design, business models
Reducing AI's negative impacts on the climate
Application and compute-related impacts
AI
Responsible AI
Capacity building
International collaboration
Impact assessment
Implementation, evaluation, and governance capabilities

Key Activities 2023

- ➔ Virtual workshop conducted in August 2023 convening experts from the broader GPAI, the RAISE Project Advisory Group, IEEE, invitees from IPCC, IPBES, and various organizations (academia, industry, international organizations, and NGOs)
    - Overall objective: to build on existing work at the intersection of AI and the environment, and find pathways to conduct impactful collaborations in the pursuit of tangible and practical projects
    - Challenges identified: Data availability, tracking, mapping, and transparency, and how to measure impact
    - Opportunities identified: integrating various knowledge systems, using GPAI's platform to provide a voice for local communities' data, identifying efficient public policies.

- ➔ Conference at the University of Geneva in September 2023 convening leading AI experts on climate action and sustainability to examine practical applications where international collaboration can be harnessed to accelerate climate action, as well as possible risks as AI technology diffuses and scales.
    - Overall objective: highlight best practices and key use cases to explore how AI systems can be deployed in the most sustainable way possible for the good of the planet.
    - Challenges identified: lack of funding, data accessibility, and communication between key stakeholders, as well as the need for long-term national planning
    - Opportunities identified: collecting data directly from the ground, tailoring public policies to regional needs and realities.

Outlook for 2024

In 2024, 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 workshops with GPAI Members, Experts, and the RAISE community, to drive deeper engagement and meaningful impact, and developing a structured cooperation program alongside IPCC and IPBES.

Responsible AI for Social Media Governance

Co-Leads

Alistair Knott
School of Engineering and
Computer Science,
Victoria University of
Wellington

Dino Pedreschi
Department of
Computer Science,
University of Pisa

Social media platforms are one of the main channels through which AI systems influence people's lives, and therefore influence countries and cultures. In 2022, the number of social media users worldwide was 4.59 billion;[1] the number is projected to be around 4.89 billion at the current moment, or 59% of the world's population. The average user spent over two and a half hours per day on social media in 2022, a figure which has been rising since 2012[2] and is projected to rise further.

Crucially, the experience of a social media user, on any given platform, is heavily influenced by AI systems that run on that platform. These platforms are largely powered by AI systems. **Recommender algorithms** learn about the interests of each user, and tailor users' content feeds accordingly. An array of **content classifiers** identify harmful content of various kinds (such as violent extremist content and hate speech), and help to moderate such content to keep platforms safe for users. This year, social media platforms have also become a medium for the dissemination of **AI-generated content**, produced by newly available tools for generating text (such as ChatGPT) and images (such as MidJourney).

Key Activities 2023

The Social Media Governance project has pursued three main work streams in 2023, one for each of the above AI influences[3].

- ➔ On **recommender algorithms**, our focus has been on finding effective ways of measuring the *effects* of these algorithms on platform users. Recommender algorithms are often optimized to keep users 'engaged' on the platform. There are well-versed concerns that this may also encourage them into areas of harmful content. The only way to properly assess these concerns is using methods currently only available within companies: namely, A/B tests that give different recommender systems to different randomly-selected groups of users, and study differences in the behavior of these groups. The EU's Digital Services Act will allow vetted external researchers to access company data and methods for the first time. We are working with the European Centre for Algorithmic Transparency, to advocate that access to A/B test data should be part of this data access.

- ➔ On **harmful content classifiers,** our project is trialling a mechanism for training classifiers outside of companies, in a semi-public domain that may offer a better model for their governance. Our first pilot project in this area is running in India, in the domain

[1] Statista (2023): Number of social media users worldwide from 2017 to 2027. https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/
[2] Statista (2023). Daily time spent on social networking by internet users worldwide from 2012 to 2023. https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/
[3] A summary of the activities and main work streams in 2023 is available here.

of political hate speech. We are running this pilot in collaboration with an academic research lab based at Jadavpur University, Kolkata, under the leadership of Prof Subhadip Basu.

➔ On **AI-generated content**, there have been many recent calls for mechanisms that allow *transparency* around such content—in particular, to identify AI-powered disinformation campaigns, which threaten to interfere with public opinion and democratic processes. Our project has argued that *AI systems for detecting AI-generated content* have an important role to play in these transparency mechanisms. But AI content detectors obviously need to be reliable in this role. We have argued the best way to achieve reliability is to require the companies that *build* AI content generators to *instrument their generators to support detection*. We have written two papers proposing that this requirement should be built into law.[4] Our proposal has already had some traction with policymakers. It was adopted by the EU Parliament, in its proposed amendments to the AI Act catering for 'foundation models'. Our proposal was also discussed at a recent US Senate Judiciary Committee Hearing on AI Oversight, at which two of our coauthors (Yoshua Bengio and Stuart Russell) gave evidence. Following this hearing, Joe Biden's recent Executive Order on AI has a useful focus on mechanisms for detecting AI-generated content. The UK's recent AI Safety Summit at Bletchley also focussed on generative AI tools. The declaration emerging from this summit stressed the potential of these tools to produce deceptive content and disinformation, which again highlights the importance of detection mechanisms.

Outlook for 2024

In 2024, we will pursue all three of the above workstreams. On recommender systems, we hope the EU's Digital Services Act will grant access to A/B test data for vetted researchers: if so, we will offer to act as vetted researchers, or to advise the researchers who are selected. On harmful content classifiers, we aim to scale up our pilot project in India, to consult a large group of content annotators sampled from the general public, and use the created dataset to train and evaluate proof-of-concept content moderation systems. On AI-generated content detection, we will be continuing advocacy efforts for our proposed detection mechanism rule. Our key focus will be on interacting with policymakers in the EU, US and UK, to pursue the technical and political issues that must be addressed in formulating a workable detection mechanism.

[4] GPAI (2023): State-of-the-art Foundation AI Models Should be Accompanied by Detection Mechanisms as a Condition of Public Release. GPAI report. Knott et al. (2023): Generative AI models should include detection mechanisms as a condition for public release. Ethics and Information Technology.

Scaling Responsible AI Solutions

Project Co-Leads

Francesca Rossi
IBM Fellow

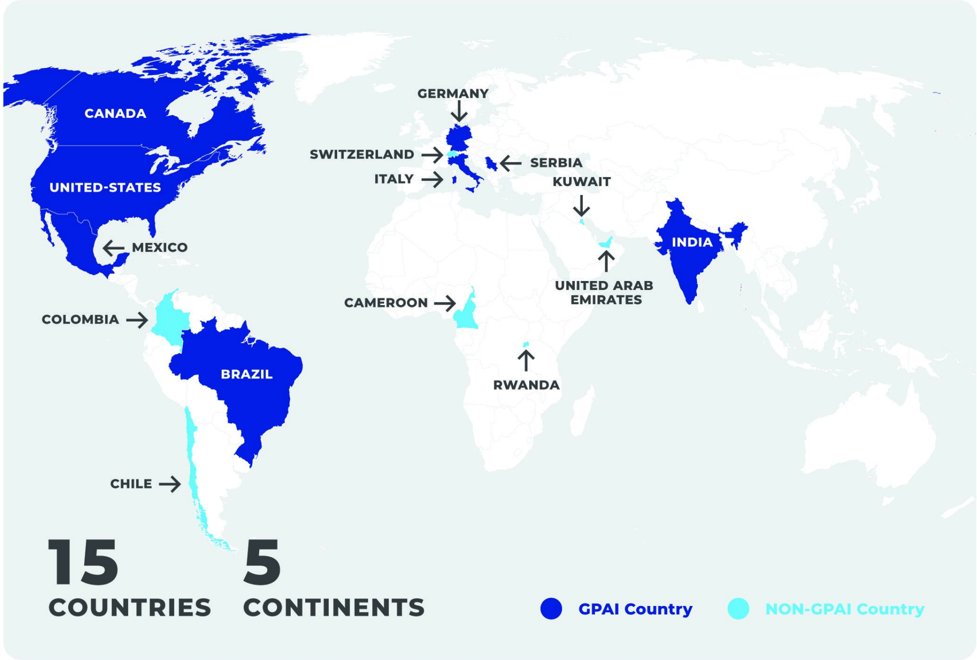
Amir Banifatemi
AI Commons / XPRIZ

Artificial intelligence (AI) systems that meet responsible standards and have positive socio-environmental and economic impacts should be supported to grow and to reach potential users and communities who could benefit from them. However, emerging AI projects have encountered challenges when it comes to practically implementing responsible AI (RAI) principles, as well as scaling. Frameworks for RAI have proliferated, but tend to remain quite high-level, without technical guidelines for implementation in various uses and contexts. At the same time, the process of scaling itself can introduce obstacles and complications to realizing or preserving RAI adherence.

The project Scaling Responsible AI Solutions (SRAIS) aims at encouraging and showcasing deployments of scalable RAI solutions. The objective **incentivizes, via public challenges, teams in various organizations to propose implementations of responsible AI solutions that would be practical, beneficial, and sustainably scalable**.

Key Activities 2023

From January to October 2023 the project set out to match teams working on RAI solutions with mentors of relevant expertise, and identify challenges that teams were facing with regards to both responsibility and scaling, and to assist in tackling these challenges[5]. In response to an initial call, 23 teams from 15 countries across five continents applied to take part in the project.



[5] See the 2023 report for further details on the work achieved here.

Five teams ultimately underwent the mentoring process, having been selected based on a range of criteria, assessing their potential contribution to the public good, their potential for institutionalizing RAI principles, and the specific scaling challenges they had already encountered. In addition, teams were selected to represent a range of country contexts (including both the Global North and the Global South) as well as a range of sectors in both the public and private spheres.

Selected Teams for 2023 Scaling RAI Solutions Program	
COMPREHENSIV – At Home Universal Primary Health Care	India
ergoCub**:** AI in Wearables and Robotics for Risk Assessment and Prevention	Italy
Jalisco's AI Forest Mapping System	Mexico
Advanced Chatbot and Voicebot Analytics Tools	Canada
Particip.ai One: Participation and Feedback Platform	Germany

The mentoring process took place across a series of initial meetings followed by three formal workshops which brought together each team and their mentors. The workshops were geared towards clearly identifying a key responsibility challenge to focus on, and developing a concrete output which took the shape or a RAI deep dive that laid out the team's response to this challenge. Teams and mentors thought together about how to ensure this output had broader relevance to other AI actors looking for practical ways to ensure they were meeting responsibility standards at different points in the scaling process. Following the production of the output (and at the time of writing this report) the teams were being supported by the mentors to implement the steps, plans, and indicators contained in their outputs. A formal evaluation committee made up of GPAI mentors has been established to monitor and report on the teams' progress.

While the participating projects in the 2023 SRAIS project were highly varied with respect to their aims and contexts, they faced very similar challenges in integrating and validating RAI principles, and scaling responsibly. Challenges that emerged for the participating teams included the establishment of robust and transparent data governance frameworks; stakeholder consultation, buy-in and the building of trust with users; safe and effective testing and experimentation; ensuring appropriateness and maintaining safety whilst scaling a solution across contexts (e.g. season, region or industry); adherence to an ethic of human-centredness (such that the application complemented the capabilities of people rather than replacing them); and user education on the appropriate use and limitations of specific AI applications.

Through the mentors' in-depth engagement with the participating projects, a series of recommendations emerged for both policymakers as well as other AI teams, to support and facilitate the responsible scaling of responsible AI. These recommendations focus on a range of areas including the facilitation of safe and accountable testing and experimentation; the need for equitable access to AI infrastructure including secure, representative datasets, as well as computational and connectivity infrastructure; the need to consider responsible AI not only at the beginning of an AI project or at the point of deployment, but throughout the lifecycle of AI—including consideration of how scaling may impact on responsibility; and the need to incentivise safety over speed, including allowing projects to fail if they are deemed to be unsafe.

Following the success of the first round of the project, the experts aim to repeat it, and expand further in subsequent years, to reach more teams in more countries, and to systematically analyze their experiences in order to produce a detailed and more technical ‘blueprint’ for scaling responsible AI solutions, that can be employed by a wide range of AI teams to benchmark their Responsible AI solutions.

Outlook for 2024

For 2024 SRAIS will continue to encourage and showcase deployments of scalable RAI solutions.

➔ Specific Objective 1: Raise awareness among governments and industry regarding the importance of Responsible AI scaling. The aim is to foster a deeper understanding of responsible AI practices, encouraging governments to incorporate them into their policies, regulations, and strategies.

➔ Specific Objective 2: Showcase how GPAI can serve as a valuable tool for governments to compare and synchronize their efforts with other countries. The project aims to demonstrate the collaborative potential of GPAI, promoting international cooperation, knowledge sharing, and exchange of best practices. By showcasing successful case studies and highlighting the benefits of collaboration, the project seeks to encourage governments to actively engage with GPAI as a platform for collective action.

➔ Specific Objective 3: Foster the funding of promising and validated Responsible AI Solutions. The project aims to create an ecosystem that connects investors, funding agencies, and organizations working on responsible AI solutions. By identifying and validating impactful AI projects, the project seeks to attract financial support and investment opportunities for the development and scaling of responsible AI solutions. This objective aims to bridge the gap between innovative ideas and the resources required for their implementation, promoting the growth and sustainability of responsible AI initiatives

Towards Real Diversity and Gender Equality in AI

Project Co-Leads

Paola Ricaurte Quijano
Tecnológico de Monterrey

Benjamin Prud'homme
MILA

AI offers a wide range of possibilities for enhancing the well-being of different groups and contributing to the UN Sustainable Development Goals. However, AI can also deepen economic, knowledge, gender and cultural divides. Still today, AI is generally designed, developed, monitored, and evaluated *without* systematic gender and diversity approaches, leading to negative consequences and becoming an obstacle to the development and adoption of Responsible AI.

This has resulted, for instance, in a) disproportionate harm caused by AI against women and marginalized groups, b) missed opportunities for AI to be more impactful, due to lack of consideration of these groups, and c) lack of inclusion in discussions on AI ethics, regulation and strategies.

The project objective is to contribute to ensuring that the AI ecosystem, and in particular States in the role of duty bearers, have the tools and commitment to incorporate effective diversity and gender equality (DGE) approaches throughout the AI lifecycle, and to demonstrate their impact and results with indicators in accordance with international standards[6]. This responds to one of GPAI's priorities, which is to understand the impact of AI on human rights.

Key Activities 2023

➔ **Literature Review**: Aiming to identify key themes, theories, methodologies, and gaps in the literature on the topic of integrating DGE into the AI life cycle. This review covers existing practices and critiques of the current discourse around, and practices related to, DGE in AI.

➔ **Regional Consultations**: Engaging localized expertise from five regions worldwide: Latin America and the Caribbean, Sub-Saharan Africa, Middle East and North Africa, North America and Europe, and Asia and the Pacific. Stakeholders included representatives from academia, civil society, industry, and government. A participatory qualitative data collection methodology was developed based on an interview guide and an iterative approach to explore emerging themes in collaboration with delivery partners. Consultations are ongoing.

➔ **Community Perspectives:** Outreach conducted by sharing a first version of the report with civil society organizations and persons self-identifying as members of marginalized

[6] See the 2023 advancement report here.

groups for additional feedback. These individuals and organizations also responded to a set of questions on the integration of DGE in AI.

- ➔ **Promising Practices and Resources**: A selection of existing initiatives or solutions aiming to integrate DGE in AI, and a more elaborate analysis through use cases. Such initiatives include technical, capacity building, policy, and community engagement.
- ➔ **Environmental Scan**: A mapping of existing initiatives that aim to include DGE in the AI life cycle, complete with labeling and annotation.

Outlook for 2024

Building on the foundational work completed in 2023, the next phase of the project's work proposes to concretise the findings from the first phase into a final report, deliver a practical and accessible DGE toolkit, and offer confidential advice to GPAI leadership as to concrete ways in which the organization could foster greater diversity. Further details are provided below:

- ➔ Developing and deploying a DGE toolkit, consisting of a curated, annotated repository of the most efficient tools identified and including explanations as to the advantages and limitations of each tool.
- ➔ Conducting a review and analysis of current GPAI practices to formulate evidence-based recommendations as to actionable ways in which the organization can foster greater DGE. The report will be addressed to GPAI leadership only and remain confidential.

Sandbox for Responsible AI

Project Co-Leads

Juan David Gutiérrez
Associate Professor
Universidad de los Andes

Aditya Mohan
National Standards
Authority of Ireland

Sandboxes and other experimental governance approaches allow organizations to test and validate t new technologies transparently and in a controlled environment.. They allow both public and private organisations to assess their services, products and procurement processes and ensure compliance with (new) regulatory frameworks and/or standards. Sandboxes may also help regulators identify possible challenges to new regulation, such as those emerging around AI and autonomous systems. This promotes the development, procurement and deployment of innovative artificial intelligence solutions that are both ethical and responsible. Despite the wide interest in sandboxes applied to AI, there are limited examples of how these are to be designed, implemented, governed and used with regards to procurement processes.

Key activities 2023

➔ **Engagement with stakeholders.** Meetings with governments, civil society organizations, academia, international organizations, professional organizations, and companies to explore the concept of "sandboxes" and other testing environments for new technologies.

- Insights: key characteristics for AI sandboxes in public procurement processes:

  (1) Experimental: to experiment, iterate and test novel technologies, business models and public innovation strategies early on before they fully fledged and enacted.

  (2) Agile: to quickly adapt to the ever changing landscape.

  (3) Anticipatory: anticipates the impact in a controlled environment.

  (4) Collaborative and stakeholder-inclusive: allows co-creation, feedback and learnings.

➔ Brussels FARI AI Institute for the Common Good conference in September 2023 convening leading experts in procurement of AI solutions to examine existing guidelines and existing use cases.

- Overall objective: highlight best practices and key use cases to explore how procurement for AI solutions can be safely deployed for government uses.
- Outcomes: international collaboration can be harnessed to accelerate due diligence in AI procurement solutions, it could also be an opportunity to identify possible risks as AI technology diffuses and scales.

➔ **Literature Review**. A research paper on AI procurement in public administration around the globe. The paper synthesizes existing resources and provides an analysis of trends, gaps, and opportunities for growth. The recommendations are divided into two groups: activities to do when procuring a particular AI system, and general activities, i.e. activities that would help to prepare for AI procurement, not necessarily in the context of a particular procurement process. The paper covers the complete life cycle of the procurement process, from setup to deployment and will offer guidance for practitioners that aim to follow process-oriented best practices.

Pandemic Resilience

Project Co-Lead

Michael O'Sullivan
Associate Professor
Faculty of Engineering
University of Auckland

This project reflects the global emergency presented by COVID-19 and the need for international and multi stakeholder collaboration for an efficient and timely response to global threats. Its overarching goal is to directly support impactful and practical AI initiatives to help in the fight against the COVID-19 pandemic. The project has produced two outputs in its first phase: (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 pandemics. In 2023 the group leveraged the living repository to gather some of the most promising initiatives to develop an AI-calibrated ensemble of pandemic spread prediction models. The outcome demonstrates a high potential for better model predictions in the face of uncertainty and provides recommendations for decision makers to strengthen their evidence-based decision making.

Key Activities for 2023

For 2023 the Pandemic Resilience project explored the use of ensemble modeling of infectious diseases to enable better data-driven decisions and policies related to public health threats in the face of uncertainty. The aim for this year’s phase was to demonstrate how Artificial Intelligence (AI)-driven techniques can automatically calibrate ensemble models consistently across multiple locations and models. The ensembling, calibration, and evidence-generation reported here was conducted by a diverse and interdisciplinary team. This diverse team co-developed and tested a collaborative ensemble model that assesses the level of use of Non-Pharmaceutical Interventions (NPIs) and predicts the consequent effect on both epidemic spread and economic indicators within specified locations. The disease of interest for this exercise was COVID-19 and its variants.

After a series of initial workshops with the selected initiatives to co-develop the project’s methodology, ,the concrete development of the ensemble model was undertaken in five main phases from January 2023 to October 2023:

1. Definition of a standardized set of inputs and outputs;
2. Adaptation of individual models to the standard;
3. Development of a calibration framework for the ensemble;
4. Deployment and testing of the ensemble across different different locations;

5. Automated calibration of the ensemble using a Genetic Algorithm (GA) metaheuristic optimization approach.

Having constructed and tested the ensemble, the study team has prepared the following report. to share key findings about the use of such ensemble models and communicate key recommendations for governments and policymakers about the future development and support of ensemble models and AI-based calibration.

Outlook for 2024

In 2024, the group will concentrate on advocating for the benefits of AI-calibrated ensemble modeling in data-driven decision making and promoting responsible AI practices to policymakers and public health officials. The envisioned future work includes exploring the creation of What-if simulations, progressing from model predictions to model prescriptions, and extending applications to other diseases. The group plans to disseminate its work at international health events, continuously experiment with the ensemble of models to generate more results, and publish findings in a scientific journal article.

Concurrently, the group is considering gauging interest and appetite for developing the AI-calibrated ensemble model further from a research prototype to a beta version pandemic-ready tool that allows potential users to test the model. This involves adding this extra development into the work plan for 2025 with the objective of enabling people to use the developed tool. One key step in this development would be to organize and run a tabletop exercise. This tabletop exercise would involve putting the model into the hands of real users for a real-life test drive, replaying pandemic scenarios and simulating decisions in a controlled environment. The group envisages potential real-life users from different jurisdictions and country members gathering together for the table top exercise, thus ensuring diverse representation of public health decision makers when testing and amending requirements for the pandemic-ready tool.

Building on this foundational work carried out over the last three years, the Working Group proposes to move forward with a new proposal called *Digital Ecosystems that Empower Communities.* This project is based on the ideas of empowering communities with digital technology and that communities who provide data for AI systems should benefit from this data. By providing templates, platforms and case studies of digital ecosystems within communities this project will streamline the adoption of digital ecosystems and the associated benefits by other communities. These community-based digital ecosystems then enable data from one community to be shared with other communities, if the contributing community wishes, so that data and/or AI models can be federated across communities to realise additional regional, national and global benefits. Hence, communities realize the benefits of digital ecosystems, retain sovereignty of their data and can contribute to larger, federated digital ecosystems with corresponding extended benefits.This project is inspired by Te Ao Māori | the Māori worldview and is grounded with the concept of digital ecosystems based on place, i.e., rooted within communities

Forward Look

For 2024, the Working Group aims to continue four of its current projects and start two new ones. The Working Group has proposed the following projects for 2024 subject to GPAI Council approval at the New Delhi Summit:

Responsible AI Working Group 2024 Projects		
RAI #1	Repositories of Public Algorithms (joint project with Data Governance)	New
RAI #2	Social Media Governance	Continuing
RAI #3	Responsible AI Strategy for the Environment (RAISE)	Continuing
RAI #4	Creating Systemic Gender Inclusion in AI Ecosystems	Continuing
RAI #5	Scaling RAI Solutions	Continuing
RAI #6	Digital Ecosystems that Empower Communities	New

We're looking forward to starting 2024 with these upcoming projects in the pipeline. We're hopeful that the next months will be productive and that 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.

Participation across our Working Group is a big part of what makes these projects true international collaborations. We would like to invite those who are interested to make personal contribution to these projects by joining our Project Advisory Groups to help shape direction, give feedback, and review research. You can express your interest to contribute by connecting with the Montreal Centre of Expertise (the CEIMIA) at info@ceimia.org.