Responsible Development, Use and Governance of AI Working Group Report

November 2020 - GPAI Montréal Summit

GPAI | THE GLOBAL PARTNERSHIP ON ARTIFICIAL INTELLIGENCE

Please note that this report was developed by experts of the Global Partnership on Artificial Intelligence's Working Group on the *Responsible Development, Use and Governance of AI. The report reflects the personal opinions of GPAI experts and does not necessarily reflect the views of the experts' organizations, GPAI, the OECD or their respective members.*

Co-Chair's Welcome

Yoshua Bengio
Founder and Scientific Director of Mila (Quebec Artificial Intelligence Institute*)*

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

The Global Partnership on AI (GPAI) was founded in 2020 to undertake and support applied AI projects and provide a mechanism for sharing multidisciplinary analysis, foresight and coordination—with the objective of facilitating international collaboration and synergies and reducing duplication in the area of AI systems governance.

We co-chair one of GPAI's four expert working groups, the Responsible Development, Use and Governance of AI Working Group (from now on, "RAI").

RAI's mandate is to foster and contribute to the responsible development, use, and governance of human-centered AI systems, in congruence with the UN Sustainable Development Goals. RAI's work is grounded in a vision of AI that is fair and respectful of human rights and democracy. Its aim is to be equitable and inclusive, and to contribute positively to the public good.

We recognize that partnerships will be essential in fulfilling our mission and maximising the impact of our work. Partnerships present the opportunity to co-create solutions to the challenges we've identified by domain specialists in a given field and the AI and technical expertise of GPAI. Moreover, they present an opportunity for coordinated, structured solutions that combine the funding, knowledge and implementation necessary to the scale of the challenges we will seek to address with our proposed committees on, for example, Climate Change, Education and Drug Discovery & Open Science. We look forward to exploring those possibilities further in the next phase of our work, and recognize too that there will be significant benefit in collaborating with our colleagues in other Working Groups on cross-cutting issues.

This report presents the work that RAI has done in the last six months, and its goals for the next semester as well as for the next 2-3 years. It is prefaced and presented by us, but it is really the product of the efforts made by RAI's 30 members and by the 9 members of RAI's Steering Committee, to whom we are extremely grateful.

As if often repeated, rightly, AI has the potential to foster the progress of our societies in many different fields, but it could also impact individuals and whole communities negatively if it is developed and used in careless or devious ways.

We thus hope that RAI will contribute to the development and implementation of international coordination mechanisms and tools that will help us, collectively, to use AI systems as a lever for meeting UN Sustainable Development Goals, as well as contribute to the identification of governance mechanisms and tools that will enable us to minimize the negative impacts of AI.

Introducing RAI

RAI has 30 members. Its international experts come from various fields, something which favors robust discussions and the emergence of diverse viewpoints. More precisely, 14 members of RAI come from the technical world (e.g. machine learning, information technologies), whereas 16 come from the social and human sciences sector and fields like communications, anthropology, literature, management, history, psychology, philosophy, international affairs, international development, journalism, economics, and political science.

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

Most members (63%) come from the academic sector, but 17% work in the private sector, 13% for non-profits and 7% in the Public Sector. A better balance should be achieved in 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.

As of November 30, members were based in 17 countries, that is Argentina, Australia, Canada (2 people), France (3), Germany (2), India (2), Italy, Japan (2), Korea, Mexico (2), the Netherlands, New Zealand (2), Singapore, Slovenia (2), Sweden, the United Kingdom (3), and the USA (3).

These members have been designated by the 15 founding members of GPAI or recommended by UNESCO. It's worth mentioning that members are designated by GPAI's member countries or recommended by international institutions, but act with full independence inside RAI.

Finally, 7 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 a panel of experts that advises the OECD.

Box 1, below, presents RAI's experts

RAI’s Members

Members of RAI (in bold, members of RAI's Steering Committee)

Yoshua Bengio (Co-Chair), Mila - Quebec Artificial Intelligence Institute
Raja Chatila (Co-Chair), Sorbonne University
Carolina Aguerre, Center for Technology and Society (CETyS)
Genevieve Bell, Australian National University
Ivan Bratko, University of Ljubljana
Joanna Bryson, Hertie School
Partha Pratim Chakrabarti, Indian Institute of Technology Kharagpur
Jack Clark, OpenAI
Virginia Dignum, Umeå University
Dyan Gibbens, Trumbull Unmanned
Kate Hannah, Te Pūnaha Matatini, University of Auckland
Toshiya Jitsuzumi, Chuo University
Alistair Knott, University of Otago
Pushmeet Kohli, DeepMind
Marta Kwiatkowska, Oxford University
Christian Lemaître Léon, Metropolitan Autonomous University
Vincent C. Müller, Technical University of Eindhoven
Wanda Muñoz, SEHLAC Mexico
Alice H. Oh, KAIST School of Computing
Luka Omladič, Institute of Applied Ethics
Julie Owono, Internet Sans Frontières
Dino Pedreschi, University of Pisa
V K Rajah, Advisory Council on the Ethical Use of Artificial Intelligence and Data (Singapore)
Catherine Régis, Université de Montréal
Francesca Rossi, IBM Research
David Sadek, Thales Group
Rajeev Sangal, International Institute of Information Technology Hyderabad
Matthias Spielkamp, Algorithm Watch
Osamu Sudo, Chuo University
Roger Taylor, Centre for Data Ethics and Innovation

Observers

Amir Banifatemi, AI Commons
Vilas Dhar, The Patrick J. McGovern Foundation
Marc-Antoine Dilhac, ALGORA Lab
Adam Murray, OECD Network of Experts on AI
Karine Perset, OECD
Stuart Russell, UC Berkeley
Cédric Wachholz, Digital Innovation and Transformation Section, Communication and Information Sector at the UNESCO

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, it intends to collaborate 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 in the light of the current international context, the GPAI Task Force has invited RAI to form an *ad hoc* AI and Pandemic Response Subgroup to support the responsible development and use of AI-enabled solutions to COVID-19 and other future pandemics, which will report to the RDUGAI. The AI and Pandemic Response Subgroup was created in July 2020. It is co-chaired by a member of RAI, Alice Oh, and by a non-member, Paul Suetens.

Work Process

RAI was created less than six months ago, in a, needless to say, quite challenging period.

It has held 5 meetings since it began its activities. These meetings were used to discuss the work that the group should undertake and the work that it had initiated.

On July 31 of 2020, members of RAI met virtually to discuss what the new RAI's mandate should be and to brainstorm about what this new body should try to accomplish in the short term.

On August 25 of 2020, members framed the first project they intended to undertake as a group. Wishing to identify gaps that RAI could address over the coming months without duplicating the different national and international initiatives aimed at fostering and ensuring the responsible development and use of AI systems, and to inform our work on the use made of AI for achieving the UN Sustainable Development Goals until now, RAI's members decided to carry out a review of national and international initiatives aimed at:

- cataloguing projects led by various stakeholders to promote the responsible research and development of beneficial AI systems and applications;

- analyzing promising initiatives that had great potential to contribute to the development and use of beneficial AI systems and applications that could benefit from international and cross-sectoral collaboration;

- and recommending new initiatives and how they could, in practice, be implemented and contribute to promote the responsible development, use and governance of human-centered AI systems.

In September 2020, RAI launched a public call for proposals to identify a consultant who could help RAI to complete that study before the end of November, in time for the December Summit. It also set up a Steering Committee comprising 9 volunteers from RAI (listed under Annex 1) to evaluate the proposals received by RAI and to supervise the work done by the winning firm or group.

Nine proposals were received by RAI before the September 20 deadline. On September 28, after two evaluation rounds, the Steering Committee selected The Future Society (TFS) as the firm that would lead RAI's first project under the Steering Committee's supervision.

TFS met with members of the Steering Committee during a kick-off meeting that took place on October 2. That meeting served to develop a common understanding of the objectives of this project, define the final work plan for the project and plan future meetings.

On October 5 of 2020, TFS presented the project's scope and timeline to members of RAI, who had the opportunity to make suggestions. This meeting was also an opportunity for the Co-Chairs to explore the possibility of creating new internal work committees that could, in the short term, tackle specific questions of high interest.

Meetings were held by the Steering Committee and TFS on October 12 and 22 to discuss preliminary versions of the report and of its different components, and to plan the intervention of TFS during meetings of the larger group.

On October 27 of 2020, TFS presented the draft version of its report to RAI. In the following days, 3 work sessions were organized by TFS to give subsets of RAI the opportunity to obtain more information on TFS's preliminary findings and recommendations and to discuss them.

The Steering Committee met two other times, on November 2 and 12, to debate the final recommendations that could be included in TFS report.

On November 8, the Co-Chairs sent a message to RAI's members to invite them to signal their interest in joining one of 5 new internal committees that RAI could possibly set up in the coming weeks. Following initial discussions and the preliminary recommendations contained in TFS's report, it was determined that these committees could work on Drug Discovery & Open Science, Governance and Transparency of Social Media, Climate Change, Education and Issues about and Means of Governance.

TFS's quasi final report was presented to RAI on November 16 and submitted on November 23. This report will be presented during the Summit. Moreover, its recommendations were used as an input to the 10-15 pages report you're reading at the moment.

Working Group Timeline

JULY

Co-Chairs' introduced (7th)

First Working Group meeting (31st): introductions and agreement on mandate; The Co-Chairs invited the members of RAI to inform them if there were particular points they would like to bring to the table.

AUGUST

Second meeting of the Working Group (25th) – discussion around the first deliverable that would be produced by RAI

SEPTEMBER

Introductory blog on RAI is published on the OECD website, including a call for proposals and terms of reference

Round 1 of the evaluation of the proposals received by RAI (24th)

Round 2 of the evaluation of the proposals received by RAI and selection of TFS (28th)

OCTOBER

Project kickoff meeting with TFS and Steering Committee (2nd)

Third meeting of the Working Group (5th) – presentation of TFS, its mandate and its work plan; discussion around the possible creation of internal work committees

Meeting of the Steering Committee and TFS (12th)

Meeting between all Co-Chairs to compare progress and discuss potential synergies (October 16th)

Meeting of the Steering Committee and TFS (22nd)

Fourth meeting of the Working Group - (27th) – presentation of TFS's first draft of its report and recommendations

NOVEMBER

Work sessions led by TFS (4th-5th)—RAI's members have the opportunity to discuss TFS's report during three work sessions organized in three different time zones

Fifth meeting of the Working Group (16th) - presentation of the last draft of TFS's report); discussion on the interest generated by the possible creation of 5 committees

TFS's report is submitted to CEIMIA (23rd)

Meeting between all Co-Chairs prior to the Summit (27th)

DECEMBER

Presentation of finalized outputs and open workshop on next projects at the Summit.

Preliminary Recommendations and Outputs for the Summit

The review of national and international initiatives conducted by TFS is the first deliverable commissioned by RAI. That review contains valuable insights and recommendations that could guide the work that RAI will do in the coming months and years.

Elaborated by TFS, with the support of RAI, in October and November of 2020, the report presents 30 promising initiatives that contribute to fostering the development or use of responsible AI or to reach the UN Sustainable Development Goals. These initiatives were then analyzed by TFS on their potential to help GPAI deliver on its objectives; on their effectiveness and alignment with the OECD AI Principles and the UN SDG agenda; on their scalability across geographies and sectors; and on whether they are as representative as possible across geographies, sectors, stakeholders and target groups. The last section of TFS's report draws on the opportunities and gaps identified by TFS to propose 4 areas for future action and 9 recommendations that will help inform RAI's agenda going forward.

In our opinion, 3 of these areas and their associated recommendations are especially pertinent:

1. AI has wide-ranging applicability and hence has the potential to influence many of the most pressing issues humanity is facing; it can be a force for good to mitigate climate change or predict the next pandemic, and it can also exacerbate global challenges as evidenced by the rise of misinformation. The breadth of potential applications of Responsible AI creates a prioritization challenge. Thus, RAI should create focused committees to address identified pressing issues.

2. To build an ecosystem with the ability to support and stimulate change, there is a need for governance tools and frameworks that promote transparency and alter incentives and behaviors throughout society to help the adoption of Responsible AI practices. There is also a need for systematic collaboration and cooperation across the ecosystem as well as a mechanism to connect cross-cutting initiatives on the domain level. Finally, for governments to implement these tools and frameworks at scale, there is a need to build capacity amongst policymakers as well as feedback loops between governments and other actors in the ecosystem. Thus, RAI should create a focused committee on governance issues and governance means.

3. Many initiatives in the Responsible AI ecosystem have struggled to collect representative input to inform their activities. This lack of inclusiveness points to a lack of capacity by initiatives, stakeholders and governments to involve a wider group in the technological transition and, hence, to co-shape innovative solutions for addressing the opportunities and the risks. Ultimately, this lack of inclusiveness risks undermining the effectiveness and credibility of many Responsible AI initiatives as well as their ability to scale. Thus RAI should develop and disseminate good Diversity & Inclusion practices.

RAI’s Focus for the Next 3-6 Months

The work conducted by TFS for RAI was extremely useful, and we now aim to identify international efforts which go beyond the work already done and catalogued in TFS’s report.

More precisely, we wish to focus on areas where current market structure and current government policies are insufficient to achieve the goals of the UN on Sustainable Development (SDGs) or other key objectives.

To achieve this, we have decided to focus on a few concrete themes of action connected to SDGs that RAI’s members are interested in and are willing to explore inside internal work committees :

1. The Committee on Drug Discovery & Open Science (which is linked to SDG 3: Good health and well-being) could examine how to create a favorable context for AI to contribute to drug discovery in an open and equitable manner, whereby international public health needs are privileged, e.g. for fighting covid-19 or antibiotics resistance. It could also see how R&D efforts should be organized and what the rules of engagement should be to ensure licensing of resulting drugs is favorable to poor and contributing countries;

2. The Committee on Climate Change (SDG 13: Climate action) could elaborate practical collaborative approaches to fight climate change using AI (e.g. to ensure that AI is making zero-carbon renewables as productive as traditional hydrocarbon suppliers) on the one hand, and to do AI in a more environmentally friendly way, on the other (e.g. to better evaluate machine learning’s environmental impact);

3. The Committee on AI and Education (SDG 4: Quality Education) could define collaborative projects whose implementation would contribute to (a) maximizing the benefits of AI for education management and delivery, empowering teaching and teachers, improving learning and learning assessment, offering lifelong learning opportunities for all, etc.; (b) addressing crosscutting issues like promoting the equitable and inclusive use of AI in education, training students so they become responsible producers and users of AI or monitoring the impacts of AI on education.

4. The Committee on Governance and Transparency of Social Media (SDG 16: Peace, Justice and Strong Institutions) could focus on elaborating principles and tools to draw the line between what is acceptable or not in terms of manipulation in advertising and social media. This will subsequently help undertake joint research initiatives towards the development of AI tools to achieve objectives like detecting fake news or bias, or flagging demagogic messaging.

We believe that in order to advance the public good in each of these areas, governance issues need to be considered which are fundamentally tied to how the proposed progress could be usefully deployed. For example, internationally funded efforts in the use of AI for antibiotics resistance drug discovery will require defining rules of engagement with private or public organizations doing the actual R&D, e.g., regarding data sharing, standards of information sharing, and licensing favorable to the public good (protecting the whole planet from the next pandemic, for example).

For this reason we have also decided to create a fifth committee, a transversal one, that will be devoted to the issue of governance in the development of public-good oriented use of AI. Beyond the focused governance issues presented above, the Committee on Issues about and Means of Governance could work on the certification, assessment, and audit mechanisms used to evaluate AI systems for responsibility and trustworthiness based on metrics such as accountability, transparency, safety, fairness, respect for human rights, and the promotion of equity.

Long-Term Vision

Other transversal topics will, in the longer term, color the work of the specific committees identified above.

One of these topics is the issue of inclusion and diversity, with a particular focus on the effects (positive or negative) that AI can have on the most vulnerable individuals and groups among us.

As we know, some people, groups or countries may end up not seeing the gains made possible by AI because they will not have access to some of the most promising tools that will be developed (e.g. only rich countries or schools may be able to use the latest AI technologies for supporting students). It is also clear that because of the way they are designed and developed, some AI tools could lead to the exclusion of people or groups that are often already fragile, to the reproduction or even reinforcement of prejudices that already exist in our societies.

The Diversity and Inclusion issue is one that RAI will address in the medium and long term. As suggested in TFS's report, RAI could help shape and spread good Diversity and Inclusion practices. This could include a strategy that helps gauge the extent to which segments of society or geographies are currently underrepresented or excluded in the Responsible AI ecosystem. Steps could also be taken to encourage open-access information and infrastructures that would help break down communications barriers between geographies, social groups, and disciplines. Processes will also have to be devised for ensuring the deployment of applications that will benefit marginalized groups and societies, that developers are drawn from more diverse backgrounds and comply with codes of conduct or to develop mechanisms whose implementation will help reduce or eliminate the negative impacts algorithmic decision can have on some of us.

Beyond the work of the 5 committees that will be launched very shortly, we also envision other topics of investigation which RAI may consider in the future, possibly spinning off the current committees or creating completely new ones which focus on other SDGs, such as peace or hunger, which could be launched in mid-2021, but also general mechanisms for international funding and governance of publicly funded AI for social good projects.

Work on governance mechanisms and processes will continue in parallel. In particular, RAI will make sure it coordinates its activities closely with those of the OECD and ONE AI's Working Group on Trustworthy AI (ONE TAI). OECD describes ONE TAI as a group that "is helping to identify promising ideas and good practices for implementing the five values-based OECD AI principles for trustworthy AI systems: 1) inclusive growth, sustainable development and well-being; 2) human rights, democratic values and fairness; 3) transparency and explainability; 4) robustness, security and safety; and 5) accountability. These practices include codes of conduct, guidelines, standards, certifications, corporate governance frameworks, risk management approaches, technical research, software tools, as well as capacity and awareness building tools. The goal is to identify practical guidance and shared procedural approaches to help AI actors and decision-makers to implement trustworthy AI, including highlighting how tools and approaches may vary across different operational contexts."

Finally, following a recommendation made by TFS in its report, RAI will work with international organizations like OECD, WHO and UNESCO to ensure it collects representative input from marginalized groups and the Global South. The role of RAI in these partnerships will be to proactively bring historically marginalized groups into these dialogues and to support initiatives that foster basic AI literacy so the public can be empowered to participate.

Annex1

Project Steering Committee Membership

Raja Chatila (Co-Chair), Sorbonne University
Dyan Gibbens, Trumbull Unmanned
Kate Hannah, Te Pūnaha Matatini, University of Auckland
Toshiya Jitsuzumi, Chuo University
Vincent C. Müller, Technical University of Eindhoven
Wanda Muñoz, SEHLAC Mexico
Dino Pedreschi, University of Pisa
Catherine Régis, Université de Montréal
Francesca Rossi, IBM Research