Scaling Responsible AI Solutions

Challenges and Opportunities

December 2023

GPAI | THE GLOBAL PARTNERSHIP ON ARTIFICIAL INTELLIGENCE

This report was developed by Experts and Specialists involved in the Global Partnership on Artificial Intelligence's project on 'Scaling Responsible AI Solutions'. The report reflects the personal opinions of the GPAI Experts and External Experts involved and does not necessarily reflect the views of the Experts' organisations, GPAI, or 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.

Acknowledgements

This report was developed in the context of the 'Scaling Responsible AI Solutions' project, with the steering of the Project Co-Leads and the guidance of the Project Advisory Group, supported by the GPAI Responsible AI Working Group. The GPAI Responsible AI Working Group agreed to declassify this report and make it publicly available.

Co-Leads:
Amir Banifatemi*, AI Commons
Francesca Rossi*, IBM Research

The report was prepared by: **Kelle Howson‡**, Independent consultant.

GPAI recognizes the dedication of expert mentors who played a pivotal role through the whole project: **Francesca Rossi***, IBM Research; **Amir Banifatemi***, AI Commons; **Venkataraman Sundareswaran***, World Economic Forum; **Anurag Agrawal***, Ashoka University; **Daniele Pucci***, Istituto Italiano di Tecnologia; **Borys Stokalski***, RETHINK; **Ulises Cortés***, Universitat Politècnica de Catalunya; **Przemyslaw Biecek***, Warsaw University of Technology; **Zumrut Muftuoglu***, Digital Transformation Office of the Presidency of the Republic of Türkiye; **Furukawa Naohiro***, ABEJA, Inc.; **Ricardo Baeza-Yates***, Institute for Experiential AI of Northeastern University, **Maria Lorena Florez Rojas†**, University of Groningen; **Kolja Verhage†**, Deloitte.

GPAI also recognizes the meaningful contribution of experts who supported the project in its different phases: **Michael O'Sullivan***, University of Auckland; **Juan David Gutiérrez***, Universidad del Rosario; **Kudakwashe Dandajena***, African Institute for Mathematical Sciences (AIMS) and University of the Western Cape; **Shameek Kundu***, Truera; **Andrea A. Jacob***, Code Caribbean.

GPAI is thankful for the dedicated work of the AI Solution teams who applied to and completed the Scaling Responsible AI Solutions mentorship program and produced meaningful outputs for this report: **Aaqib Azeem**, from Wysdom.ai; **Suresh Munuswamy**, from COMPREHENSIV; **Eduardo Ulises Moya-Sanchez**, **Abraham Sánchez U.Moya**, **Raul Nanclares Da Veiga**, **Alexander Quevedo Charon**, **T. Camacho**, **Ulises Jiménez Pelagio**, **A. Piña Gobierno** from Jalisco's AI Forest Mapping System; **Jascha Stein**, **Ronald Strauss** from Particip.ai One; and for the ergoCub team, **Daniele Pucci**, **Lorenzo Rapetti**, **Enrico Valli**, from the Istituto Italiano di Tecnolgia (IIT), and **Cristina Di Tecco**, **Matteo Ronchetti**, **Francesco Draicchio and Giovanna Tranfo** from the Istituto nazionale Assicurazione Infortuni sul Lavoro (INAIL).

GPAI would also like to thank the following individuals who provided valuable support to enrich the report: **Paola Ricaurte Quijano***, Tecnológico de Monterrey; **Stuart J Russell***, UC Berkeley; **Seydina Moussa Ndiaye***, Open Training for the Reinforcement of Competences for Employment and Entrepreneurship in the Digital Sector (FORCE-N); **Celine Caira****, OECD AI; **Przemyslaw Biecek***, Warsaw University of Technology; **Zumrut**

Muftuoglu*, Digital Transformation Office of the Presidency of the Republic of Türkiye; **Venkataraman Sundareswaran***, World Economic Forum; **Daniele Pucci***, Istituto Italiano di Tecnologia.

GPAI would like to acknowledge the tireless efforts of the colleagues at the International Centre of Expertise in Montréal on Artificial Intelligence (CEIMIA) and of the GPAI Responsible AI Working Group. We are grateful, in particular, for the support of **Arnaud Quenneville-Langis**, **Laëtitia Vu,** and **Stephanie King** from CEIMIA, and for the dedication of the Working Group Co-Chairs **Raja Chatila***, Sorbonne University and **Catherine Régis***, Université de Montréal.

* Expert
** Observer
† Invited Specialist
‡ Contracted Parties by the CofEs to contribute to projects

Citation

GPAI 2023. Scaling Responsible AI Solutions: Learning from AI teams to identify and address challenges in Responsible AI, Report, December 2023, Global Partnership on AI.

Table of Contents

Executive Summary

Artificial intelligence (AI) seems to present solutions to many challenges across different domains. However, there is now a widespread understanding of the range of potential risks and harms to people and the planet that AI can produce if conceived, designed, and governed in irresponsible ways. In response to this, many proposals, frameworks and laws have been advanced for the responsible development and use of AI systems. In tandem, more and more AI 'solutions' are emerging around the world, which attempt to contribute to the public good, whilst upholding best-practice standards of responsibility.

It is important that AI systems that meet responsible AI (RAI) best practices and have positive socio-environmental impacts are supported to grow and reach potential users and communities who could benefit from them. However, nascent AI projects have encountered challenges when it comes to practically implementing RAI principles, as well as scaling. Key RAI challenges include mitigating bias and discrimination, ensuring representativeness and contextual appropriateness, transparency and explainability of processes and outcomes, upholding human rights, and ensuring that AI does not reproduce or exacerbate inequities. Frameworks for RAI have proliferated, but tend to remain at a 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 realising or preserving RAI adherence.

From January to October 2023 the Global Partnership on Artificial Intelligence (GPAI)'s Working Group on Responsible AI undertook a project in response to these challenges. The project, called Scaling Responsible AI Solutions (SRAIS), set out to match teams working on RAI solutions with mentors with relevant expertise, in order to identify challenges teams were facing with both responsibility and scaling, and assist in tackling these challenges. In response to an initial call for participation, 23 teams from 14 countries on 5 continents applied to take part in the project. The project was guided in particular by OECD's Recommendation on Artificial Intelligence (2019).

Five teams ultimately went through the mentoring process, having been selected based on a range of criteria, including their potential contribution to the public good, their potential for institutionalising 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.

The mentoring process took place across a series of meetings and workshops. There were two initial workshops involving all the mentors and all the teams, to share experiences and gain a sense of the general challenges teams were facing. After that, each team had three one-on-one sessions with their assigned mentors. These sessions were geared towards clearly identifying a key responsibility challenge to focus on and developing a customised RAI deep dive that laid out the team's response to this challenge. Teams and mentors thought together about how to ensure this RAI deep dive 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 RAI deep dives (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 RAI deep dives. 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 raising and maintaining the resources necessary to keep the project running once deployed; 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; and sustaining adherence to RAI principles over time.

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 AI solutions. 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 analyse 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.

Introduction

There is an increasing awareness amongst policymakers and the public, that a responsible approach to artificial intelligence (AI) is needed to minimise harm arising to people, societies and environments as a result of the large-scale deployment of AI systems, and to ensure the beneficial and sustainable use of AI. However, as AI has rapidly advanced, and quickly come to influence and intermediate many different aspects of our lives, policy has struggled to keep pace, and it has proved practically difficult to implement best practice standards of responsible AI.

The development of AI has exhibited a high degree of market concentration due *inter alia* to uneven access to capital and data assets; and first mover advantage. This has made it difficult to systematically institutionalise responsible AI (RAI) principles across the sector, especially as smaller players seeking to develop explicitly responsible AI solutions have faced greater barriers to scaling and gaining a foothold. While several frameworks outlining what should constitute RAI have emerged globally, there are insufficient incentives to develop and adopt RAI, as well as insufficient practical guidance as to how to apply and adhere to principles of RAI in diverse contexts and projects.

Nascent RAI applications have proliferated, including outside the traditional centres of computational and technological power. The successful scaling of such initiatives across diverse contexts is critical to challenge the currently uneven, highly concentrated and generally under-regulated trajectory of AI development. However, these initiatives require support from funders, technical experts and policymakers in order to scale responsibly. Moreover, adopters need much greater guidance and oversight with respect to the appropriate and responsible use of AI technologies.

A key concern with the lack of supranational consensus on the responsible governance of AI means that understanding of "how to operationalize high-level principles such that they translate to technology design, development and use…" (Cole et al., 2022, 1) is limited. The vacuum of guidance and oversight in terms of actual implementation has served to enable narrow interpretations of ethics and responsibility in certain instances.

This report summarises a project undertaken from January to October 2023 by the Global Partnership on Artificial Intelligence (GPAI), in response to the challenge of adoption and scalability of RAI. The project, called Scaling Responsible AI Solutions (SRAIS), matched teams from five different countries who were at different maturity stages of developing AI solutions, with experts and specialists in RAI.[1] The teams received tailored mentoring, aimed at supporting them both to integrate best practice standards of RAI, and to scale their AI solution.

Teams gained an opportunity to collaborate with GPAI experts and draw on insights from various perspectives, define and adopt RAI performance metrics and showcase results. The overall objective of the project was to produce tangible outcomes towards scaling RAI.

The following section outlines major elements of responsible AI frameworks, and elaborates on the conception of responsible AI which underpinned this project. The report then goes on to discuss some of the structural challenges identified to scaling responsible AI in the wider literature. Following that we introduce the various phases of the project, and then go on to introduce and

[1] Note, to differentiate these participating teams from the overall SRAIS project within GPAI, they are referred to in this report as teams or participating projects.

discuss each of the five selected projects, including their aims; the challenges they each identified at the outset of the SRAIS project; and the focus and preliminary outcomes of their engagement with the GPAI experts. After that we provide a series of recommendations for AI teams and policymakers, based on the insights of this project, to better enable the scaling of RAI solutions. Finally we give an overview of the next steps for the SRAIS project.

Understanding Responsible AI

Standards, recommendations and guidelines for responsible AI have proliferated in recent years. These have been advanced by industry, NGOs, national and regional policymakers, and international and multilateral bodies. This widespread interest in establishing frameworks for developing and deploying AI responsibly throughout the entire AI lifecycle,[2] has generated a wealth of innovation and constructive debate. However, these frameworks have been predominantly developed in the context of the Global North and may not fully take into account Global South challenges and priorities (Jobin et al, 2019). This has also resulted in a varied and sometimes confusing landscape of standards and approaches. Within this landscape are competing vocabularies for articulating ethical concerns and governance priorities for AI, varyingly termed *inter alia* "AI ethics", "trustworthy AI", "responsible AI" (Schiff et al., 2021). This lack of shared understanding in itself can act as a barrier to developing and scaling AI applications which uphold equity and human rights (Munn, 2022).

As such, interpretations and operationalisation of responsible principles for AI have in large part been focused on considerations around transparency, explainability, privacy and security for individual users. There has been less attention paid in practice to questions of equity, fairness, human rights and justice, including throughout the AI lifecycle (conception, design, development, deployment, and monitoring and evaluation), and at the collective and societal levels (GPAI, 2022a). These latter principles involve much greater nuance, complexity and contextualisation to operationalise.

The GPAI Responsible AI Working Group understands responsible AI as AI that is "human-centred, fair, equitable, inclusive and respectful of privacy, human rights and democracy, and that aims at contributing positively to the public good (GPAI, 2020)." This understanding of RAI has been developed in reference to the UN Sustainable Development Goals (SDGs) as well as RAI literatures, and major international frameworks which have established high-level principles for RAI.

This section provides an overview of the principles of RAI which underpin the work of the GPAI RAI Working Group, and the Scaling RAI Solutions project. We draw in particular on the OECD's Recommendation on Artificial Intelligence (2019). We discuss each element of RAI in turn, in order to give context to the rationale, methods and outcomes of the Scaling RAI Solutions project.

Human-centredness

The growth of AI systems have led to fears of adverse impacts on people and society. Identified risks and harms to people of AI systems include reduced agency over our lives as difficult-to-understand, or "black box" algorithms exert increasing influence over our day-to-day activities. They may also include the displacement of activities previously reserved for people, such

[2] OECD (2019) defines the AI system life cycle as involving: "i) 'design, data and models'; which is a context-dependent sequence encompassing planning and design, data collection and processing, as well as model building; ii) 'verification and validation'; iii) 'deployment'; and iv) 'operation and monitoring'. These phases often take place in an iterative manner and are not necessarily sequential."

as creative and intellectual activities, as well as work tasks, jobs or entire professions, in addition discrimination and non-inclusiveness have been identified as key concerns arising from the use of AI systems.

While such harms are possible, AI itself does not necessarily have predetermined outcomes, and its potential harms must be mitigated. It is important to understand AI as a general purpose technology that can have both positive and negative outcomes depending on how it is designed, built and deployed, by whom, and for what objectives. As such, calls have grown to find ways of making sure that AI is put to use in a way that benefits humanity first and foremost.

Human-centred AI refers to AI systems over which users can have a high-level of understanding and control. Rather than eroding the need for human creativity and intelligence, human-centred AI augments and enhances human lives and capabilities (Xu, 2019; Shneiderman, 2021). Stanford's Human Centred AI Initiative proposes three pillars of human-centred AI: First, AI should "incorporate more of the versatility, nuance and depth of the human intellect"; second the development of AI should be guided by ongoing research on its impact on human society, and third, that "the ultimate purpose of AI should be to enhance our humanity, not diminish or replace it" (Li & Etchemendy, 2018). Reference to the importance of human-centred AI appears in the OECD Recommendation of the Council on Artificial Intelligence (2019), which states that "actors should implement mechanisms and safeguards, such as capacity for human determination," and references the need for AI to augment human capabilities and enhance human creativity (p. 7).

Fairness

Upholding the principle of fairness in the design and use of AI systems can encompass a broad range of sociotechnical considerations, many of which intersect with other aspects of RAI. Existing frameworks for RAI touch on fairness in several respects, including the need for proportionality, transparency and explainability, the ability to meaningfully contest outcomes, security and safety of users and data subjects, as well as adherence to labour standards and decent work. We may also see unfairness stemming from the way in which the data used to train AI systems is governed, alongside the data collected by and through the use of AI systems.

Ensuring fairness begins at the earliest stages of the conception and design of an AI application. The principles of proportionality and do no harm require designers to carefully consider whether their application is appropriate and proportionate for the context in which it is intended to be used, and for the problem it attempts to solve, or whether it is necessary to achieve its "legitimate aims and objectives" (UNESCO, 2021, p. 20). For instance, does the problem require an AI solution, or is there a different approach that may be more well-suited, and carry fewer risks? This requires designers to undertake rigorous risk assessments, considering possible risks to human beings, communities, societies or ecosystems, and putting in place measures to prevent all possible harms identified.

Fairness also requires any stakeholder potentially affected by an AI system throughout its lifecycle (including users, but also workers involved in building the system), to be empowered to meaningfully contest any adverse impacts or harms they experience in relation to the system. The ability to meaningfully contest harms and challenge decisions necessitates transparency and explainability with respect to how systems are built and deployed. This includes making information available in an accessible format, on how data is collected and used, how and why algorithms produce particular outcomes, and how AI systems influence and participate in decision-making processes. In the context of generative AI, transparency might require a disclaimer to accompany all content

generated by AI, and such a disclosure requirement is present in the draft EU AI Act (European Commission, 2021).

Transparency and explainability is crucial not only for individuals to challenge outcomes, but also for national and international regulation and accountability mechanisms to function effectively. As such, it must be sensitive not only to individual-level outcomes, but also to collective-level outcomes, for instance with regard to how outcomes may differ by gender. However, it is also important to note that transparency and explainability needs to be balanced with other rights and freedoms, most notably privacy and security—specifically, transparency should not entail the making available of sensitive information (either personal or non-personal), which could adversely impact individuals or communities.

Another important way that unfairness arises in the AI lifecycle is with respect to issues of labour standards, job quality, and decent work. This touches on questions of how AI technologies are deployed in the management and intermediation of work (Cole et al., 2023; GPAI, 2022b); possible industrial disruption entailed by the entry of AI technologies which result in the displacement of human labour (see “human-centred” above); as well as the conditions of the millions of workers involved in building and maintaining AI systems, who are disproportionately low-wage and precarious workers located in the Global South (Gray & Suri, 2019; Howson et al., 2022a; Anwar & Graham, 2022). Ensuring fairness in the development and use of AI requires concerted attention to articulate and uphold minimum standards of fair work for all workers, the building of fair AI production networks, ensuring accountability and human oversight in algorithmic decisions that affect workers, the protection of worker data, attention to how AI might impact on issues of health and safety in the workplace, and the advancement of collective worker voice (GPAI 2022b).

Equity

An extremely large body of research has demonstrated the propensity of AI systems to reproduce social biases (Raji et al., 2022; Buolamwini and Gebru, 2018), and embed structural and historical injustices, (Birhane, 2022; Noble, 2018), including the erasure or marginalisation of communities such as indigenous groups, language groups, and others (Kukutai et al., 2020). This has been shown to occur in predictive models, large language models, and large statistical models amongst others. This occurs through a number of mechanisms, including the use of unrepresentative or biassed datasets to train AI systems, inadequate consultation and contextual sensitivity in the design and use of AI systems, as well as a lack of representativeness amongst those conceiving and developing AI systems. As a result, AI systems carry a very high risk of encoding, reproducing and exacerbating dimensions of inequality within and between countries, including along gendered, racial, linguistic, caste, and abled lines. In order to mitigate the risk of inequitable outcomes, responsible AI solutions must not only ensure representativeness and promote diversity within datasets, but also within the teams responsible for identifying the problem and defining and developing the solution. In addition, there must be provision for monitoring, auditing and mitigating equity and discrimination issues that arise after deployment and through scaling. This includes built-in frameworks for individuals and groups to contest discriminatory outcomes.

Finally, to avoid perpetuating structural and historical injustice on a new, digital scale, RAI approaches must also go beyond diversity and representativeness, to strive for the shared governance of data and other AI infrastructures amongst all those who contribute to and are impacted by them (for instance through open source datasets, data trusts, physical connectivity infrastructure, as well as expanded access to AI literacy and skills; broadening the base of people who can develop AI, and participatory methodologies for designing AI systems) (Singh &

Gurumurthy, 2021; GPAI 2022c). This calls for ongoing and primary attention to be paid to who benefits from AI systems, and to ensure that the benefits of AI are shared equitably across stakeholders. This calls for the advancement of such objectives as shared ownership and governance of AI infrastructure), as well as equitable access to devices and connectivity, within and between countries. Many commentators have called for an expanded understanding of data not as a proprietary asset of large corporations, but as a socially-produced resource, or a commons, the use and benefits of which should accrue to those who produce it (i.e. everyone).

Human Rights, Democracy, Sustainability and Contribution to the Public Good

The OECD recommendation includes reference to the need for AI to uphold human rights, and democracy. The rights most frequently invoked in AI ethics discourses centre on privacy and data protection, which are often seen as most at risk of being impinged by AI. Alongside privacy and data protection, The OECD recommendation also refers to the need for AI actors to uphold internationally recognised labour rights.

Prabhakaran et al. (2022) argue that the doctrine of human rights provides a normative framework for grounding progress on AI ethics and responsibility. Human rights provide an explicit, universal set of values, which enable AI ethics discourses to move away from reliance on implicit norms and values. They also reorient the focus away from the systems themselves and towards a more human-centred approach.

In addition to human rights, RAI should also uphold democracy, and not impinge on our ability to live in peaceful and just societies. This includes guarding against mis and disinformation, which has arisen through the use of algorithms and platforms and has already impacted the health of democracies. It goes beyond this however, to encompass full consideration of peace, inclusiveness and interconnectedness. The OECD recommendation states that AI actors should: “respect the rule of law, human rights and democratic values, throughout the AI system lifecycle” (p. 7).

This also points to the importance of environmental sustainability to RAI. AI systems can have tremendous impacts on the natural environment, including through water and energy usage—the latter of which can contribute significantly to carbon emissions. The OECD recommendation states that AI actors should: “proactively engage in responsible stewardship of trustworthy AI in pursuit of beneficial outcomes for people and the planet” (p. 7).

Rationale for the SRAIS Project: Challenges with Scaling Responsibly

Many actors around the world are highly motivated to use advances in AI technologies to benefit communities, solve social and environmental problems, and make people’s lives better. Uses of AI which aim to contribute to the common good, are sometimes seen to be inherently responsible. However, as outlined in the section above, ensuring an AI application adheres fully to the various dimensions of responsibility and harm reduction is a complicated task with many intersecting considerations, which requires significant expertise and oversight. Ethical intentions do not automatically ensure beneficial outcomes. Therefore, risk and ethical impact assessments are required.

A key issue is that access to the necessary multidisciplinary expertise and oversight to realise RAI deployment is not available to most developers, and there is a lack of consensus amongst various

actors including funders and policymakers regarding minimum standards of responsibility and ethics in AI. There is a need to bridge the gap between theory and practice in RAI, and developers aiming to scale an AI solution need a blueprint which guides them to do so responsibly.

Moreover, there are a number of structural barriers to scaling which in some cases can serve to disincentivise adherence to RAI principles. In other words, in the current landscape of under-regulation of AI, it may be easier to scale for those who do not pay as strict attention to ethics and responsibility. To further complicate this issue, the process of scaling itself can add new, often unanticipated social and environmental risks and harms. As such, teams face challenges in both ensuring their solutions are designed responsibly, as well as in scaling, and they also face distinct challenges in *scaling responsibly.*

The rationale for the Scaling RAI Solutions project is to assist teams with realising responsibility not just at the point of design, but at all stages of the AI lifecycle, and to help truly responsible AI solutions overcome scalability challenges whilst preserving responsible standards. Figure 1 illustrates this space, whereby GPAI experts aim to support teams to progress along axes of scalability and responsibility simultaneously.

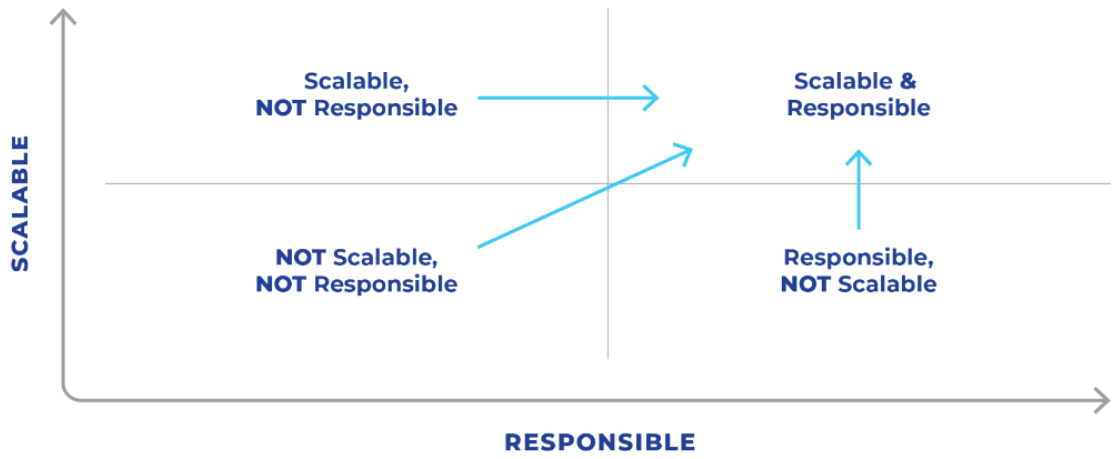

Figure 1. Balancing scalability and responsibility

It is, however, important to state one caveat to the above. This is that not everything needs to be scaled to be legitimate, or to have impact. Some solutions may be necessarily limited to a specific function in a specific context, and this may be appropriate and valid. In some cases it may be more appropriate to adapt and replicate particular solutions for different purposes or in different contexts, rather than to apply the same solution to different contexts or uses. The Scaling AI Solutions project focuses on solutions which have a demonstrated need to scale (for instance to achieve pubic good objectives), in acknowledgement of the fact that scalability is a particular challenge for responsible solutions, and that scaling responsible solutions where appropriate is necessary to challenge unevenness, concentration, and patchy integration of ethics in the AI landscape.

General Scalability Challenges for Nascent Applications

The global landscape of digital technology ownership is highly concentrated and uneven. In particular, the ability to access and utilise large stores of data as well as computational infrastructure (a key prerequisite for developing safe and effective AI) has been concentrated amongst a select

few platform companies, overwhelmingly located in the United States and China. In 2019 over 90% of the market capitalisation value of the 70 largest platform companies was located in China and the US, with Europe accounting for just 4%, and Africa and Latin America together for 1% (UNCTAD, 2019).

The typical business model of global platform companies has been centred on market dominance, via a few key strategies (Kumar, 2020). Hallmarks of the big platform business model include the extraction and appropriation of valuable data from communities (Sadowski, 2019), the extension of global value chains into service sectors (Howson et al., 2022), the use of continuous injections of venture capital to cross-subsidise supply and demand to rapidly enrol users (Peck & Philips, 2020), the intermediation of transactions, and exploitation of network effects (Langley & Leyshon, 2017), the fierce guarding of proprietary, "black box" algorithms resulting in information asymmetries between owners and users (Pasquale, 2015), the surveillance and manipulation of users (Zuboff, 2019), the exploitation of precarious outsourced labour (Gray & Suri, 2019; Tubaro et al., 2020) and the circumvention of existing regulation, including through geographical disembeddedness (Katta et al., 2020), as well as the deployment of massive lobbying efforts to enter into markets and protect market positions (Lewis et al., 2022).

These strategies are squarely aimed at establishing control over markets, which in turn enables big players not only to lead innovation, but also to easily acquire promising start-ups. The resulting oligopolistic nature of the data/AI-driven sectors makes it very difficult for small players aiming to scale themselves to compete—and even harder for those hailing from marginalised geographies, communities or social groups. Even if not deployed at scale, artificial intelligence is inherently a *scaled* phenomenon, in that it requires very very large quantities of data to operate, and to provide accurate and safe outcomes. This reality offers a fundamental advantage to those who have already been able to accumulate large data assets, and has underpinned widespread calls from different parts of the world, for innovations in inclusive data governance through (for example) open source data, data trusts, and data commons (Grossman et al., 2016; O'Hara, 2019), as well as conceptions and protections of national and indigenous data sovereignty (Hummel et al., 2019). As such, key challenges for new players to enter and scale include access to both quality data, as well as computational infrastructures (cables, servers, devices, electricity, etc.).

Another barrier to scaling is the uneven access to investment in the platform/AI sector (Harrison et al., 2020), which at a global level is underpinned by numerous factors including the strength of public institutions, auditing and accounting standards, fiscal health (Benali & Galfiki, 2021), as well as geographical biases or linguistic or cultural norms. Crowdfunding (or raising small injections of funding from a wide and diverse audience, often via a digital platform) is sometimes seen as a 'democratising' solution to this issue. However, research by Gallemore et al. (2019), found that the success of crowdfunding initiatives on the platform Indiegogo were still influenced by location—with rural initiatives less likely to be successful. Langley & Leyshon (2017) also argue that crowdfunding might replicate rather than disrupt uneven patterns of access to capital.

Importantly, with respect to funding for scalability, a key challenge for new entrants is that of ongoing operational costs. Access to funding may diminish after the point of deployment. Due to a general (mis)conception that AI systems are fairly self-sustaining, and not requiring of significant human input to function, stakeholders are often at risk of underestimating the ongoing operating costs of AI systems, including the costs of monitoring, maintenance and improvement, as well as the costs of, for instance, keeping servers running. These costs can increase significantly as a system is scaled.

Alongside issues of market access and funding, developers face complex technical challenges when scaling AI initiatives. One issue is that of interoperability—whereby datasets, algorithms and software are unable to communicate and exchange with each other. Another is the absence of frameworks to facilitate the porting of data by users and parties between different platforms and AI ecosystems. This becomes a greater challenge to scaling across different sectors for instance the public and private sectors, as well as across borders. The proprietary nature of most AI development can hamper innovation in the public good, as systems are enclosed and exclusive, unable to share information and function in a connected way. Lehne et al. (2019) argue that the absence of interoperability negatively impacts the use of AI to its full potential in health settings, with implications for patient well-being worldwide. However, interoperability also entails risks that must be carefully monitored and navigated, especially in terms of data protection, security and safety of users and impacted communities.

Challenging this status quo requires coordinated national, supranational and international regulatory efforts, which bring together affected communities, including users, workers, legislators and industry to establish clear frameworks and guidelines. However, the Scaling AI Solutions project has demonstrated to us the wide breadth of initiatives currently trying to navigate this complex and often unfavourable terrain, and to do so while upholding RAI standards.

Specific Challenges With Scaling Responsible AI Solutions

The above section outlines general structural, cost-related and technical challenges to scaling AI initiatives. However, integrating frameworks of responsible AI introduces specific challenges, as well as new opportunities with regard to scalability. The above section outlines general challenges in scaling AI initiatives. However, integrating responsible AI frameworks introduces both challenges and opportunities with respect to scalability. First and most straightforwardly, incorporating measures of social, economic, environmental impacts into understandings of success requires additional resources and capabilities—meaning responsible AI applications are more complicated and costly to scale, compared to AI applications which are solely focused on generating value.

In addition, governance challenges make it difficult for developers to operationalise and validate RAI standards in practice (Cole et al., 2023). This is due both to a general lack of regulation for RAI, as well as competing understandings of responsibility and ethics in global regulatory frameworks that do exist for AI. Although there are emerging high-level principles and values for what constitutes RAI (as outlined above), there is sometimes a lack of clarity with regard to how these can be practically implemented in different use cases, sectors and contexts. Although many projects may have an in-principle commitment to RAI, in order to scale an RAI initiative they need to be able to demonstrate adherence to the range of RAI standards via an independently-verifiable framework which enables auditing and monitoring of projects against established metrics. This becomes important for key aspects of scaling, including access to different markets and jurisdictions, access to funding, as well as building trust with stakeholders to facilitate adoption. However the current landscape is characterised by emerging regulatory standards in some jurisdictions, with different degrees of complexity, but little clarity with regard to compliance, or designation of responsibility.

Also key to establishing trustworthiness and safety of AI applications, is the ability to rigorously test them in a controlled way prior to deployment and scaling (Aghajari et al., 2023). Testing should focus not only on whether an application works as intended, but also, crucially, on its potential impacts on people, societies and environments. Depending on the application, controlled testing may call for the involvement of human participants, in order to gain a full understanding of its potential impacts. As with other innovation, research and development (for instance in the medical

field), there is a critical need for standardisation, resources, oversight and approval bodies to facilitate testing and experimentation in AI development, in order to ensure that any AI application being deployed for public or private use has been demonstrated to be safe and responsible. AI testing protocols should mirror and build on best practice research ethics standards used elsewhere. While testing protocols need to be extremely robust, rigorous and accountable to independent bodies, they should also be accessible, navigable, and streamlined as much as possible for developers—especially to promote equal access to testing and experimentation capabilities. However, as it stands, many AI applications are released for public use with little testing or experimentation, and their negative impacts become visible only after scaling, as they are experienced by users.

To further complicate the need for testing and exploration to ensure responsibility and trustworthiness, while developers may give due consideration to RAI principles in the early stages of product design, it is also important to note that the process of scaling itself can endanger or undermine responsible and ethical standards even if they are present in early prototypes. For instance, biases in datasets may only be identified when implemented at scale at the point at which they cause real-world harms. This is due to the fact that smaller samples of users (for instance within a single region or country) may not include sufficient diversity and representativeness for biases to be identified. Harms and concerns arising from AI applications may differ across social groups, communities and geographies, therefore scaling an application across contexts may introduce unanticipated contextual concerns and harms. In addition, as outlined previously, meaningful stakeholder consultation and even co-design of applications with intended beneficiaries is key to ensuring AI systems are responsible. However, implementation at scale (across different regions, industries, social groups, etc.) may make meaningful stakeholder consultation more difficult as there is greater (physical, social, cultural) distance between the designers and impacted communities.

Finally, while the previous section discussed general barriers and challenges to raising funding for nascent AI initiatives, it is also important to understand how funding concerns impact on and interact with values and principles of RAI. In the uneven and highly competitive landscape of venture capital, and without sufficient regulatory guardrails, raising the capital necessary to scale can introduce incentives which are at odds with RAI principles in some instances—as a result of obligations to investors and shareholders. For instance, adhering to best practice RAI principles throughout the AI lifecycle can add to costs, and lengthen timeframes. This can and should be balanced by institutionalising minimum standards for RAI in regulation, as well as raising awareness and literacy amongst investors, end users and the public at large around the need for RAI, which will create market incentives for responsibility.

Phases of the SRAIS Project

Selection of Teams

A call for participation was issued in April 2023. It invited applications from teams who had reached at least the pilot stage with an AI application and were experiencing challenges related to responsible AI adoption and scaling. Prospective participating teams were asked to provide information on: their application domain; a description of their application including the problem them were trying to solve, the specific uses of AI and data within the application, intended impact of the application and metrics of success; whether an AI responsibility assessment had been conducted or was planned; and challenges encountered or anticipated in development, adoption, and scaling.

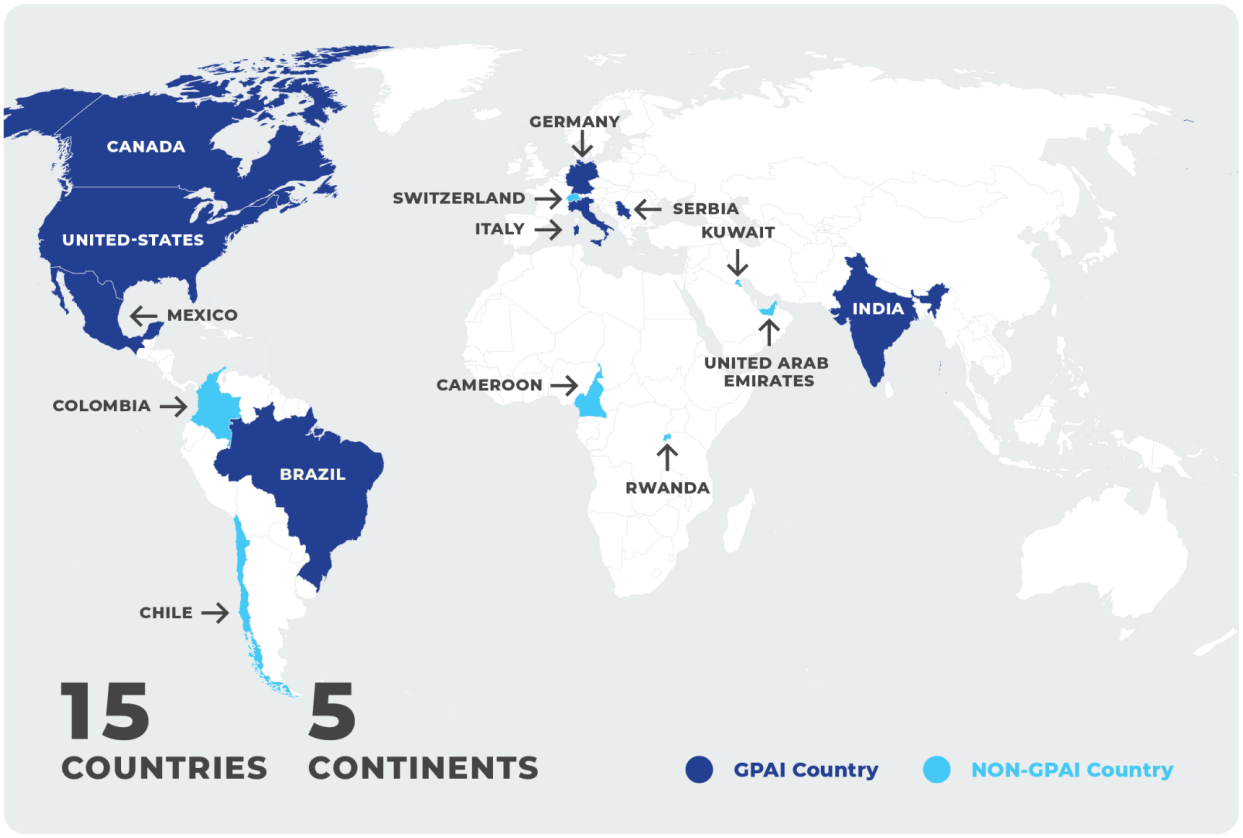

Figure 2. Countries from where applications were received for the SRAIS project

AI Team	Country	Topic	GPAI Focus Area	Document generated by the AI team as a result of the mentorship program
COMPREHENSIV – At Home Universal Primary Health Care	India	Healthcare Access	Global Health	Challenges and strategies for responsible data collection and management for future AI healthcare applications
ergoCub: Artificial Intelligence in Wearables and Robotics for Assessment	Italy	Workers Well-Being - Healthcare	Human Rights	Trustworthy AI guidelines for wearables in healthcare and industry
Jalisco's AI Forest Mapping System	Mexico	Environment	Resilient Society	Key challenges and potential solutions for developing a responsible decision support system based on a classification model
Wysdom AI: Advanced Chatbot and Voicebot Analytics Tools	Canada	AI Systems Audit	Resilient Society	Responsible AI pillars for conversational systems analytics
Particip.ai One: Participation and Feedback Platform	Germany	Democratic Participation	Human Rights	Guidelines for responsible use of AI voice and chatbot technology for people-centred participation and feedback

Figure 3. Overview of participating teams

23 teams applied, from 15 countries across 5 continents. Teams were selected based on a three-tier prioritisation framework. The highest priority were applications which were both grounded in responsible AI principles and faced challenges with scaling. The second priority were applications that had demonstrated potential for scaling, but needed to assess and validate their alignment with RAI principles. The third priority were applications that faced challenges both with respect to scaling, and also with ensuring and demonstrating responsibility.

In addition, evaluators used several further criteria to assess applications for participation, which mirrored the prioritisation framework included in the call for proposals. These included the teams' ability to address challenges; their awareness of RAI issues and desire to scale responsibly; how well they structured their problem definition; the potential impact of their proposed solution and the significance of impact with respect to application area, size of impacted population, inclusion of underserved countries and communities, and ability to disseminate widely; the suitability of their proposed solution for addressing the identified problem; the feasibility (scalability) of the proposed solution; the presence of mechanisms to address RAI issues; consideration of data governance and indigenous rights; and the presence of a risk assessment framework.

Finally, evaluators gave consideration that participating projects were spread across a range of geographies and sectors—with particular consideration given to the balance between Global North and Global South teams, as well as gender representativeness. As such the five teams selected came from five different countries, with a balance of global North and South teams. Alongside this

regional diversity, they had a range of aims and sectoral focus areas, across private and public sector applications.

Seven teams were initially selected, but two subsequently needed to withdraw due to capacity and funding constraints. The five participating projects are detailed below, along with the objectives that were identified for each project by the GPAI mentoring team, at the outset of the Scaling Responsible AI Solutions project.

Mentoring Process

After the participating projects had been selected, a series of preliminary meetings took place in which the participating projects had a chance to introduce themselves to each other and discuss the responsible scaling challenges they faced. Thereafter the mentoring process commenced. Each participating project was matched with a group of GPAI experts whose expertise was particularly relevant to their challenges and aims (ranging from 2-6 experts per team).

The mentors and the participating projects had three sessions together. The first session was dedicated to identifying and agreeing on a key RAI deep dive to be produced by the conclusion of the three sessions. The RAI deep dive for each team would constitute a two page document focusing on responding to a key RAI challenge identified collectively by the teams and mentors. Where relevant to the participating project, the documents included clear KPIs for responding to identified RAI challenges. These documents are available in the appendices to this report. During the subsequent mentoring sessions, the GPAI experts guided the teams to think about their challenges in new ways, and offered input and feedback to the documented deep dives, to assist teams in the process of scaling responsibly.

Implementation and Evaluation

Following the conclusion of the three mentoring sessions and the production of the RAI deep dive to guide teams to identify and respond to their key challenges in scaling responsibly, teams are (at the time of report writing) being supported to produce a RAI plan. The RAI plan will include clear steps and measurement frameworks—tailored for each team—towards realising the objectives and KPIs formulated through the mentoring process and production of the two page document. The teams will then have access to robust evaluation of their RAI plans by an evaluation committee made up of GPAI experts. Following this evaluation, evaluators will issue a certificate, or "GPAI Experts' Certificate of Responsible AI Progress" to the teams who successfully complete the work for their RAI plans based on the mentorship outcome, i.e. the RAI deep dive.

The Participating Projects: Aims, Challenges, and Outcomes of the GPAI Mentorship Phase[3]

Wysdom Smart AI Analytics Tools — Team from Canada

Wysdom is a Canadian company, offering chatbot analytics to clients. Wysdom.ai utilises AI to evaluate intent-based conversational bots, to gain insights into and address challenges with their performance and behaviour, in order to improve user experiences. Wysdom's analytics focuses on helping companies to optimise conversational flow and bot responsiveness.

The advent of generative AI chatbots presents a key scalability challenge for companies like Wysdom. Applying analytics to generative AI requires much greater capacity to deal with growing data volume and complexity in order to provide real-time analysis and insights. In addition, generative AI introduces new risks, including in relation to transparency and accuracy of responses. A key challenge lies in developing robust frameworks to algorithmically detect and prevent harmful and biased outcomes in generative AI chatbot interactions. There is also a need to find ways to prioritise data protection, consent management and transparency whilst adapting analytics systems to respond to generative AI.

Insights and Outcomes of the Mentoring Process

The GPAI mentors agreed that it will be important for Wysdom to think much more systematically about how their application may intersect with RAI considerations. However, the mentors also noted at the outset of the process, that in doing so Wysdom is well-positioned to take advantage of a significant opportunity, to assist companies with the responsible adoption of generative AI, and to contribute to the dissemination and implementation of RAI principles. They identified this as a valuable focus for the mentoring programme. As such, the mentors and the Wysdom team agreed that the key focus of their participation in the SRAIS project should be to guide the use of the Wysdom chatbot analytics system for the identification and mitigation of RAI risks.

It was noted that compared to some of the other participants Wysdom has a relatively large team and advanced capabilities which should enable them to define and monitor KPIs. Wysdom has been providing their bot management service since 2017, and has a team of approximately 70 employees. They have established partnerships with many of the largest players in the tech landscape, including Google, Microsoft and Amazon.

In the preliminary discussions, the mentors raised several RAI considerations that Wysdom should integrate into their analytics. Due to the nature of the application, data and human-centred challenges came to the fore.

RAI Deep Dive: Responsible AI Pillars for Conversational Systems Analytics

It was decided that Wysdom's RAI deep dive would be a document laying out a series of RAI indicators that could be applied to evaluate the performance of chatbots, alongside performance-related indicators already being applied by Wysdom. Ultimately thirteen RAI indicators were identified: Scalability and performance; bias detection and mitigation; user privacy and data

[3] Note: The content of this section has been supplied by the participating teams, who have confirmed that the information provided is complete and accurate.

protection; transparency and explainability; ethical interactions; security; multilingual and multicultural considerations; user education; regulatory compliance; user consent; copyright protection; A/B testing and evaluation and sustainability. The document produced by the Wysdom team (Annex A) lays out each challenge, along with a brief discussion of how it can be detected/measured, and mitigated.

COMPREHENSIV: A Digital Platform for at Home Universal Primary Health Care, Data Life Cycle Management Challenges and Strategies — Team from India

COMPREHENSIV is a smartphone application designed to be used by trained field personnel to screen for and manage a large range of early-stage disease conditions, in real time, based on images of the conditions, along with other relevant images (such as of living situations) which provide context for sociodemographic, economic, WASH (Water, Sanitation, Habitat and Hygiene), nutrition and disability status. It is aimed to be easily operated by any trained field worker, including in instances of low digital literacy. The use of icon-based communication and image-based screening is considered to be important in overcoming language limitations and improving field usability.

COMPREHENSIV—or the concept of home-based primary health care on a digital platform for resource constrained setting—started as a PhD project in 2008. It progressed through work undertaken by faculty and students in India's first postgraduate (MSc and PhD) program in Health Informatics since 2013 and transitioned into a startup company called Hi Rapid Lab with due approvals in 2021.

Currently COMPREHENSIV has been rolled out in specific regions in which Hi Rapid Labs has networks with existing service providers. The data collected via the application is subject to institutional ethics approvals and individual consent. However, expanding to other regions and countries has been identified as a responsible scalability challenge. This arises particularly in relation to the need to ensure appropriateness and applicability of the tool in different local contexts, including with regard to cultural considerations, and amongst population groups with different risks and needs.

Insights and Outcomes of the Mentoring Process

In preliminary discussions, the GPAI team identified a range of areas to focus on with Hi Rapid Labs, with respect to scaling COMPREHENSIV responsibly. It was identified that there should be greater consideration of how the tool complements the work of existing healthcare workers, to enhance their capabilities. There was also a sense that the role of AI in this project should clearly highlight the social benefit of specific AI uses, and be understandable in the settings where it is (or will be) being used. The use of AI within the tool should also be transparent, and its parameters and outcomes explainable and able to be understood by users, especially those living in resource constrained settings. Multiple multimedia communication pathways, beyond traditional research publications, were discussed as potential solutions.

There was also discussion of ensuring responsibility within the data collection protocols for COMPREHENSIV, both in terms of the datasets used to train the AI elements of the system, as well as the data collected from end healthcare users. This emerged as the key goal and focus of the mentoring process. It was noted that although the COMPREHENSIV team is in the relatively early

stages of maturity, there is a need to ensure strong data governance protocols from the earliest stages, as data is collected to further develop the solution. In particular, there is an important need to ensure both privacy and security, particularly if images are sensitive and/or personally identifiable, as well as a need to ensure representativeness of datasets. The Hi Rapid Lab team highlighted the multiple informed consent options that are built into COMPREHENSIV, the universal screening and service approach that substantially reduces bias. The discussion led to expanded understanding of the data life cycle and end cycle processes and the need to develop a robust framework even in the early stages.

The team identified the need for the framework to be acceptable to regulators and future clients for applications beyond research. Due to the nature of the application and its use in healthcare settings, there is a need for identity to be sufficiently preserved in data in order to be appended to record. This requires robust protection protocols, and a clear articulation of responsibility principles for such data, especially because it is collected outside a formal healthcare service relationship.

RAI Deep Dive: Challenges and Strategies for Responsible Data Collection and Management for Future AI Healthcare Applications

COMPREHENSIV's RAI deep dive was the development of a trustworthy and compliant data governance framework. The mentors advised the team to produce a document outlining and justifying the type of data being collected, as well as how informed consent would be obtained for this data collection. The key focus areas were ethical data acquisition; responsible data storage and use; and responsible algorithm development and data sharing practices. The resulting document articulating COMPREHENSIV's responsible data governance framework is available as Annex B.

Jalisco's AI Forest Mapping System — Team from Mexico

The Western part of Mexico's Jalisco region is a deforestation hotspot. This project aims to develop AI models that can monitor and detect illegal deforestation as early as possible, to enable authorities to identify the most critical/urgent restoration and conservation needs, and to respond more rapidly. The region in question is part of the avocado production belt, and avocado plants and other crops being planted illegally in forested areas is a serious challenge for the government—especially due to Jalisco's importance as a highly biodiverse area, as well as its value in terms of carbon stocks.

In response to this problem, the team is developing AI models based on Object Based Image Analysis (OBIA), which uses open satellite data from the European Spatial Agency and the missions Sentinal-1 and Sentinal-2, as well as elevation data from the Shuttle Radar Topography Mission. In preliminary tests, the model had 80% accuracy. However, the team has not yet been able to deploy the model at scale, due in part to challenges with dealing with data complexity, and labelling and processing the data. The region the project seeks to monitor is "megadiverse", meaning that the model needs to be adept at generating results across highly varied landscapes, as well as different seasons, accurately and without bias. The team noted that there is a risk that variability in the season and region could generate bias in the model which may only be detectable after scaling. However, achieving this level of accuracy in diverse conditions requires training the model with a very large quantity of representative, labelled data—a costly and labour-intensive undertaking. The accuracy of the model is critically important to ensuring responsible deployment—due in particular to the fact that the tool is intended to be used by authorities to enforce regulatory provisions (including potentially informing legal and justice-related processes such as prosecution and fines).

In addition, the team is still working on creating a more user-friendly interface, in order to be able to deploy at scale.

With respect to deployment and scaling the team also identified challenges and costs related to the ongoing maintenance and monitoring of the model. These include integrating an automatic data pipeline, including data preprocessing (cleaning, etc.), and evaluation of data quality and model metrics. Finally there is a need to include human in the loop supervision to critical tasks with low AI-model accuracy or confidence.

Insights and Outcomes of the Mentoring Process

In their preliminary discussions, the GPAI mentors noted that it was positive that the project had a clear purpose, which was in line with a specific sustainable development goal, and aimed to contribute to the good of people and the planet. However, they also identified critical areas the team would need to focus on in order to ensure responsible deployment and scaling. In particular, it would be important to set clear parameters for the appropriate uses of the model, as well as to be clear on its limitations. For instance, a model with 80% accuracy may be very useful in providing aggregate data to gain insights into general trends, but may not be suitable to inform decision making in activities related to policing, fines, prosecution etc.

The mentors suggested that the focus of their work with the Jalisco team should be on translating the model into a system that can be used to support human decision making in a responsible, proportionate, unbiased and contestable way. Moreover, they identified possible responsible and beneficial uses for the Jalisco team's model outside of its application for policing deforestation. This could include assisting with comparative monitoring of land use to inform systems of carbon credits for instance.

RAI Deep Dive: Key Challenges and Potential Solutions for Developing a Responsible Decision Support System Based on a Classification Model

From these discussions, Jalisco's RAI deep dive emerged, as a guideline for translating the AI models into a decision-support system whilst upholding RAI principles. The document focused on RAI challenges that arose in the course of realising this objective, and ways to overcome them. The key challenges identified were: The robustness of the AI model and its ability to make accurate predictions within seasonal and regional variability; the ease of use of the interface as well as multilingual considerations, and the need for an adequate complaints and dispute resolution mechanism; the need for the application to be tailored to the requirements of different stakeholders (including different levels of government); and the need to better interrogate and mitigate the socio-environmental risks of the application, for instance the risk of wrongly accusing individuals and communities of participating in deforestation activities.

While the risks identified were specific to the Jalisco application, issues such as ensuring training data is able to account for variability in use settings, as well as the need for multistakeholder governance and dispute resolution, are also of much broader relevance to many AI actors and regulators around the world. Through the mentoring process and the production of the RAI document, the Jalisco team was able to identify potential responses and measures to mitigate RAI risks.

Particip.ai One — Team from Germany

Particip.ai One is a telephone voicebot-based participation and feedback platform, intended for use by citizens, employees and consumers to voice their feedback and participate in decision-making processes. The project aims to improve citizens' ability to participate in government processes and to foster government accountability; to bolster employees' ability to provide internal feedback within their organisations; and to enable customers to offer feedback and suggestions for the improvement of goods and services.

The team decided to use a telephone voicebot (as opposed to digital text-based platform), as it was perceived that a voicebot helps to overcome challenges related to literacy, better enabling users to engage in dialogue and express their opinions. The team judged that the use of voicebot technology via telephone improves accessibility and encourages inclusion of those facing barriers to the use of text-based and digital platforms, including people without internet access, with low literacy, as well as those with disabilities such as vision-impairment or dyslexia.

Particip.ai uses AI across a wide range of tasks and functions. NLP algorithms are used to interpret spoken language. AI algorithms are also used to analyse spoken feedback and extract insights related to participants' sentiments, to inform decision-making based on vast data volumes. The system may direct complex conversations to a human agent.

Insights and Outcomes of the Mentoring Process

At the outset of the mentoring process, the team identified scalability challenges as mostly infrastructural, alongside the need to sustain user engagement and reinforce data protection and security protocols. The key challenges they saw related to accommodating a growing user base and user load, as well as handling larger volumes of data and effectively managing and analysing the increasing volume of feedback. Achieving this at scale requires greater infrastructural and computational capacity. The team also felt that as the tool is scaled to more users, it will be important to find ways of sustaining user engagement—proposing strategies such as personalised experiences, gamification and targeted notifications to achieve this, though it is important to note that such strategies have also been associated with risks and harms to users in some contexts. Finally the team pointed to the need to customise the system to the diverse needs of different organisations and communities, which will present an increasing challenge as it is scaled to different contexts.

Alongside these scalability challenges, the mentors drew attention and focus to the complex responsibility challenges facing solutions like Particip.ai One. These included ensuring representativeness, diversity and inclusivity, especially for tools intended to be used to inform decision-making that will affect diverse groups and communities. This underscores the need to ensure that AI processes within particip.ai are completely unbiased and non-discriminatory, responsive to the needs of all users of different genders, backgrounds, language groups, accents and abilities. A key risk is that certain voices and perspectives are over- or underrepresented in the analytics and insights generated by the system, which could lead to marginalisation or exclusion of vulnerable communities from a range of processes and decisions which directly affect them, and have adverse impact on already vulnerable groups.

In addition, because data collected by the platform is intended to directly inform governance and decision-making, the processes by which this occurs must be transparent, explainable and easily

understood, and contestable by users. This requires clear communication about data usage, platform policies and decision-making processes.

Partip.ai One's primary users are likely to be governments and employers. The mentors noted that it is very important that users are given appropriate guidance on how to use the tool responsibly. For instance in workplace settings, Particip.ai One aims to allow users who are potentially vulnerable, or in a relationship of subordination with their primary clients, to express opinions which may be negative, and may result in risk or harm if their personal identity was revealed to clients. For example, an employee giving negative feedback about an employer may be vulnerable to retaliation including termination. For this reason it is critically important that user privacy is protected at all times, and that informed consent is obtained.

Finally on a broader scale, because it collects data to inform governance decisions which impact peoples' lives, Particip.ai One's model could have beneficial or harmful impacts on democratic processes, labour governance, and consumer protection, depending on the extent to which RAI principles are adhered to. It is therefore very important that the model is designed, deployed and managed in a way that centres its most vulnerable users, considering not only their individual rights, but their collective rights (for instance, data sovereignty, the protection of marginalised group identities, workers' rights to collective bargaining and due process, and the ability to meaningfully contest collective, societal and economic harms arising from the use of the platform).

Ensuring the responsible use of the tool by primary users thus formed the key focus of the mentoring process.

RAI Deep Dive: Guidelines for Responsible Use of AI Voice and Chatbot Technology for People-Centred Participation and Feedback

In light of the potentially significant positive contribution of this tool to enhanced community participation in government, management and commercial decision-making; but also the significant harms that could arise if responsible principles are not integrated, the GPAI mentors identified a need to assist the Particip.ai One team to develop a set of clear recommendations about the responsible and appropriate use of their tool to be communicated and discussed with their clients. It was agreed that the team would design and implement a set of guidelines and exemplary practices for the responsible use of the tool, aiming to bring about social benefits for the end users. The guidelines would be high-level enough to be useful to a range of different organisations in different parts of the world.

Part of this included the need to work towards a framework for the self-declaration of the voice and chatbot, to ensure that users were fully informed that they were interacting with a machine and set expectations clearly from the beginning to avoid confusion for the final user. A key question related to this was whether Particip.ai One would embed a technical solution in the chatbot's code to disclose that it is a bot or provide the client with the option to include a disclaimer.

In addition, the guidelines focused on data privacy, security and compliance; ethical interactions and content creation; feedback and accountability; and continuous review and improvement. The team's guidelines and recommendations to clients, their self-declaration framework and their data lifecycle framework are available as Annex D.

ergoCub: Wearables and Robotics for Assessment, Prediction and Reduction of Biomechanical Risk in the Workplace — Team from Italy

ergoCub aims to apply embodied AI technologies to prevent musculoskeletal disorders in workers, and for use in healthcare settings. ergoCub notes that musculoskeletal diseases have been identified as the most common occupational disease globally. This team, based in Italy, seeks to intervene in this problem, through the development of wearable technologies, as well as humanoid robots. The intention is to monitor and predict workers' physical conditions, and provide real-time alerts using sensor-equipped suits, and to enable proactive and preventative risk management including by delegating certain high-risk tasks to humanoid machines—or enabling "collaboration" between workers and robots.

The project is funded by the Istituto Nazionale Assicurazione Infortuni sul Lavoro (INAIL,the National Institute for Insurance against Accidents at Work), and coordinated by the Instituto Italiano di Tecnologia (Italian Institute of Technology, IIT). It started in June 2021, and has been funded for a period of three years. Currently the team is in the process of conducting in-field tests with workers, to validate and refine their prediction algorithms. The team anticipates deploying the AI-powered wearable sensor system in real scenarios by the project's conclusion. The team has also developed a prototype for their humanoid robot, and noted that they expect the robot to be a starting point for future AI-powered collaborative robots.

In terms of scalability, the ergoCub team perceived their challenges to be specifically related to stakeholder adoption, and processes for ethical approval. The team identified the fact that workers may be resistant to the introduction of such technologies in their workplace, due to privacy concerns. If this occurs, the ergoCub developers see it as a misinterpretation of the intent of the technology. They note that management and unions both need to be involved in promoting the technology to workers, and that these parties should be "properly instructed" on the benefits of the technology. The developers pointed to the fact that a proportion of the population is "always contrary" to the use of new technology, but suggested that this can be mitigated by the involvement of "change management experts". However, the mentors noted that there are legitimate concerns in deploying wearables that must be taken seriously.

It was also perceived by the team that requirements for ethical committee approval of trial protocols may seriously delay the project. The team needs to carry out several field-trials of different iterations of the technology, with workers. Each trial requires approval by an ethical committee, which the team noted was "time consuming". However the mentors noted that these approvals are important to obtain, even if they take time. With respect to upholding principles of responsibility, the team saw worker privacy as the central consideration for wearables deployed in the workplace.

Insights and Outcomes of the Mentoring Process

At the outset of the mentoring processes, the mentors identified a number of RAI concerns and considerations for ergoCub which went beyond privacy and individual data protection. The mentors suggested that the mentoring process should focus on developing a comprehensive set of tailored RAI indicators to guide the development and scaling of wearables in healthcare and industry—encompassing *inter alia* considerations around proportionality and appropriateness, equity and non-discrimination, informed consent, job quality and decent work, stakeholder participation,

consultation and co-design, and surveillance and data governance. It was noted that there was a need to find “common ground” for RAI indicators across different use cases in healthcare settings and industry, given their differences.

It will be important for the team to carefully consider which workplace contexts may be appropriate for the use of ergoCub wearables and robots, and which may not, to ensure that the technology is only deployed in situations where it meaningfully assists in achieving a legitimate objective. The team had identified a time pressure, which derives from the period of funding for the project. However, the mentors noted that it is extremely important that enough time is allowed for proper ethical oversight protocols to ensure appropriateness. This points to an issue in terms of how funding conditions and structures can often incentivise the rapid development of AI applications, which is sometimes in conflict with RAI and ethics considerations.

With respect to the principles of equity and non-discrimination, the mentors identified a need to focus on the training datasets used for ergoCub’s systems to ensure they are representative and will not result in differential or discriminatory outcomes for certain groups of users. Particularly with respect to wearables which collect biomechanical data from human users, to make predictions which may influence peoples’ job outcomes, it is extremely important to ensure gender sensitivity and responsiveness—that the system does not have differential impacts for women or other vulnerable or marginalised groups, such as the transgender community or persons with disabilities.

Finally, the mentoring process assisted the ergoCub team to develop RAI indicators that incorporate principles of decent work and job quality, given the technology’s potential to impact peoples’ experiences of work and their job outcomes. Particularly as it is likely that the wearables and humanoid robots are most likely to be adopted as a managerial intervention, given that there is often an uneven power relationship between employers and workers. Many workers vulnerable to musculoskeletal injuries are low wage, precarious workers. The RAI principles developed for the ergoCub project need to include ways to mitigate against the risks of increasing managerial surveillance leading to increasing job intensity, algorithmic management and discipline, the extraction and monetisation of worker-data (even if deidentified), as well as the risk of work fragmentation, increased precarity, or labour displacement as a result of humanoid robots taking on tasks and roles previously reserved for human workers.

Moreover, the mentors identified a need for the ergoCub team to give greater consideration to principles of human-centred AI, to ensure that their application is fully co-designed with its intended beneficiaries, and also to ensure that it enhances and augments human capabilities, rather than replacing them (i.e. that the humanoid robots are truly collaborative). To realise these outcomes, it is not sufficient to have a positive intent behind the technology, but there must be robust responsible AI indicators, guidelines and guardrails for the project.

RAI Deep Dive: Trustworthy AI Guidelines for Wearables in Healthcare and Industry

It was agreed that ergoCub would produce a document providing key indicators for addressing some of the above issues specifically for AI processing data from wearables leaving the RAI issues due to the use of humanoid robots aside. The mentors suggested that this document should contain (1) contextual information regarding the settings where the wearables would be deployed, and the intended beneficiaries, (2) a systematised risk analysis, including risk categories, and identification of high risks, and (3) proposed KPIs for measuring the quality of risk mitigation through product design, data policy, testing, etc. The resultant paper produced by the team (see Annex E) utilised

the United States' National Institute of Standards and Technology's Risk Management Framework, to "map, measure, and manage" RAI risks that might arise through the use of the wearables technology. The document covered privacy and data security, fairness, interpretability, accountability and acceptability.

Recommendations

Below are a series of policy recommendations arising from our experience during the SRAIS project, challenges encountered by participating teams, and outcomes of the mentoring process. These recommendations seek to move beyond existing RAI frameworks, by focusing on both policy and practical steps that would not only help to put RAI principles into practice, but would help validated RAI applications to gain a foothold and scale, in a landscape dominated by large players and varying adherence to responsibility. While the recommendations here are tailored towards AI teams and policymakers as the actors with the most direct responsibility for the impacts of AI systems, it is also acknowledged that there are a range of other stakeholders and actors who also have influence on the responsible scaling of AI systems, including users, funders, NGOs, etc.

For AI Teams

- **Responsibility throughout the AI lifecycle:** Ensure due consideration is given to best practice principles of responsible AI at each stage of the AI lifecycle (avoid trying to retrofit RAI at the point of deploying or scaling).
- **Narrow conceptualisation:** Prior to embarking on development of a solution, and in particular prior to engineering and testing, identify the specific purpose of the application as narrowly as possible.
- **Stakeholder identification and consultation:** Identify all relevant stakeholder groups, including: Primary and secondary users of your application (for instance, if an application is to be used by public health workers to support better health outcomes for patients, the primary users are public health workers, and the secondary users are patients); intended beneficiaries (who may differ from users) and any other groups who might be impacted by the deployment of the application. Undertake a needs analysis which includes consultation with all relevant stakeholders, to gain a clear understanding of their perspective on the problem, their needs, priorities, and concerns.
- **Contextual appropriateness:** Clearly identify the range of contexts in which the application might be deployed as it scales. For each context (and in consultation with local stakeholders), interrogate contextual appropriateness, with respect to linguistic, cultural, social, economic, ecological conditions and priorities.
- **Impact assessments:** Undertake rigorous impact assessments, which interrogate the potential impacts (including unintended impacts) of the application with regard to individual and collective rights, gender, decent work, democracy, and the environment.
- **Tailored RAI indicators:** Develop a set of comprehensive and clearly articulated RAI indicators specific to your project—across the various facets of RAI including human-centredness, fairness, equity, human rights, democracy and public good, by which the responsibility of the solution can continue to be monitored and evaluated.
- **Articulation of RAI adherence:** Strong adherence to frameworks of ethics and responsibility can set you apart in competing for public and private sources of funding—clearly articulate and emphasise the ways in which your project advances RAI in your engagement with funders and other stakeholders.

- **Controlled testing:** Undertake rigorous testing of your application in a controlled environment, which not only assesses aspects like performance, user experience and engagement, but also safety, adherence to RAI principles, and potential social and environmental risks and harms.
- **Ethical oversight of testing:** During testing (especially if testing with human subjects) ensure third-party accountability to a relevant oversight body, and ensure the implementation of best practice standards of research ethics including informed consent and data protection.
- **Use in human-decision making:** Carefully consider the opportunities and limitations of specific AI models to support human decision-making, and clearly communicate appropriate use and limitations to stakeholders.[4]
- **Data governance:** Establish robust frameworks and guidelines for the responsible governance of data prior to the commencement of any data collection.[5]
- **Safety guidance for users:** Ensure that users have clear and comprehensive information about the AI system, and particularly on its safe and appropriate use, as well as purposes that it should not be used for.[6]
- **Responsibility goes beyond intentions:** It is important to remain attentive to instances where an AI deployment that emerges from responsible intentions to contribute to the public good, may still inadvertently produce harmful outcomes. Even when a problem is clearly identified and a project has a clear social or environmental good, RAI related risks and harms can arise.[7]
- **Monitoring and evaluation:** After deployment of the solution, conduct regular monitoring and evaluation against the project's RAI indicators, including consultation with stakeholders, and make the results available to all stakeholders. In monitoring and evaluation of AI systems, include dedicated RAI KPIs, alongside evaluation of performance, user engagement, user experience etc.[8]

For Policymakers

- **Standards-setting:** Where applicable, pursue participatory standards-setting at the national and international levels of the responsible design, use, adoption and governance of AI. Work towards clear and operationalizable national and international frameworks and guidelines for RAI.
- **Regulate for competitiveness:** Where applicable, introduce appropriate regulation to challenge monopolisation of data assets, AI infrastructure and public-interest systems. Ensure sufficient competitiveness of commercial enterprise in the AI landscape.
- **Safe, public data infrastructures:** Consider public investment in and accountable governance of open, representative and safe datasets, as well as other infrastructures underpinning AI development and use—focusing on equitable access. Where appropriate, advance interoperability of digital and AI systems to enable wider access, collective innovation and knowledge sharing.

[4] This emerged from the experience of the Mexico team in thinking about how to apply their model to support decision-making affecting communities and individuals within variable contexts.
[5] This issue arose in the experience of the India team, which needed to collect (potentially sensitive) personal data from individuals and households as part of the development of their solution.
[6] This recommendation emerged from the experience of the Germany team, who the mentors prompted to produce guidelines for appropriate and safe use.
[7] For instance, the Italy team needed to mitigate the risk of the harmful surveillance of employees via their wearable intervention to monitor work-related health and safety risks.
[8] This recommendation stems particularly from the experience of the Canada team, in exploring ways to incorporate RAI principles into their system for evaluating chatbots.

- **Oversight and enforcement:** Explore options for establishing and sufficiently resourcing oversight and enforcement bodies for RAI.
- **Controlled testing:** Governments may consider playing a role in the facilitation of controlled testing and experimentation in AI, to ensure safety and responsibility prior to deployment and scaling. Regulate and set standards for AI experimentation and testing, with appropriate oversight.
- **Collective recourse:** There may be a need for governments to establish frameworks for recourse for community and society-level harms of AI systems, as well as individual-level harms.
- **Funding structures for RAI:** Consider how funding structures and conditions could better avoid disincentivizing abandonment or significant reform of a project. For instance, if an application is found in the engineering and testing phases to have responsibility concerns, consider ways to disincentivise the deployment of the application in that form.
- **Safety over speed:** Governments may consider exploring ways to ensure that funding structures and conditions do not unduly incentivise rapid deployment and scaling at the expense of responsibility.
- **Operating costs:** To scale RAI solutions, teams need support not only with developing and validating RAI, but also with ongoing running costs, and the costs of ensuring continued adherence to RAI principles. Policymakers may wish to take this into consideration in institutional procurement frameworks.
- **Innovation in AI evaluation:** Consider how frameworks for the systematic evaluation of the responsible behaviour and performance of AI systems can be supported by emerging technologies.[9]
- **Data collection frameworks:** Consider how governments may support and facilitate safe data collection practices during the earliest stages of innovation and development in AI.[10]
- **Public-private consultation and communication:** Consider possibilities for strengthening consultation between AI teams and government actors.[11]
- **Possibilities for partnership on RAI:** Consider possibilities that exist for governments to ensure responsible principles are at the heart of public-private AI partnerships.[12]
- **Responsibility with citizen voice platforms:** Where policymakers are considering the use of citizen voice/democratic participation platforms, consider how to incorporate transparency, explainability and responsible co-governance of platform infrastructures deployed in this process.[13]

Next steps

While the participating projects in the 2023 SRAIS project were highly varied with respect to their aims and context, they faced very similar challenges in integrating, 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

[9] This emerged from the experience of the Canada team.
[10] This recommendation was underscored by the experience of the India team.
[11] The Mexico team struggled to consult with different levels of government on their regulatory decision-making support tool, which highlighted the need for this recommendation.
[12] The Italy team's project was formed as a collaborative public-private partnership to address a public health issue.
[13] As highlighted by the experience of the Germany team.

complemented the capabilities of people rather than replacing them); and user education on the appropriate use and limitations of specific AI applications. Interrogating and mitigating these issues in a systemic and accountable way was identified as critical to validate the use of the various solutions, and to engage constructively with users, funders and regulators.

The mentoring process with the GPAI experts focused on assisting teams to develop clear frameworks for identifying, monitoring and mitigating various RAI risks and harms. The production of concrete RAI deep dives through the mentoring process is critical to enable continued progress on institutionalising RAI after the conclusion of the SRAIS project. At the time of writing the report, while the mentoring workshops have concluded, the GPAI experts and the participating teams are focusing on implementation and evaluation of RAI standards and principles. The GPAI experts will remain available to teams to guide them on implementation. In addition, an evaluation committee of GPAI experts will be established, to formally assess progress with the RAI plans and provide further recommendations to teams.

The SRAIS project will now be repeated on a yearly cycle. The first round in 2023 allowed us to gain insights into the range of challenges faced by AI teams and how the GPAI RAI Working Group can become ready to scale responsibly. Following the success of the first round of the project, the experts aim to expand it further in subsequent years, to reach more teams in more countries, and to systematically analyse 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.

Bibliography

Aghajari, Z., Baumer, E. P., Hohenstein, J., Jung, M. F., & DiFranzo, D. (2023). Methodological Middle Spaces: Addressing the Need for Methodological Innovation to Achieve Simultaneous Realism, Control, and Scalability in Experimental Studies of AI-Mediated Communication. Proceedings of the ACM on Human-Computer Interaction, 7(CSCW1), 1-28.

Anwar, M. A., & Graham, M. (2022). The digital continent: placing Africa in planetary networks of work (p. 288). Oxford University Press.

Benali, M., & Ghalfiki, J. E. (2021). Access to venture capital in Africa: the role of public institutions and corporate governance. International Journal of Business Performance Management, 22(2-3), 257-272.

Birhane, A. (2022). Automating ambiguity: Challenges and pitfalls of artificial intelligence. arXiv preprint arXiv:2206.04179.

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency (pp. 77-91). PMLR.

European Commission. (2021). Proposal for a regulation of the European Parliament and of the Council: Laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative Acts. 2021/0106 COD. https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF.

Gallemore, C., Nielsen, K. R., & Jespersen, K. (2019). The uneven geography of crowdfunding success: Spatial capital on Indiegogo. Environment and Planning A: Economy and Space, 51(6), 1389–1406. https://doi.org/10.1177/0308518X19843925.

GPAI. (2020). Responsible Development, Use and Governance of AI Working Group Report, Report, November 2022, Global Partnership on AI, Montréal.

GPAI. (2022a). Data Governance Working Group – A Framework Paper for GPAI's Work on Data Governance 2.0, Report, November 2022, Global Partnership on AI, Paris.

GPAI. (2022b). AI for Fair Work: AI for Fair Work Report, November 2022, Global Partnership on AI.

GPAI. (2022c). Data Justice: A Primer on Data and Economic Justice, Report, November 2022, Global Partnership on AI.

Gray, M. L., & Suri, S. (2019). Ghost work: How to stop Silicon Valley from building a new global underclass. Eamon Dolan Books.

Grossman, R. L., Heath, A., Murphy, M., Patterson, M., & Wells, W. (2016). A case for data commons: toward data science as a service. Computing in science & engineering, 18(5), 10-20.

Harrison, R. T., Yohanna, B., & Pierrakis, Y. (2020). Internationalisation and localisation: Foreign venture capital investments in the United Kingdom. Local Economy, 35(3), 230–256. https://doi.org/10.1177/0269094220924344.

Howson, K., Johnston, H., Cole, M., Ferrari, F., Ustek-Spilda, F., & Graham, M. (2022a). Unpaid labour and territorial extraction in digital value networks. Global Networks.

Howson, K., Ferrari, F., Ustek-Spilda, F., Salem, N., Johnston, H., Katta, S., ... & Graham, M. (2022b). Driving the digital value network: Economic geographies of global platform capitalism. Global Networks, 22(4), 631-648.

Hummel, P., Braun, M., Tretter, M., & Dabrock, P. (2021). Data sovereignty: A review. Big Data & Society, 8(1), 2053951720982012.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature machine intelligence, 1(9), 389-399.

Katta, S., Badger, A., Graham, M., Howson, K., Ustek-Spilda, F., & Bertolini, A. (2020). (Dis) embeddedness and (de) commodification: COVID-19, Uber, and the unravelling logics of the gig economy. Dialogues in Human Geography, 10(2), 203-207.

Kukutai, T., Carroll, S. R., & Walter, M. (2020). Indigenous data sovereignty.

Kumar, A. (2020). Monopsony capitalism: Power and production in the twilight of the sweatshop age. Cambridge University Press.

Langley, P., & Leyshon, A. (2017a). Platform capitalism: the intermediation and capitalization of digital economic circulation. Finance and society., 3(1), 11-31.

Langley, P., & Leyshon, A. (2017b). Capitalizing on the crowd: The monetary and financial ecologies of crowdfunding. Environment and Planning A: Economy and Space, 49(5), 1019–1039. https://doi.org/10.1177/0308518X16687556.

Lehne, M., Sass, J., Essenwanger, A., Schepers, J., & Thun, S. (2019). Why digital medicine depends on interoperability. NPJ digital medicine, 2(1), 79.

Lewis, P., Davies, H., O'Carroll, L., Goodley, S., Lawrence, F. (2022, July 11). The Uber whistleblower: I'm exposing a system that sold people a lie. The Guardian.

Li, F-F., & Etchemendy, J. (2018). Introducing Stanford's Human-Centered AI Initiative. https://hai.stanford.edu/news/introducing-stanfords-human-centered-ai-initiative
Munn, L. (2022). The uselessness of AI ethics. AI and Ethics, 1-9.

Noble, S. U. (2018). Algorithms of Oppression: How Search Engines Reinforce Racism. New York University Press.

OECD. (2019). Recommendation of the Council on Artificial Intelligence. OECD/LEGAL/0449.

O'Hara, K. (2019). Data trusts: Ethics, architecture and governance for trustworthy data stewardship.

Pasquale, F. (2015). The black box society: The secret algorithms that control money and information. Harvard University Press.

Peck, J., & Phillips, R. (2020). The platform conjuncture. Sociologica, 14(3), 73-99.

Prabhakaran, V., Mitchell, M., Gebru, T., & Gabriel, I. (2022). A human rights-based approach to responsible ai. arXiv preprint arXiv:2210.02667.

Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., & Denton, E. (2020, February). Saving face: Investigating the ethical concerns of facial recognition auditing. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (pp. 145-151).

Sadowski, J. (2019). When data is capital: Datafication, accumulation, and extraction. Big data & society, 6(1), 2053951718820549.

Schiff, D., Borenstein, J., Biddle, J., & Laas, K. (2021). AI ethics in the public, private, and NGO sectors: A review of a global document collection. IEEE Transactions on Technology and Society, 2(1), 31-42.

Shneiderman, B. (2021). Human-centered AI. Issues in Science and Technology, 37(2), 56-61.

Tubaro, P., Casilli, A. A., & Coville, M. (2020). The trainer, the verifier, the imitator: Three ways in which human platform workers support artificial intelligence. Big Data & Society, 7(1), 2053951720919776.

UNCTAD. (2019). Digital Economy Report 2019: Value Creation and Capture: Implications for Developing Countries.

UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence. Paris.

Xu, W. (2019). Toward human-centered AI: a perspective from human-computer interaction. interactions, 26(4), 42-46.

Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power: Profile books.

ANNEX

Annex A: RAI Deep Dive — Team from Canada: Formulating Responsible AI Pillars for Conversational Systems Analytics

By: **Aaqib Azeem**, Wysdom.ai
Date: October 12th, 2023

In our rapidly evolving technological landscape, the integration of artificial intelligence (AI) and natural language processing has given rise to intelligent chatbots and voicebots. These AI-driven conversational agents are now pervasive across industries, offering improved customer service, operational efficiency, and user engagement. However, as these AI technologies continue to shape our interactions, the imperative to cultivate responsible AI solutions becomes increasingly evident.

Responsible AI isn't merely a matter of ethics; it's a commitment to ensuring that AI benefits society at large and mitigates the potential for harm. As AI systems become more integrated into our daily lives, the consequences of irresponsible AI can be significant. Ensuring the responsible development, deployment, and usage of AI technologies is not just a moral obligation; it's essential for building trust and sustaining the growth of AI-driven industries.

Within the realm of responsible AI, certain topics take center stage due to their profound impact on the ethical and effective use of AI-powered conversational agents. These topics encompass various facets of responsible AI, from fairness and transparency to security and compliance. Addressing these challenges isn't just a matter of mitigating risks; it's a strategic imperative for organizations looking to provide AI solutions that are dependable, secure, and ethically sound.

In this context, this paper explores 13 key pillars in developing responsible AI chatbot and voicebot analytics tools. It outlines potential mitigation strategies tailored to address these specific topics of responsibility, all of which are critical for shaping the future of AI-powered conversational experiences.

RAI Pillars for Conversational AI

1) Scalability and Performance
Description: As the volume of conversations surges, ensuring that the analytics platform remains high-performing and scalable under peak loads is crucial.
Mitigation: Wysdom.ai plans to employ cloud-based solutions and autoscaling to dynamically allocate computational resources based on demand. Utilizing efficient data storage and processing capabilities will enable the platform to handle growing data loads efficiently while maintaining real-time analysis.
Measurement (KPI): Monitor system responsiveness and efficiency, track server utilization, and measure response times.

2) Bias Detection and Mitigation
Description: Detecting and eliminating bias in chatbot responses to ensure fairness and inclusivity.
Mitigation: Wysdom.ai plans to develop advanced algorithms that identify and rectify biased responses. Continuous monitoring of chatbot interactions and using diverse training data can help reduce bias in responses, enhancing the system's fairness.

Measurement (KPI): Measure the reduction in biased responses, track user feedback on fairness.

3) User Privacy and Data Protection
Description: Protecting user data and ensuring compliance with data protection regulations.
Mitigation: Wysdom.ai plans to implement robust data encryption, access controls, and audit trails to secure user information. Transparency in data handling practices and obtaining informed user consent will build trust and address privacy concerns.
Measurement (KPI): Monitor compliance with data protection laws and track the number of data breaches.

4) Transparency and Explainability
Description: Making the chatbot's decision-making process transparent to users.
Mitigation: Wysdom.ai plans to develop explainability mechanisms to provide insights into how the chatbot arrives at its responses. Utilizing interpretable AI models and visualizations can help users understand the rationale behind the chatbot's actions.
Measurement (KPI): Gather user feedback on the clarity of responses and measure user satisfaction with explanations.

5) Ethical interactions
Description: Ensuring that chatbots interact with users in a responsible and appropriate way.
Mitigation: Wysdom.ai plans to implement ethical guidelines and rules within the chatbot's interaction behavior.
Measurement (KPI): Track instances of unethical interactions and monitor user complaints related to ethics.

6) Security
Description: Protecting chatbot systems from security threats, including data breaches and malicious attacks.
Mitigation: Wysdom.ai plans to invest in robust security measures, including regular security audits, intrusion detection systems, and penetration testing. Secure coding practices and user authentication mechanisms are essential for protecting against threats.
Measurement (KPI): Monitor the reduction in the number of security incidents and the effectiveness of security measures.

7) Multilingual and Multicultural Considerations
Description: Handling various languages and cultural nuances effectively in diverse regions.
Mitigation: Wysdom.ai plans to employ multilingual AI models and cultural awareness modules to ensure accurate and culturally sensitive interactions. Collaboration with linguists and cultural experts can enhance the chatbot's language and cultural capabilities.
Measurement (KPI): Evaluate multi language accuracy and user feedback related to cultural sensitivity.

8) User Education
Description: Ensuring that users are aware of chatbots' capabilities and limitations..
Mitigation: Wysdom.ai plans to provide user-friendly guides and tutorials to educate users about the chatbot's functionalities. Clear communication about when users are interacting with a chatbot versus a human agent can manage user expectations effectively.
Measurement (KPI): Assess user awareness through surveys and user knowledge retention.

9) Regulatory Compliance

Description: Staying compliant with data protection and AI ethics regulations.
Mitigation: Wysdom.ai plans to maintain a dedicated compliance team that continuously monitors and adapts to evolving regulations. Conducting regular compliance audits and maintaining comprehensive records will demonstrate commitment to regulatory adherence.
Measurement (KPI): Track regulatory compliance status and the absence of legal actions.

10) User Consent
Description: Obtaining informed consent for data collection and allowing the user to opt out of interacting with the AI system and speak to a human instead.
Mitigation: Wysdom.AI plans to implement user-friendly consent mechanisms and seamless transition to human agents.
Measurement (KPI): Monitor user consent rates, track the number of successful transitions to human agents, and gather feedback on consent processes.

11) Copyright Protection
Description: Protecting intellectual property and adhering to copyright laws, especially with regard to generated or utilized content.
Mitigation: Wysdom.ai plans to implement content filtering mechanisms to prevent the generation or use of copyrighted materials without proper authorization. Legal consultation can help navigate complex copyright issues.
Measurement (KPI): Monitor copyright-related complaints and legal actions.

12) A/B Testing and Evaluation
Description: Continuous evaluation of chatbot performance for identifying and addressing issues.
Mitigation: Wysdom.ai plans to conduct regular A/B testing and gather user feedback to assess and improve the chatbot's behavior. Data-driven insights can drive continuous enhancements.
Measurement (KPI): Compare behavioral metrics between versions and assess user satisfaction across all pillars.

13) Sustainability
Description: Considering the environmental impact of AI systems.
Mitigation: Wysdom.ai plans to optimize energy consumption during both training and usage phases by utilizing energy-efficient hardware and adopting green computing practices. Sustainability initiatives can align with responsible AI efforts.
Measurement (KPI): Measure energy consumption reduction and environmental footprint.

Summary

In today's fast-paced digital landscape, virtual agent analytics solutions have become indispensable for businesses. Ensuring that these solutions tackle responsible AI challenges in chatbot and voicebot analytics effectively is of utmost importance. By putting the suggested strategies into practice, companies can make certain that their advanced chatbot and voicebot analytics tools not only work seamlessly at scale but also operate in an ethical and legally compliant manner. This not only builds trust but also ensures that AI serves society and businesses in a responsible and ethical way. It is a path to providing innovative, secure, and ethically sound AI-driven conversational experiences that contribute positively to society and businesses.

Annex B: RAI Deep Dive — Team from India: Challenges and Strategies for Responsible Data Collection and Management for Future AI Healthcare Applications

By: **Suresh Munuswamy**, Public Health Foundation of India, Indian Institute of Public Health - Hyberabad and Hi Rapid Lab Private Limited; **Shreya Ramakrishnan**, Indian Institute of Public Health – Hyderabad; **Bhavya Tanuku**, Hi Rapid Lab Private Limited; **Priyadarsini Suresh**, Hi Rapid Lab Private Limited.
Date: October 17th, 2023

COMPREHENSIV is a customizable survey and service digital platform for at home primary health care in resource limited lingusitically diverse communities. It is designed to source field data, understand the data and deliver the customized interventions in real time. COMPREHENSIV is designed around a question unit with 5 customizable parts: (1) icon based question to overcome language; (2) image capture with gps location and metadata to provide context and traceability; (3) further icons as response options to maintain language neutrality; (4) contextually linked to pre-programmed interventions like short animated videos or text in real time; (5) revisit date prompt. A trained healthcare personnel oversees the process
In view of the broad scope of the problem, large amounts of traceable personal private data being acquired, and the recent Indian digital personal data protection act 2023 (1), it was decided to examine and evolve this project from the view of responsible data acquisition, storage, use, and sharing. How proposed AI modules would be developed and validated from the data was also discussed, but in view of the multiplicity of potential applications, this was a secondary RAI deep dive.

Key Challenges and Potential Solutions

1) Ethical data acquisition. COMPREHENSIV has been approved by a registered internal Human Research Ethics Committee of the Indian Institute of Public Health. The challenge is in its transition as a research mode project with the proposed output being a new version of an existing application, now enhanced with embedded AI corresponding to the functionalities described above, to an academic start-up. The solution is in a comprehensive well-documented consent process. Module wise informed consent options will be provided where the service provider can explain at each step. People can opt out of the entire service or part of the service depending on their requirements. Multiple opt out options, right to access, right to be forgotten, right to be informed, record of processing will be provided to people at different stages of the service.

2) Responsible Data Storage and Use. The challenge is in creating appropriate time-bound and purpose-limited governance frameworks compatible with new legal requirements brought about by the DPDPA 2023. The solution decided upon is to first create a responsive oversight system that can guide the investigators beyond an ethical committee alone. A Data Governance Committee (DGC) with independent members and a framework will be created to oversee and manage: ethical and consent approved data sourcing; clear data traceability and identification; access management; data analysis, service based on data; data sharing; purpose limitation. Every data point collected and processed will be documented, justified and limited to specific pre agreed use. Post the purpose all data will be reviewed, de identified and deleted if agreed upon. Among initial steps determined to be necessary for legal compliance and for scalable responsible AI, a personal portal will be created where a participant can view all the personal data that is sourced; purpose the data has been

collected, applications it is used for; along with a consent manager where if there is a new purpose the data needs to be used, he or she can consent for. The consent manager will also have an option to hold normal data, option to hold health related data without any service, any analysis, delete data – in which case a de-identified version will be created immediately and shared to DGC for review. If no request is received, then the data will be held for a period as per the country's service policy.

3) Responsible algorithm development and data sharing practices. Development of ML algorithms will start when the project reaches a threshold of 25% data sourcing and validated prospectively. There were no existing plans to share data with others and all development was planned inhouse. If data needs to be shared for any reason, as in third party validation, a subset of data randomly segregated at different levels, de-identified (safe harbour method) will be created, shifted to a separate in-house server and access shared.

Conclusions

Focused academic AI start-ups may emerge from large open-ended academic research projects that collect large digital personal datasets for public health RAI deep dive. Such a transition requires additional consideration about data-governance and responsible AI, as highlighted above, to be compliant with legal and responsible AI requirements.

Reference

Digital Personal Data Privacy Act, Ministry of Electronics and Information Technology, Government of India, 2023 accessed from (https://www.meity.gov.in/writereaddata/files/Digital%20Personal%20Data%20Protection%20Act%202023.pdf)

Annex C: RAI Deep Dive — Team from Mexico: Key Challenges and Potential Solutions for Developing a Responsible Decision Support System Based on a Classification Model

By: **Eduardo Ulises Moya-Sanchez**, **Abraham Sánchez**, **Raul Nanclares Da Veiga**, **Alexander Quevedo Charon**, **T. Camacho**, **Ulises Jiménez Pelagio**, **A. Piña**, Gobierno de Jalisco, the U.S. Forest Service and the National Forest Commission of Mexico (CONAFOR) and the US Forest Service (USFS).
Date: October 9th, 2023

According to [1] the leading causes of deforestation in Mexico are conversion into pastures for livestock farming, agriculture (avocado orchards and agave crops), and illegal logging. Reducing the deforestation areas is imperative to preserve biodiversity, water and mitigate the climate change impact [2]. This problem could be more critical if you consider trade-offs between agricultural production and forest conservation.

An early warning system (EWS) for deforestation can be an effective tool for the timely detection of forest disturbance. Our EWS for deforestation uses satellite imagery from ESA's Copernicus program. It uses several AI models with the object-based classification (OBIA) method: one to generate the forest mask (the area we consider forest), and another to detect deforestation.

The results from our EWS are used as part of the knowledge database for an Environmental Decision Support Systems (EDSS), powered by AI, to improve decision making in the context of tackling deforestation. In this context, we present some key challenges and potential solutions based on our experience going from the EWS to an EDDS.

Key Challenges and Potential Solutions

1. **AI model robustness.** The model robustness could be seriously affected by the data variation. Mexico is a megadiverse country and thus data from different ecoregions can have huge variability. Moreover, some ecosystems are also affected by seasonality effects that can have a major impact in model accuracy. Class separability can be improved if an AI model is trained for each ecoregion [4]. In our case, the three different monitoring areas (two in the Yucatan Peninsula and one in the State of Jalisco) use three different models for forest disturbance detection. We still need to improve our detections in regions where season alternance has a bigger effect, like the area in the state of Jalisco. In the future we would like to explore the inclusion of weather data (e.g. precipitation) to increase the robustness of the models against seasonality effects.

2. **EDSS Deployment.** In our opinion the main aspects to consider in this regard are: Use of ease and a dispute mechanism. In our experience with previous EDSS the user interface for web platforms has to be very simple and in some cases, elderly population will need in person attention to be able to use the EDSS. We also need to consider the different languages spoken in the area and adapt the user interfaces accordingly. The dispute mechanism will allow citizens affected by the EDSS resolutions to present a complaint in which they can fill in the reasons they think the decision is wrong. This implies the EDSS will need a team of reviewers to review the allegations and act accordingly. If a complaint is accepted the data in the EDSS has to be updated (usually the forest mask will be modified).

On the AI side, it is necessary to evaluate the performance of the models, with the proper metrics, after the deployment and keep updating them to ensure they are functioning adequately.

3. **Multi-stakeholder project**. In deforestation tackling projects we usually have to deal with actors in different government levels, farmers, loggers, civil society organizations, indigenous communities and so on. Government organizations usually demand a very high level of accuracy in the detection so they can take legal actions on the actors causing deforestation. This will usually require human validation of the deforestation alerts and the use of auxiliary data to ensure the deforestation is illegal. Moreover, different government institutions will require different products generated by the EDSS (e.g. a general report on deforested area and where it is occurring at municipality level or a very precise report of the deforested area with coordinates in a construction box format). Thus the outputs of the EDSS have to be tailored to the needs of the different stakeholders. We need to consider all these issues but without compromising the timeliness of the alerts.

4. **Socio-environmental implications**. In some cases the use of remote sensing can forgo other aspects which are almost impossible to infer from simple observation of satellite imagery. These aspects are usually related to territorial planning and local agricultural practices which usually require other auxiliary data or expert knowledge to avoid wrongly accusing the population/communities living in the area. This can be alleviated by refining the forest masks using data from land management plans, the knowledge of the local authorities or working with the communities or social organizations in the monitoring area. Organizing workshops with key stakeholders is a good way of solving this problem.

References

[1] Goldstein, A., Erickson, H., Gephart, N., & Stevenson, S. (2011). Evaluation of land use policy and financial mechanisms that affect deforestation in Mexico.
[2] Knoke, T., Hanley, N., Roman-Cuesta, R. M., Groom, B., Venmans, F., & Paul, C. (2023). Trends in tropical forest loss and the social value of emission reductions. *Nature Sustainability*, 1-12.
[3] CONAFOR (2020). Estimación de la tasa de deforestación en México
para el periodo 2001-2018 mediante el método de muestreo. Documento Técnico.
Jalisco, México
[4] Tulbure, M. G., Hostert, P., Kuemmerle, T., & Broich, M. (2022). Regional matters: On the usefulness of regional land‐cover datasets in times of global change. *Remote Sensing in Ecology and Conservation*, *8*(3), 272-283.

Annex D: RAI Deep Dive — Team from Germany: Guidelines for a Responsible Use of AI Voice- and Chatbot Technology for People-Centered Participation and Feedback

By: **Jascha Stein**, Particip.ai and People-Centered Internet; **Ronald Strauss**, Particip.ai.
Date: October 9th, 2023

Particip.ai provides a state-of-the-art Participation and Feedback Platform, aiming to transform the landscape of user engagement and data collection. For the second time, the

UNESCO AI Group listed Particip.ai as one of the GLOBAL TOP 100 AI solutions[14] for achieving the United Nations 17 Sustainable Development Goals with status Extremely Promising". As we operate across multiple jurisdictions and industries, the complexity of adhering to varied regulatory landscapes becomes evident. The nuances of international laws, coupled with the diverse user expectations, underscore the importance of forging and maintaining trust. Trust is a cornerstone of our relationship with both our clients and the end-users. These Guidelines, therefore, not only serve for ethical and transparent use but also as a testament to our commitment to fostering trust, ensuring responsible and meaningful interactions on our platform. In this regard, as a customer of the platform, we encourage you to follow these guidelines to ensure a responsible use of the platform.

1. **Transparency and Explainability**
   - 1.1. Always inform end-users when they are communicating with an AI system.
     Example Dialogue: "Hello! I'm a bot here to assist you. How can I help you?"
   - 1.2. Provide users an easy option to transition from the AI voicebot to a human representative.
     Example Dialogue: "Would you like to continue with me, or would you prefer to speak with a human representative?"
   - 1.3. Make only accurate and factual statements regarding the capabilities of the AI voicebot.
     Example Dialogue: "I can help with basic inquiries and troubleshooting. For complex issues, I can connect you to a specialist."
   - 1.4. If required by the user, provide explanations of the voicebot response.
     Example: "The data I provide is obtained by this computation …"

2. **Data Privacy, Security, and Compliance:**
   - 2.1. Obtain explicit consent from users prior to recording, processing, or storing voice interactions.
     Example Dialogue: "Do you mind if I record this conversation for quality and training purposes?"
   - 2.2. If asked or legally necessary, clearly communicate the intent behind data collection.
     Example Dialogue: "We collect this data to improve our services and provide you with better support. We won't share it with third parties without your consent."
   - 2.3. If asked or legally necessary, inform users about where their data is processed.
     Example Dialogue: "Your data is processed in our European data centers, adhering to GDPR regulations."

3. **Ethical Interactions and Content Creation:**
   - 3.1. As the architect of the voicebot's dialogues and interactions, ensure they are respectful.
     Example Dialogue: "Thank you for your input. I'm here to help and respect your preferences."
   - 3.2. Design voicebot responses to handle responsibly any aggressive or negative user utterances.
     Example Dialogue: "I'm not able to fulfill your request, please rephrase or let me know if you'd prefer a human agent."
   - 3.3. Design the voicebot in a way that avoids displaying harmful, or offensive content.
     Example Dialogue: "I strive to provide accurate and helpful information. If there's an error, please let us know so we can rectify it."

[14] https://ircai.org/top100/entry/particip-ai-ccc-crises-contact-center-and-helpdesk-for-people-in-need-or-with-disadvantages/, accessed November 2023

3.4. Design the voicebot without human-like responses that show emotions, such as: *“I’m sorry…”, “I feel bad”, “I understand how you feel”.*

4. **Feedback and Accountability:**
   4.1. Incorporate mechanisms for users to report biases or errors.
   Example Dialogue: "If you find any inconsistencies or biases in my responses, please provide feedback by [feedback process]"
   4.2. Be accountable for issues or damages.
   Example Dialogue: "Thank you! This will be reviewed and addressed promptly by [name of the responsible organization]"

5. **Continuous Review and Improvement:**
   5.1. Periodically revisit and refine interactions.
   Example Dialogue: "Thank you for the feedback. We continuously work on improving our system to serve you better."
   5.2. Leverage feedback mechanisms.
   Example Dialogue: "Your insights are valuable to us. We use them to make our service more efficient and user-friendly."

Conclusion

Embracing these guidelines is not just a commitment to ethical technology deployment but a promise to the users. In the event of a non-compliance with these guidelines, we reserve the right to implement appropriate dialogs by default, which are customizable to a certain extent. Adherence ensures that the AI voicebot platform serves its audience responsibly, transparently, and efficiently.
These are guidelines but stay abreast of specific regulations and best practices for a responsible use of voicebots. Also consider the constant evolution of the technology which may raise additional issues which should be handled in a responsible way.

References

[1] Overview and inspiration: World Economic Forum: Chatbots RESET, https://www.weforum.org/reports/chatbots-reset-a-framework-for-governing-responsible-use-of-conversational-ai-in-healthcare, accessed September 2023

[2] Framework for guidelines: Microsoft RAI Impact Assessment Template, https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-RAI-Impact-Assessment-Template.pdf, accessed September 2023

Annex E: RAI Deep Dive — Team from Italy: Trustworthy AI Guidelines for Wearables in Healthcare and Industry

By: **Daniele Pucci**, Istituto Italiano di Tecnologia; **Lorenzo Rapetti**, Istituto Italiano di Tecnologia; **Enrico Valli**, Istituto Italiano di Tecnologia; **Cristina Di Tecco**, Istituto Nazionale per l'Assicurazione contro gli Infortuni sul Lavoro; **Matteo Ronchetti**, Università di Pisa; **Francesco Draicchio**, Italian National Research Council; **Giovanna Tranfo**, Istituto Nazionale per l'Assicurazione contro gli Infortuni sul Lavoro.
Date: October 4th, 2023

The Need of Wearables in Industry and Healthcare

Every year, Europe spends 3.3% of its GDP to prevent and treat occupational injuries [1][2]. Despite these efforts, 60% of European workers experienced a musculoskeletal disease (MSD) at least once throughout their life, thus contributing to a total of 120M people dealing with chronic MSD diseases [3][4]. And this condition is not of Europe only. Worldwide, work-related MSD is the most common occupational health disease in industry [5][6], which calls for the envisioning of an industry digital transformation aiming to reduce the impact of work-related musculoskeletal diseases and disorders.

Besides the significant burden of MSD on economies, it is essential to address not only prevention but also rehabilitation. Traditional rehabilitation methods, while valuable, often involve time-consuming physical therapy sessions, medication, or even surgery in severe cases. These approaches are effective but may not always be accessible, convenient, or tailored to individual needs. As a result, there is a growing recognition of the need for more innovative and patient-centric rehabilitation solutions. The need for effective rehabilitation strategies and the integration of innovative technologies in this context cannot be overstated.

The urgent need to envision new systems to monitor humans at work to improve ergonomics and/or during rehabilitation processes to streamline recovery time paves the way for wearable devices [7]. Smartwatches, sensorised shoes, smartclothes via e-textiles, and smartglasses are only a few examples of wearables. When targeting industry and healthcare applications, for example, the ergoCub https://ergocub.eu/ project develops sensorised shoes and suits complemented by the processing AI for human health indicators, which are smart wearables also referred to as iFeel https://ifeeltech.eu/. So, a wearable device is a compact and typically body-worn technology designed to collect and analyze various physiological and environmental data. AI is integrated into wearables to process multimodal information, enabling them to retrieve and analyze relevant data efficiently. This empowers wearables to provide personalized insights, adapt to user behaviors, and enhance their overall functionality. For example, wearable sensors can promote telemonitoring and telerehabilitation reducing the pressure on health systems and allowing a more continuous follow up in the rehabilitation process. Telemonitoring and telerehabilitation applications can be built upon wearable sensors ranging from the simple collections of vital parameters to continuous monitoring of subject's movement and posture, through distributed inertial sensors and sensorized garments, with the aim of enhancing ergonomics, optimizing rehabilitation, and promoting overall wellbeing [8]. Real-time data aid in preventing the recurrence of MSDs, by providing feedback to workers about their posture to minimize the risk of injury and provide healthcare professionals with data to assess rehabilitation progress and treatment plans.

State-of-the-Art on Standardization and Trustworthy AI for Wearables

Despite wearables devices nowadays proliferate in various market segments with the goal of monitoring different indicators of the human physiological and physical state, standardizations for their hardware architectures as well as guidelines for the trustworthy AI applications that process wearable data are still lagging [9]. A series of high-level guidelines to promote Trustworthy AI have been presented by an independent group of experts in the *Ethics Guidelines for Trustworthy AI* document of the European Commission [10]. In this study, Trustworthy AI is approached with three different components: lawful, ethical, and robustness. Despite the proposed guidelines are pivotal indications for the development of AI systems, specific applications still need to be defined, and how to apply these guidelines to the specific case of wearables remains an open problem.

An effort to standardize terminologies and high-level architectures for wearable devices was recently made by the IEEE Consumer Technology Society when in February 2022 approved the IEEE Standard for Wearable Consumer Electronic Devices, also referred to as IEEE Std 360TM-2022 [11]. The standard, however, mainly focuses on the technical requirements and testing procedures for different natures of wearables, by leaving aside how applications - that process data from the wearable devices - shall deal with trustworthy AI principles. In other words, the IEEE Std 360TM-2022 deals only with the first pillar of trustworthy AI, which can be viewed as the framework for managing AI risks composed of three pillars: *technical requirements* (reliability, robustness, generalizability, resilience, security), *socio-technical requirements* (interpretability, explainability, free of bias, privacy), and *social requirements* (transparency, accountability, fairness).

The process of standardizing trustworthy AI systems requires a common ground. An effort to define common terminology to describe the design, engineering and usage of AI systems is presented in the *ISO/IEC 23053, Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)* [12]. How to develop and manage AI systems, however, is still an open problem - as per today, October 2023 - being the *ISO/IEC DIS (Draft International Standard) 42001 Artificial Intelligence Management System* still under development [13]. Analogously, third parties that aim at the certification of AI systems still struggle to obtain a defined landscape where to operate – see, e.g., the *NLP report MS 45,* titled *Certification of Machine Learning Applications in the Context of Trustworthy AI with Reference to the Standardization of AI Systems* [14]. This work-in-progress concerning the state of the art on trustworthy AI systems standardization and certification processes clearly also impacts AI software processing wearable data.

The NIST AI Risk Management Framework (AI RMF 1.0)

In January 2023, the National Institute of Standards and Technology (NIST) published the *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* [15], whose goal is "*to offer a resource to the organizations designing, developing, deploying, or using AI systems to help manage the many risks of AI and promote trustworthy and responsible development and use of AI systems*". It is a voluntary, rights-preserving, non-sector-specific, and use-case agnostic framework that provides organizations and individuals with methodologies that aim at increasing trustworthiness of AI systems. The core of the framework is composed of four *horizontal functions*: **govern**, **map**, **measure**, and **manage**. The **govern** function aims at establishing a high-level culture for AI risk management. The **map** function represents all activities to recognize and identify risks. The **measure** function serves to assess and track risks along the AI system lifecycle, often by defining Key Performance Indicators (KPIs). The **manage** function prioritizes risks to address depending on the projected impact. So, these four functions are high-level activities that either individuals or organizations may implement for increasing the awareness of the AI risks during the entire process of designing, developing, and deploying AI systems.

The four sector-agnostics horizontal functions abstract the specificities that each application of the AI system possesses depending on the use-case at hand. In the AI RMF 1.0, the application-specific needs required to increase trustworthiness of the application-specific AI system are referred to as **AI RMF Profiles,** which are *"implementations of the AI RMF functions, categories, and subcategories for a specific setting or application based on the requirements, risk tolerance, and resources of the Framework user"*. Profiles describe how risks can be managed at various stages of the AI lifecycle in the specific sector. What follows lays down trustworthy guidelines for wearable AI applications following the AI RMF Profile structure in two scenarios: industry and healthcare.

The AI RMF Wearable Profile for Industry and Healthcare

Govern

AI risk management must be governed by the executive board of the company/institute in which the wearable system is developed. The govern function shall include activities related to: i) explanation and enforcement of the risk management culture, ii) application of international recognized standards in the whole AI system lifecycle, iii) realization of internal procedures for the compliance to the applicable international norms, iv) definition and implementation of policies, regulations, and principles for the responsible development and deployment of the product, v) definition of the working groups and relative responsibilities in terms of AI risk generation, vi) management and definition of periodic review and assessment of the implemented policies, vii) investigation of new procedures and policies that are required for the development and deployment of future products. While the govern sub-functions are horizontal with respect to the activities of the company/institute, their application shall be tailored to the specific product lifecycle and to the development teams involved in the product life cycle. This allows a better structure of the internal processes by specifying customized development frames for each working group with respect to AI risk responsibilities.

For example, the lifecycle of a potential wearable medical device includes several groups of actors: i) hardware and firmware developers, ii) Product designers, iii) AI application layer software developers, iv) Information and Communication Technology (ICT) experts, v) GDPR responsible, vi) Principal Investigators (PI), vii) post-processing engineers and researchers, viii) clinical trials managers. Each group shall be properly instructed for AI risks that might impact their specific responsibilities. Hardware and firmware developers, product designers, and application layer

software developers shall get acquainted with risks related to *acceptability*. ICT experts, GDPR responsible, and post-processing engineers and researchers shall be informed about risks related to *privacy and data security*. PI and clinical trial managers shall be instructed for *bias and fairness* risks. Post-processing engineers and researchers, clinical trials managers, and PI shall be acquainted with risks related to *interpretability and accountability*.

Below are details about the functions *Map*, *Measure*, and *Manage* concerning the major risk categories for wearables in industry and healthcare: **i) *Privacy and Data Security*, ii) *Fairness*, iii) *Interpretability, iv) Accountability*, v) *Acceptability***.

Privacy and Data Security

Map – When sensitive and personal data are collected by wearable devices, risks are related to data privacy and ownership. Privacy violations such as data breaches and unauthorized access to databases must be prevented. Additionally, determining the rightful owner of the data generated by wearable systems is essential, data ownership affects the responsibilities related to data management and protection.

Measure – *Data Breach Rate*: *number of data breach occurrences measured in a specific period of time.* Employ specialized software tools to assess system vulnerability through techniques like penetration testing and vulnerability assessments.

Manage – Privacy-enhancing technologies and data minimization methods, such as de-identification and aggregation for specific model outputs, play a crucial role in crafting AI for wearable systems helping safeguard sensitive data throughout its lifecycle. Moreover, the establishment of a robust framework for informed and explicit consent, also defining data ownership, is essential. Therefore, a safe escalation procedure should be available for the subjects to address cases of consent extortion.

Fairness

Map – Data collection, especially during clinical studies with wearables, may suffer from limited representation of the real application field population. This limitation can lead to the development of biased models, resulting in algorithmic bias and erroneous outcomes. The application of wearables should reasonably address physical differences between people (e.g., sizes, gender related) to the same extent as other personal elements of the work environment (e.g., protective garments and devices).

Measure – *Data Bias: statistical metrics (e.g., mean and variance) of the parameters (e.g., height and weight of subjects) governing the generation of the anthropometric models.* Evaluating data bias requires a thorough diversity analysis of the data source used in system training.

Manage – To address biases and promote fairness it is important to favor balanced statistical representation (e.g., prioritizing diversity, equity, and inclusion). When this is not possible, bias mitigation can be enforced via mathematical models. Additionally, mitigating harmful biases does not automatically ensure fairness, as wearable systems with balanced predictions across demographic groups may still be inaccessible to individuals with disabilities or those affected by the digital divide.

Interpretability

Map – AI algorithms developed for wearable solutions may produce results that are complex and difficult to interpret. This risk may lead to making wrong decisions, which might be critical when it concerns with subjects' health.

Measure – *Quantitative Assessment: statistical metrics (e.g., percentage) of comprehension level of the AI-based wearable system outcomes.* Automated satisfaction surveys can assess the capability

of the user to interpret the outcomes.

Manage – Interpretability of wearable system outcome can be promoted by providing a description of the system function to the final users, according to their knowledge and skills. Transparency and user trust is further developed by enabling the system to communicate the reasons and underlying data that lead to specific outputs.

Accountability

Map – When decisions rely on AI algorithm outcomes, it becomes challenging to assign responsibility and accountability in the event of errors related to software calculations. Additionally, AI-based systems must adhere to applicable norms and regulations, which can be particularly demanding in sectors like healthcare due to stringent requirements.

Measure – Assumption: AI outcomes are always interpreted by users, e.g., doctors for healthcare and risk managers for industry. *Quantitative Analysis: statistical metrics (e.g., percentage) of the number of times users take decisions different than those indicated as most probable by the AI system.*

Manage – Accountability should monitor the *quantitative analysis* metrics and investigate reasons of potential increase mismatches with internal audits.

Acceptability

Map – While technology can have a positive impact on the lives of workers and patients, low wearability of systems can hinder their acceptance. In a similar way, scarce acceptability can derive from perception and aesthetic evaluation. Additionally, poor usability, especially for non-technical users, can discourage adoption. Those acceptability concerns can lead to data shortages, impacting the quality of analysis and potentially resulting in data bias and the inclusion of an incorrect population.

Measure – *System Usability Scale (SUS): definition of a usability score for the wearables in the industry and healthcare application.* Questionnaires constitute the principal qualitative assessment tool to measure the SUS. *Consent Rate*: *measurement of the number of users voluntarily adopting wearable technology.*

Manage – Acceptability of a wearable system can be addressed following an iterative approach where customized surveys might be used to evaluate the wearable system acceptance during the iterative improvement of the system.

[1] Report: The value of occupational safety and health and the societal costs of work-related injuries and diseases. https://osha.europa.eu/en/publications/

[2] Public spending on incapacity https://data.oecd.org/socialexp/public-spending-on-incapacity.htm

[3] Report: Work-related musculoskeletal disorders: prevalence, costs and demographics in the EU. https://osha.europa.eu/en/publications/

[4] Musculoskeletal health in the workplace, https://doi.org/10.1016/j.berh.2020.101558

[5] Hossain, M.D., Aftab, A., Al Imam, M.H., Mahmud, I., Chowdhury, I.A., Kabir, R.I. (2018). Prevalence of work related musculoskeletal disorders (WMSDs) and ergonomic risk assessment among readymade garment workers of Bangladesh: A cross sectional study. PLOS ONE, 13(7): e0200122. https://doi.org/10.1371/journal.pone.0200122

[6] Franco, G. (2010). Work-related Musculoskeletal Disorders: A Lesson from the Past. Epidemiology, 21(4), 577-579

[7] [16] Alberto R, Draicchio F, Varrecchia T, Silvetti A, Iavicoli S. Wearable Monitoring Devices for Biomechanical Risk Assessment at Work: Current Status and Future Challenges-A Systematic Review. Int J Environ Res Public Health. 2018 Sep 13;15(9):2001. doi: 10.3390/ijerph15092001. Erratum in: Int J Environ Res Public Health. 2018 Nov 16;15(11): PMID: 30217079; PMCID: PMC6163390.

[8] Chini G, Fiori L, Tatarelli A, Varrecchia T, Draicchio F and Ranavolo A (2022). Indexes for motor performance assessment in job integration/reintegration of people with neuromuscular disorders: A systematic review. Front. Neurol. 13:968818. doi:10.3389/fneur.2022.968818

[9] Ranavolo A, Ajoudani A, Cherubini A, Bianchi M, Fritzsche L, Iavicoli S, Sartori M, Silvetti A, Vanderborght B, Varrecchia T, Draicchio F. The Sensor-Based Biomechanical Risk Assessment at the Base of the Need for Revising of Standards for Human Ergonomics. Sensors (Basel). 2020 Oct 10;20(20):5750. doi: 10.3390/s20205750. PMID: 33050438; PMCID: PMC7599507.

[10] European Commission, Ethics guidelines for trustworthy AI, report, study - Publication 08 April 2019 - https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[11]IEEE Standard Association, IEEE Standard for Wearable Consumer Electronic Devices--Overview and Architecture - https://standards.ieee.org/ieee/360/6244/

[12] ISO/IEC 23053:2022, Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML) - https://www.iso.org/standard/74438.html

[13] Information technology — Artificial intelligence — Management system - https://www.iso.org/standard/81230.html

[14] National Physical Laboratory (NPL) - https://eprintspublications.npl.co.uk/9683/1/MS45.pdf

[15] NIST, AI 100-1 - https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf