Scaling Responsible AI Solutions

Building an international community of practice and knowledge-sharing

November 2024

GPAI / THE GLOBAL PARTNERSHIP ON ARTIFICIAL INTELLIGENCE

This report was planned prior to the integration of the Global Partnership on Artificial Intelligence (GPAI) and the Organisation for Economic Co-operation and Development (OECD) mid-2024. Consequently, the report was not subject to approval by GPAI and OECD members and should not be considered to reflect their positions.

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.

Project Co-Leads:
Amir Banifatemi[*], Cognizant
Francesca Rossi[*], IBM Research
Venkataraman Sundareswaran*, SAAZ Micro

This report was prepared by **Kelle Howson**[‡], Consultant, who led the overall content development based on the work done by Experts and Teams who participated in the project. Special thanks are extended to **Chahinez Bensari**[‡], Consultant, for their meaningful contributions to the copywriting and drafting of content specifically related to the African track.

GPAI is grateful for the dedication of expert mentors from the global track who played a pivotal role through the whole project: **Borys Stokalski**[*], RETHINK; **Anurag Agrawal**[*], Ashoka University; **Stéphanie Camaréna**[†], Source Transitions; **Caroline Gans Combe**[†], INSEEC; **Gilles Fayad**[†], IEEE; **Benjamin Larsen**[†], World Economic Forum; **Monica Lopez**[*], CIfAI; **Zümrüt Muftuoglu**[*], Digital Transformation Office of the Presidency of the Republic of Türkiye; **Furukawa Naohiro**[*], ABEJA, Inc.; **Sandro Radovanović**[*], University of Belgrade; **Nava Shaked**[*], HIT Holon Institute of Technology and **Amitabh Nag**[*], BHASHINI.

GPAI is also grateful for the dedication of expert mentors from the African track who played a pivotal role through the whole project: **Shakira Barbirye**[‡], Infectious Diseases Research Collaboration (IDRC); **Mamadou Samba Camara**[‡], UQAD; **Idy Diop**[‡], UCAD; **Ibra Dioum**[‡], LITA; **Jean Louis Fendji**[‡], Afroleadership; [‡], **Niyongabo Julius**[‡],University of Burundi and Olivia University, Bujumbura ; **Ismael Kone**[‡], Virtual University of Cote d'Ivoire; **Neema Mduma**[‡], Nelson Mandela African Institution of Science and Technology; **Lawrence Nderu**[‡], Jomo Kenyatta University of Agriculture and Technology (JKUAT); **Tsopze Norbert**[‡],University of Yaounde 1 (UY 1); **Elisabeth Oseku**[‡], HASH; **Joel N. Nwakaire**[‡], ATPS; **Ugochi Okengwu**[‡], University of Port Harcourt; **Mesfin Fikre**[‡], Addis Ababa University and **Austin Waffo**[‡], University of Yaounde 1 (UY 1).

GPAI is grateful for the meaningful contribution of experts who supported the project in its different phases: **Andrea A. Jacob*,** Code Caribbean; **Kudakwashe Dandajena*,** African Institute for Mathematical Sciences (AIMS) and University of the Western Cape; **Ulises Cortés***, Universitat Politècnica de Catalunya; **Ricardo Baeza-Yates*,** Ricardo Baeza-Yates, Expert, Institute for Experiential AI of Northeastern University.

GPAI would also like to thank the expert participants of the Working Session on Recommendations, who provided valuable insights to enrich this report: **Shakira Barbirye**[‡], Infectious Diseases Research Collaboration (IDRC); **Gilles Fayad**[†], IEEE;

Jean Louis Fendji‡, Afroleadership; **Zümrüt Muftuoglu***, Digital Transformation Office of the Presidency of the Republic of Türkiye; **Joel N. Nwakaire**‡, ATPS; **Ugochi Okengwu**‡, University of Port Harcourt; **Sandro Radovanović***, University of Belgrade; **Borys Stokalski***, RETHINK and **Austin Waffo**‡, University of Yaounde 1.

GPAI wants to underline the organisations who contributed in delivering quality trainings to the expert mentors from the African track: **FAIR Forward - AI for All**‡; **GIZ - Deutsche Gesellschaft für Internationale Zusammenandeit (GIZ) GmbH**‡, **African Observatory of Responsible AI**‡ and **Ladysmith**‡.

GPAI is thankful for the dedicated work of the AI Solution teams who applied to and completed the global track of the Scaling Responsible AI Solutions mentorship programme and produced meaningful outputs for this report: **Pawel Potakowski**, **Max Salamonowicz**, **Maciej Lewandowski**, from Migam; **E. Ulises Moya Sánchez**, **Abraham Sánchez**, **Alejandro Zarate**, **Edgar Villareal**, from the Gobierno de Jalisco; **Izabella Krzeminska**, **Michal Butkiewicz**, **Artur Bajll**, **Alicja Kasicka**, **Jakub Rzeznik**, **Emilia Lesiak**, **Michal Szczerbak**, **Maciej Jonczyk**, **Damian Boniecki**, **Adam Konarski**, **Pawel Tuszynski**, from Orange Innovation Poland; **Dr. Dinesh Gupta**, Dr. Shweta Birla Dhakonia, from the Translational Bioinformatics Group, ICGEB, **Dr Viney Gupta**, from the RP Center of Ophthalmology; **Wojciech Ozimek**, from One2Tribe; **Jan Twardowski**, **Urszula Mosiej**, from Websensa; **Nitin Singh**, **Ranjan Relan** from Agent Analytics.

GPAI is also grateful for the dedicated work of the AI Solution teams who applied to and completed the African track of the Scaling Responsible AI Solutions mentorship programme and produced meaningful outputs for this report: **Andre Kevin Nyemb Ndjem Eone**, **Velda Nbayen**, **Wilfried Girlan Wayou**, **Ledoux Kouam**, **Hilary Ngouadje**, from Greenlive Agriculture; **Ugochi A. Okengwu**, **Akpughe Hillard Azino**, **Eruotor Taiye**, **Eyinanabo Odogu**, **Sunday Nwovu**, from AI solutions for Multi-Crop Leaf Disease Detection; **Sepele Cyrille,Ted Bekolo,Taira Manuel**, **Nedaouka Joachim**, from Kit4council; **Brian Halubanza**, **Chansa Chomba**, **Selina Kadakwiza**, **Kondwani Kabaghe**, **Chipulu Nshenda**, from Human Wildlife Conflict Management System using AI powered IoT; **Lawrence Nderu,Ruth Oginga**, **John Michael Rono**, **Maryanne Mwihaki**, **Derrick Matindo Obwatsa**, **Millicent Shiatikha**, **Sonia Lomo**, **Daniel Muiruri**, **Felista Mogire**, **Bonface Ingumba**, from Harnessing the capabilities of Large Language Models (LLMs) to create awareness on Data protection Laws in Kenya, Uganda and Rwanda; **Victor Ogunbiyi**, **Oluwabukola Dada**, **Joel Danjuma**, **Fasobu Adekemi**, **Odewunmi Bolaji**, from Dawn Ai Study; **Teegwende Zougmore**, **Ayodele Awokoya**, **Billy Peter Munyenyembe**, **Daphne Machangara**, from Non-Intrusive Fish Weighing; **Birba Bienvenu Emmanuel**, **Kinig Yilomou Ingrid**, **Kone Abdoul Kader**, **Armel Emmanuel Sogo**, **Zombre Christian**, from Delia: AI-Powered Chatbot and Voice Health Assistant; **Stephane Fadanka**, **Elisée Jafsia**, **Bawang Bawa**, **Ruqaiya Sattar**, **Agien Petra**, fromStrengthening responsible AI Pipeline for Mboalab AI4 Health Project; **Walelign Tewabe Sewunetie**, **Hailemariam Abebe**, **Tesfamariam Mulugeta**, **Surafel Tilahun**, from Large Language Models for Sexual, Reproductive, and Maternal Health Rights; **Victoria Mukami**, **Consolata Gakii**, **David Mugo**, **Marilyn Ronoh**, **Stanley Ndegwa**, from BESHTE: A Chatbot to enhance HIV testing, status awareness, and status disclosure among youth in Kenya; **Kansiime Noel**, **Hassan Bahati Mukisa**, **Kesandu Uchenyi**, **Mark Tushemerirwe**, **Barbara Atukunda**, **Lilian Mirembe**, **Moses Ariong**, from Regulatory AI; **Sebastian Mwaura**, **Mary Munyao**, **David Munene**, **Ruth Wangui**, **Fred Omollo**, **Lucy Njoki**, **Samuel Maloba**, **Fabian Cheruiyot**, from AI-Driven Carbon Credit Calculations for Electric Motorcycle Fleets.

GPAI extends its gratitude to **Matthew Smith** and **Gillian Dowie** from the Canadian International Development and Research Centre (IDRC) for the trust attributed to the Scaling Responsible AI Solutions project teams and methodology in funding the African track pilot as part of the AI4D programme.

Finally, GPAI would like to acknowledge the tireless efforts of colleagues at the International Centre of Expertise in Montréal on Artificial Intelligence (CEIMIA). We are particularly grateful to **Arnaud Quenneville-Langis** and **Laëtitia Vu** from CEIMIA for ensuring the project management and coordination of the project's global track, and to **Thomas Hervé Mboa Nkoudou**, **Brigitte Ondigui** from CEIMIA, and **Winston Ojenge‡**, African Centre for Technology Studies for ensuring the project management and coordination of the project's African track. We are also thankful for the dedication of the Working Group Co-Chairs **Amir Banifatemi***, AI Commons and **Francesca Rossi***, IBM Research.

* Expert of GPAI's Responsible AI Working Group
** Observer at GPAI's Responsible AI Working Group
† Invited Specialist
‡ Contracted parties by the CEIMIA

Citation

GPAI 2024. *Scaling Responsible AI Solutions: Building an international community of practice and knowledge-sharing*, Report, November 2024, Global Partnership on AI.

Table of Contents

Executive Summary

This report marks the conclusion of the second year of the Scaling Responsible Artificial Intelligence Solutions (SRAIS) project, an initiative of the Responsible AI (RAI) working group of the Global Partnership on Artificial Intelligence (GPAI). In 2024 the project has grown in scope and impact, and has taken strides towards consolidating a global network of collaboration and knowledge-sharing. This network is focused not only on responsibility in the development of AI-based systems, but more uniquely on the intersection between scalability and responsibility.

The process of scaling an AI-based application presents distinct challenges in terms of adherence to RAI principles. These include the need for responsible approaches to data and cultural integration in new places of operation; the risk of bias amplification as an application gains a larger and more diverse user base; the additional resource demands of responsible technical and operational expansion; the need to navigate varying legal and regulatory frameworks; and the imperative of assessing and mitigating the potential complex societal, developmental and environmental impacts of a given AI-based system in all of its intended use contexts.

Navigating this range of issues places additional demands on teams looking to establish and scale their AI-based offerings, and these demands can risk placing those committed to fully adhering to responsible principles, at a disadvantage. This is compounded by overall market concentration in the global AI ecosystem—which makes it difficult for small players and those outside the global centres of technological innovation to compete. There is a clear need for interventions which support and enhance the scalability of RAI approaches, in order to raise adherence to responsibility standards across the board.

The SRAIS project seeks to advance this objective by connecting teams working on AI-based applications around the world with expert mentors, who together undergo a structured process designed to identify and overcome key challenges they face at the responsibility/scalability nexus. Alongside the concrete outcomes of mentoring for individual teams, this methodology is designed to elicit more generalisable insights, lessons and recommendations for stakeholders involved in scaling responsible AI applications globally.

In 2024 the SRAIS project, with dedicated support from the International Development Research Centre (IDRC) of Canada, introduced a special track, focused on African AI teams. This was part of our effort to expand our geographical reach to low- and middle-income country contexts which face disproportionately high barriers to locally-driven digital innovation and entry into global markets for AI-based applications. Africa in particular has been underrepresented in the development and ownership of AI-based systems globally (despite contributing significant volumes of data and labour to AI value chains).

A total of 21 teams from eleven countries took part in the SRAIS mentorship cycle this year, with 13 of these in the “Africa Track”. This report captures the experiences and outcomes of the mentoring process for individual teams, as well as broader lessons learned and widely applicable recommendations.

While each participating team faced a unique combination of challenges relating to their specific applications, the context in which they were working, and their scaling goals, commonalities were identified across many teams, which have been captured in the thematic discussion and lessons learned sections of this report. These include: (1) regional differences and unevenness in access to resources (like data and compute), infrastructure, technical capacity, and enabling regulatory frameworks, which particularly disadvantaged Global South teams including those in the Africa track; (2) an overall shortfall in awareness and literacy around the spectrum of responsible AI principles amongst stakeholders—while there is general recognition of issues related to privacy and accuracy, there is less familiarity with other concerns around data governance, cultural integration, and rights violations that may arise across the AI lifecycle and sustainability; and (3) a disproportionate focus on technical "fixes" to enhance compliance with responsibility standards, which do not necessarily engage meaningfully with the more nuanced ethical, social and political aspects of responsibility in different contexts. In light of these insights and others drawn from this year's project cycle, this report advances a series of concrete recommendations aimed at AI teams, policymakers, funders and third party auditors, to simultaneously strengthen the responsibility and scalability of AI-based applications.

Introduction

This year marks the second cycle of the Scaling Responsible Artificial Intelligence Solutions (SRAIS) project, an initiative of the Global Partnership on Artificial Intelligence (GPAI),[1] coordinated by the Montreal Center of Expertise in Artificial Intelligence (CIEMIA). This year, the project has expanded its scope and impact. This report summarises the activities of the project in 2024, including how its rationale, theory of change and methodology have been further consolidated, as well as the concrete contributions the project has made to supporting responsible AI projects around the world to scale responsibly, and the insights gained from these engagements.

The SRAIS project addresses a specific need in the AI field for which there is currently insufficient support. While significant work has been dedicated to defining and developing frameworks for responsible AI (RAI) at the global, regional and national levels, many developers of AI applications encounter challenges when seeking to operationalise them, particularly when scaling their applications in a way that preserves RAI principles. Key processes involved in scaling AI—such as growing an application's user base, making the application available in different countries or contexts, and expanding its uses—may pose unique social and environmental risks, even if the application was initially designed and developed in line with RAI principles.

In recognition of this challenge, the SRAIS project was conceived to provide hands-on support and expert guidance to selected teams around the world entering the deployment and scaling phases. The overarching objectives of the project are two-fold. First the project seeks to assist teams in clearly defining the key challenges they face in scaling their applications responsibly, and in developing an actionable approach to addressing these challenges. This is achieved through a structured series of engagements with dedicated expert mentors over the course of several months. Second, the project aims to generate and capture new insights into the experiences of those looking to deploy and scale RAI applications responsibly around the world. This is in order to contribute to the global knowledge-base on the opportunities and challenges posed by AI for diverse stakeholders; as well as to distil general recommendations for both RAI developers and policymakers, and to support the scaling of RAI at a broader level.

These dual objectives can be broken down to a more granular series of outcomes, namely: (1) providing opportunities to deploy and scale RAI solutions, (2) evaluating and showcasing results, (3) encouraging cross-functional collaboration, (4) helping develop performance metrics, and (5) operationalising the RAI framework across different uses and contexts within GPAI and beyond.

Introducing the SRAIS African Track

In 2024 the project had the ambition of reaching more countries and territories, especially in places excluded from, or facing particular challenges in, the local development and scaling of AI-based applications. It is precisely within low- and middle-income country contexts that the potential of responsible AI could be most profound, with respect to addressing key

[1] GPAI structure - integration with OECD etc.

sustainable development challenges to improve lives. However actors in these contexts often face higher barriers to accessing and realising the potential of technological tools for development. In recognition of this, a particular effort was made to reach more teams and mentors from the Global South to enhance international collaboration and to include diverse perspectives and contextual insights. As part of this effort, this year we have introduced a special track of the project focused specifically on Sub-Saharan Africa, which has run in parallel with the global project. In 2024 the “Africa Track” work of the SRAIS project has been undertaken in partnership with the International Development Research Centre (IDRC) of Canada.

Africa, with its diverse regions and burgeoning technological landscape, is already home to numerous existing AI initiatives. Despite the commendable efforts of these initiatives, they often grapple with challenges related to scaling responsibly (Eke et al., 2023). These challenges have sometimes been characterised in the literature as representing a phenomenon of digital colonialism. Digital or data colonialism refers to regional inequalities in terms of how representation, labour and value is distributed in the production and consumption of digital goods (Couldry & Mejias, 2019; Coleman, 2018). For instance, while the African continent is highly integrated into global digital and AI production, providing an enormous amount of data and labour resources into the production of digital technologies, this integration has oftentimes taken place on an inequitable and extractive basis (Anwar & Graham, 2022). Moreover, African identities, languages and cultures are often not reflected within the digital and AI-based systems which are developed in the Global North and used in African countries.

The value created by the development and use of AI technologies has been distributed unevenly between world regions, with Africa so far gaining the smallest share. A 2019 UNCTAD report suggested that at that time Africa and Latin America together accounted for 1% of the market capitalisation value of the 70 largest digital platforms in the world (UNCTAD, 2019). It is clear that there is enormous energy, potential and capacity for innovation within the locally-owned African AI ecosystem. However structural barriers to becoming established and scaling are higher for the large majority of African startups, than for those in other places. It is critical for even and sustainable development that African-owned AI initiatives are able to scale regionally and globally—and that Africa’s integration into the global AI ecosystem does not follow an inequitable path of the extraction of data resources and exploitation of low-waged labour.

Against this backdrop, the “Africa Track” of the SRAIS project aims to work in collaboration with developers of responsible AI-based systems in Africa, to support them towards navigating and dismantling barriers to scaling. Key objectives were to foster capacity building and community engagement centred around the principles of Responsible AI. This strategic vision extends beyond regional boundaries, with the aim of connecting African experts to the broader international Responsible AI community. In 2024 the SRAIS Africa Track was supported by the International Development Research Centre (IDRC) of Canada.

This report brings together the activities and outcomes of both the “global” and the “Africa” tracks of the project in 2024, and presents some comparative/integrated analysis of challenges and opportunities faced by the teams across the two tracks. A more detailed overview and analysis of the activities of the SRAIS Africa Track 2024 will also be available in a dedicated report to be published later.

Following two cycles of the SRAIS project, we have clearly demonstrated that it is not enough for an application to simply *set out* to contribute to a social or environmental good; or to incorporate RAI principles in the conceptualisation and design phases of a project. Nor is it enough to ensure technical excellence. These are all imperatives, but if an AI project is to meaningfully help to solve pressing problems and create value for local communities, whilst mitigating risks and harms to people and environments, responsibility must be maintained while the project scales, and this introduces new challenges and considerations for RAI frameworks to grapple with. The following section identifies and summarises common practical considerations and challenges encountered in the course of scaling AI responsibly.

Challenges and considerations in scaling responsible AI

As the 2023 SRAIS report detailed, some of the challenges that RAI applications face when attempting to scale are structural and beyond the direct control of most small AI developers (GPAI, 2023). These include the very high degree of market concentration in the sector, whereby infrastructure (like data storage and management solutions or compute resources) is owned by a small handful of corporations. In addition, access to much-needed capital is highly uneven—and particularly lacking for those outside the global tech industry centres, like Silicon Valley, London and Shenzhen. The incentives created by this market concentration coupled with existing regulatory environments often reward ‘first-movers’, rapid deployment, proprietary and enclosed systems, and ‘monetisation’ potential. In turn this can disadvantage and disincentivize slow and careful development and testing, as well as alternative ownership and governance models based on principles of collaboration, open-access, solving real-world problems, and the creation of value for local communities.

High-level regulatory responses are required to challenge market concentration in AI, in order to enable more small-scale operations to gain a foothold. During this SRAIS project cycle, we have identified some policy recommendations in this vein. Yet, despite this challenging environment of market concentration, the advent of AI and generative AI has captured peoples’ imaginations, and all over the world people are conceiving of ways to harness AI responsibly, to improve lives around them. Because it brings together a diverse global community of AI practitioners, experts and policymakers who are committed to the responsible development and use of AI-based systems, the Global Partnership on AI is uniquely positioned to intervene in support of such visions, and to provide an institutional springboard for them to be realised and to grow. The SRAIS project has been at the forefront of these efforts, and through this work we have deepened our understanding of the practical challenges and considerations arising from the scaling of responsible AI.

In order to maintain responsibility through the scaling process, applications must be inherently responsible, or adhere to principles of responsible design and governance prior to scaling. Although conceptualisations of what constitutes responsibility in AI research and development are subject to continuous evolution and debate, there are now well-established global frameworks for RAI design and governance, including the OECD's (2019) Recommendation of the Council on Artificial Intelligence; IEEE's (2019) Ethically Aligned Design, UNESCO's (2022) Recommendation on the Ethics of Artificial Intelligence, IEEE/ISO/IEC 24748-7000-2022, and the European Union's AI Act 2024. These frameworks outline agreed-upon characteristics of responsible AI systems, and share cross-cutting themes and principles. Broadly these encompass (1) fairness, which refers primarily to ensuring that AI-based systems treat everyone fairly, and are inclusive and free from bias, (2) data protection and privacy, or the safeguarding of user data from predatory or harmful uses, (3) robustness, safety and accuracy, or ensuring that the outputs of AI systems are trustworthy and reliable for the purposes for which they are being used, (4) human-centred design, or a commitment to the use of AI to improve and enhance (rather than replace or subjugate) human capabilities and creativity, (5) transparency, explainability and accountability, or ensuring that users are adequately informed of when they are interacting with AI systems, why the AI-based system has made a specific decision or recommendations, and the limitations and risks of AI-based systems, as well as determining who is responsible if risks or harms occur, and how users may seek recourse for harms suffered, and (6) sustainability, or assessing and mitigating the environmental impact of AI-based systems (for instance their energy and water usage). In addition the EU AI Act classifies uses of AI based on their level of risk, with certain very risky types of AI systems prohibited outright—including systems which are inherently manipulative or deceptive in order to change users' behaviour; systems of biometric profiling which infer sensitive attributes (such as race or political opinion); as well as certain methods of social scoring, real-time facial recognition, and inferring emotions in specific settings.

While the 2023 SRAIS report provided a deep-dive into RAI principles with reference to these frameworks, in the remainder of this section we focus on the practical considerations and requirements that emerge at the intersection of responsibility and scaling, and point to several key priorities for ensuring responsibility during scaling. Traditional scaling generally involves expanding into new geographies, reducing costs, enhancing accessibility and building user bases. In the field of AI, additional unique challenges and imperatives come to the fore when undertaking these activities. The below summary of specific responsible scaling challenges draws on a forthcoming article written for GPAI by SRAIS project co-lead Amir Banifatemi.

Data and cultural integration

Data and cultural integration here refers to the need to incorporate additional sources of data into an AI-based system's training in order to make sure it remains accurate and safe when expanded into new contexts. For example, a system which has been developed to predict weather patterns in a specific region based on historical data from that region would not be accurate when scaled to new regions with different climatic conditions, without access to dedicated data from those contexts. Likewise, this manifests in the context of algorithmic invisibility when data collection and algorithmic systems fail to adequately capture or represent local cultural contexts, leading to the systematic exclusion or misrepresentation of different populations or local perspectives, knowledge systems, and ways of life in digital spaces. This technological bias is particularly evident in how search engines, content recommendation systems, and AI-based models trained primarily on Western datasets often overlook or misinterpret African cultural nuances, perpetuating digital divides and knowledge inequities. Whilst these are relatively straightforward examples of the problems that may arise when an AI-based system's training data is not representative of the context it is being used in, more subtle and often more pernicious harms can occur when training data is not representative of cultural values, norms, languages, and knowledge systems in contexts to which an AI-based system has been scaled. In such cases AI-based systems may generate irrelevant, inappropriate or even unsafe outputs.

Bias amplification

Moreover as an AI-based system is scaled there is a greater risk of bias amplification—or that a lack of representativeness in datasets may lead to discriminatory real-world outcomes. This often occurs not only due to contextual or cultural differences, but also due to the presence of underlying biases in the aggregate of existing socially-produced data available on the internet, which stem from pre-existing social relations. As an AI-based system is used by more and more parties for more varied purposes, the consequences of underlying biases become all the more serious. Gender or social biases are particular areas of concern, and it is critical that the expansion of AI-based systems ensures responsiveness to contextual gender and social norms, perspectives and inequities.

Legal and regulatory integration

Ensuring that AI-based systems comply with all relevant legal frameworks during scaling is a challenging, but nonetheless essential task. Traditional cross-border enterprises whose operations are not based predominantly on digital platforms or the exchange of digital products, typically encounter the need to establish legal and physical entities in the countries they expand to, and as such to navigate and comply with local regulations. However, AI-based systems can be deployed to myriad locations with much less friction whilst being managed remotely. This can lead to a lack of clarity as to prevailing jurisdiction, compliance matters, intellectual property issues, licensing, the ownership of AI-generated content, and the rights and responsibilities of users and operators. The lack of clarity is compounded by the fact that legal frameworks to explicitly govern AI-based applications are absent or still emerging in much of the world (Adams et al., 2024). Greater global cooperation and harmonisation in the development of AI regulation may assist in facilitating the scaling of RAI. However, it is also critical that AI operators ensure compliance with all applicable extant legal frameworks including in spheres of competition, property, labour, taxation, etc. whilst scaling their applications.

Technical and operational expansion

To scale effectively, it is helpful for AI-based systems to be able to integrate smoothly with other technologies and platforms. 'Interoperability' can be achieved through the use of industry standards, and helps with improving the useability and functionality of AI-based systems. Achieving interoperability can be challenging, however, as it requires industrial cooperation and cohesion, in a sector that—as described above—has often been characterised by the close guarding and gatekeeping of infrastructure and commercial property such as algorithms and platforms. Scaling AI-based systems can also present other technical challenges, such as ensuring operations can handle higher user loads and increased data volumes without compromising performance.

Labour and economic considerations

The use of AI-based systems will have varied impacts on job markets, livelihoods and economic opportunities around the world. In contexts with high unemployment, it is critical to ensure that AI-based applications are not used to 'fill gaps' in service provision in a way that forecloses on livelihood opportunities for local people. In contexts with low unemployment, the use of AI-based systems has led to fears about the displacement of labour by technology or degradation of wages and working conditions. At the same time, rapid growth in the AI sector may be creating opportunities for those with relevant skills (for instance in coding and software development) who are located in places where technological development is concentrated.

Finally, it is becoming increasingly recognised that the development and maintenance of AI-based systems requires large quantities of what is termed by some to be "low skilled" labour—in the form of entering and labelling data, monitoring and correcting outputs, content moderation, and many other tasks. This requirement becomes particularly pronounced as a system is scaled—as the need for "human-in-the-loop" and data labour will inevitably increase. In many instances this human labour is critical for ensuring that the outputs of an AI-based system are safe, fair and compliant. Paradoxically, these tasks are typically outsourced to large pools of workers in low- and middle-income countries and characterised by very low wages, high degrees of 'algorithmic management' and subordination, high labour intensity, and a lack of access to statutory labour protections. In light of the highly complex and contextual implications of AI-based applications for the world of work, when scaling AI systems it is critical to recognise both the additional labour inputs that will likely be required, and the conditions of this labour, alongside the differential labour impacts of the deployment of AI systems in different places.

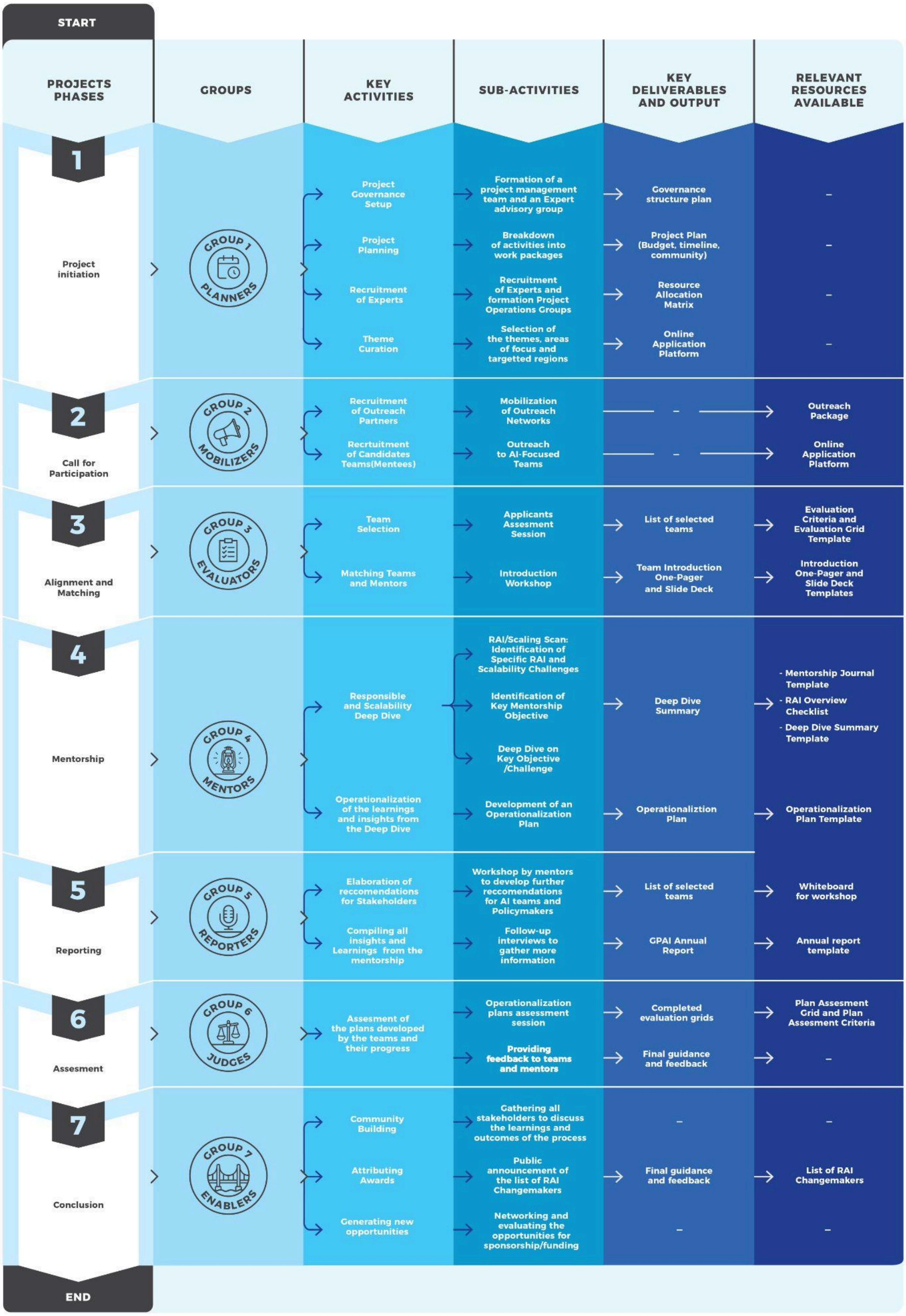
The general challenges and considerations that arise in scaling responsible AI as detailed above are present in different ways in the experiences of many of the teams who have participated in the SRAIS project over the course of 2023 and 2024. Following a brief discussion of the methodology of the project we go on to present the stories of the 2024 participating teams, and how their approach to scaling responsibly in light of these challenges has been refined through the SRAIS mentoring process.

Methodology

The SRAIS project is focused on the dual objectives of, first, driving real-world impact by applying collective expertise to existing endeavours, and second, gathering insights and building understanding of the realities and challenges faced by actors around the world in scaling responsible AI, which can be applied more broadly by practitioners and policy makers to help overcome barriers to scaling responsible AI. The key methodology we apply to both of these objectives is that of structured mentorship. Connecting AI teams to mentors who are international experts in their fields, to work collectively towards a defined goal, provides a pathway to overcoming key responsible scaling challenges. These include operationalising high-level responsibility principles in a specific application; and identifying, defining and demonstrating metrics of success; as well as simply opening up access to global spaces, dialogues and institutions at the forefront of responsible innovation, which is so often lacking for developers in the majority world.

The SRAIS mentoring process is designed to be highly tailored to the needs of each team, whilst at the same time producing lessons and outcomes that have broader value and applicability to other actors in the RAI space, including developers and startups, and policymakers.

Beyond immediate practical support and knowledge-sharing, the linkages established through the mentoring relationship contribute to the formation of a diverse global community of practice dedicated to the dissemination and growth of responsible AI. In this way, mentoring forms the basis of the SRAIS project’s theory of change, by both directly and indirectly supporting the realisation of a critical mass of explicitly responsible AI projects which are able to scale and compete with existing AI players.



START

PROJECTS PHASES	GROUPS	KEY ACTIVITIES	SUB-ACTIVITIES	KEY DELIVERABLES AND OUTPUT	RELEVANT RESOURCES AVAILABLE
1 Project initiation	GROUP 1 PLANNERS	Project Governance Setup	Formation of a project management team and an Expert advisory group	Governance structure plan	-
		Project Planning	Breakdown of activities into work packages	Project Plan (Budget, timeline, community)	-
		Recruitment of Experts	Recruitment of Experts and formation Project Operations Groups	Resource Allocation Matrix	-
		Theme Curation	Selection of the themes, areas of focus and targetted regions	Online Application Platform	-
2 Call for Participation	GROUP 2 MOBILIZERS	Recruitment of Outreach Partners	Mobilization of Outreach Networks	-	Outreach Package
		Recrtuitment of Candidates Teams(Mentees)	Outreach to AI-Focused Teams	-	Online Application Platform
3 Alignment and Matching	GROUP 3 EVALUATORS	Team Selection	Applicants Assesment Session	List of selected teams	Evaluation Criteria and Evaluation Grid Template
		Matching Teams and Mentors	Introduction Workshop	Team Introduction One-Pager and Slide Deck	Introduction One-Pager and Slide Deck Templates
4 Mentorship	GROUP 4 MENTORS	Responsible and Scalability Deep Dive	RAI/Scaling Scan: Identification of Specific RAI and Scalability Challenges	Deep Dive Summary	- Mentorship Journal Template - RAI Overview Checklist - Deep Dive Summary Template
			Identification of Key Mentorship Objective		
			Deep Dive on Key Objective /Challenge		
		Operationalization of the learnings and insights from the Deep Dive	Development of an Operationalization Plan	Operationaliztion Plan	Operationalization Plan Template
5 Reporting	GROUP 5 REPORTERS	Elaboration of reccomendations for Stakeholders	Workshop by mentors to develop further reccomendations for AI teams and Policymakers	List of selected teams	Whiteboard for workshop
		Compiling all insights and Learnings from the mentorship	Follow-up interviews to gather more information	GPAI Annual Report	Annual report template
6 Assesment	GROUP 6 JUDGES	Assesment of the plans developed by the teams and their progress	Operationalization plans assessment session	Completed evaluation grids	Plan Assesment Grid and Plan Assesment Criteria
			Providing feedback to teams and mentors	Final guidance and feedback	-
7 Conclusion	GROUP 7 ENABLERS	Community Building	Gathering all stakeholders to discuss the learnings and outcomes of the process	-	-
		Attributing Awards	Public announcement of the list of RAI Changemakers	Final guidance and feedback	List of RAI Changemakers
		Generating new opportunities	Networking and evaluating the opportunities for sponsorship/funding	-	-

END

Figure 1. Roadmap for the deployment of the SRAIS project

Structure and phases of the mentoring process

The mentoring approach we take is one of constructive and collaborative partnership between mentors and teams. Mentors aim to gain a holistic understanding of teams' objectives and challenges, advise on the applicability of RAI principles to their work, and to guide them on overcoming challenges and implementing these principles. In this way the methodology of the SRAIS project constitutes a highly applied approach to RAI, where mentors and teams exchange knowledge based on real-world experiences and applications, with a clear focus on measurable impact and outcomes. In 2024, the mentoring process took place from June to November, following the selection of participating teams, and involved engagements between teams and mentors, and the production of defined outputs.

Selection of participating teams

Prospective teams who were in the process of building an AI-based application and had reached at least the pilot stage responded to a public call for participation, and a final selection was made by a sub-committee of mentors alongside the project leads.

The selection process paid particular attention to both the extent to which applications are grounded in RAI principles, and the scaling aspirations and challenges faced by these applications.

In addition, the selection process aimed to achieve a balance of topic, geographical and gender representativeness amongst participating teams, as well as representation from both the public and private sectors. This system is intended to ensure the mentoring process covers a wide range of perspectives to be as impactful as possible for participating teams from a scalability standpoint, whilst making sure that a commitment to responsibility remains central.

Eleven teams applied to take part in the 2024 mentoring process under the global track, and Thirty-one under the African track, representing a total of sixteen countries. Eight teams were selected in the global track, and thirteen in the African track, with eleven countries ultimately represented amongst the participating teams.

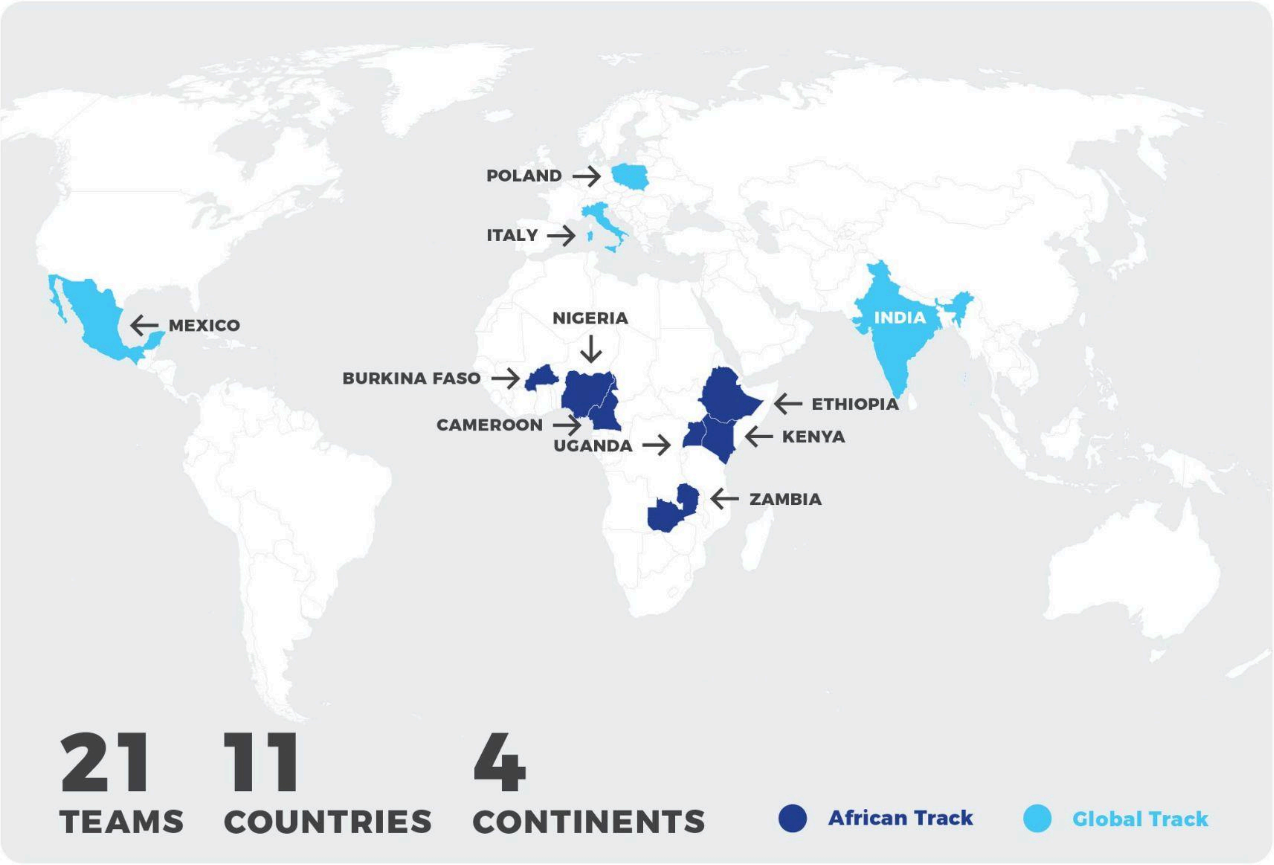


Figure 2. Countries of provenance of participating teams

Direct mentoring

Mentors in the SRAIS project held six formal '*mentorship sessions*' with participating teams across the cycle of the project. The purpose of these sessions was to help identify specific responsibility and scalability challenges facing teams, and to provide tailored guidance to support teams in overcoming these challenges. Each participating team was matched with a lead mentor and supporting mentors, with consideration given to the relevance of mentors' specific expertise and experience to teams' contexts and projects.

The early mentoring sessions were designed to help teams to clearly identify and prioritise the most pressing and important SRAIS challenge/s they face. For instance mentors explored with teams aspects such as operationalising particular RAI principles; developing key performance indicators (KPIs) with which to assess their adherence to RAI principles in testing, or as an application scales; developing clear guidance for responsible use of their application by third party users; or articulating their responsible use case to funders, institutional adopters, policymakers, or other stakeholders.

Guidance on written outputs

Mentors then worked together with participating teams to develop a brief written output which set out the teams' approach to addressing a key challenge identified. Templates were provided for these outputs in order to ensure they were clear, structured and relevant to broader scaling RAI conversations. As such, mentors assisted teams to produce both a "deep-dive" discussion of the challenges they faced in scaling responsibly, as well as an actionable, documented plan or approach to overcoming a key challenge, consisting of a series of detailed and concrete steps for implementation, a detailed timeline and metrics by which to evaluate progress. The following section of this report summarises the outcomes of the mentoring process for each team—specifically the key challenges and action plans captured in their written outputs.

In line with the SRAIS project's objectives of knowledge sharing, cross-functional collaboration, and building the base of technical guidance for the operationalisation of RAI, in each case, it was important that the output has applicability beyond just the participating project. While the mentoring process serves an immediate practical purpose in assisting teams to overcome their identified challenges, the intention is also that stakeholders can learn from and apply insights contained in the outputs in other contexts and uses into the future.

Identifying policy and practice recommendations

Mentors then worked with teams, project leads and GPAI Expert evaluators to identify recommendations for both other AI developers and for policymakers and legislators, that arise from their experiences of responsibility and scalability challenges. These recommendations are provided at the end of this report.

Monitoring and evaluation of outcomes

Following the conclusion of the structured mentoring process and the publication of this report, a sub-committee of GPAI Experts was responsible for monitoring and evaluating the implementation of the goals and targets laid out in each team's plan. The sub-committee liaised with lead mentors and teams to access evidence and documentation demonstrating progress on targets, identify where targets are not being met, and provide clear feedback and suggestions to teams. These evaluation results help mentors and teams conclude their participation in the project with clear and validated feedback and advice from a trusted community of RAI Experts.

The participating projects: Aims, challenges and outcomes of the mentoring cycle

Overview of participating teams

Global track

AI-Focused Team	Country	Topic	GPAI Area of Focus	Challenge addressed by the SRAIS Deep Dive Summary and Plan
ASLAC automatic sign language avatar creation - societal and technological innovation disrupting	Poland	Universal Access	Human Rights	To develop a responsible data collection process that allows the use of data with personal information (video) while guaranteeing that no personal features make it to the training sets.
Artificial Intelligence-Based Referral System for Patients with Diabetic Retinopathy in Jalisco	Mexico	Health	Global Health	To ensure that outputs and decisions of the system align with local needs, to develop long-term strategies for secure and compliant data storage and sharing and to tailor the application to the specific requirements of clinicians.

AI-Focused Team	Country	Topic	GPAI Area of Focus	Challenge addressed by the SRAIS Deep Dive Summary and Plan
Personality AI for Hypersonalisation with user's privacy protection (privacy by design)	Poland	Personalization	Data Governance	To develop a transparent and explainable hyper-personalization system component that respects user privacy while continuously monitoring and adapting based on user behaviour without external data transmission.
AI-based predictive models for diagnosing angle dysgenesis on ASOCT scans in glaucoma diagnostics and treatment decision support	India	Health	Global Health	To minimise any bias in the training dataset by ensuring diversity with balanced class samples from various demographics and geographical locations, accounting for confounding factors such as age, gender, and ethnicity. Include explainability for AI forecast outcomes.
Employee Performance Management, Learning & Development	Poland	Professional Development	Future of Work	To ensure the system provides maximum transparency, so that users can understand the recommendations and rewards provided to them, and understand when the targets and tasks are assigned to them directly by managers or automatically by the system.
LLM-Based Knowledge Management & Communication System Ready for Corporate Deployment	Poland	Knowledge Management	Innovation and Commercialisation - Data Governance.	To develop comprehensive, audience-specific guidelines for implementing and using an enterprise-specific AI-based system in a

AI-Focused Team	Country	Topic	GPAI Area of Focus	Challenge addressed by the SRAIS Deep Dive Summary and Plan
				responsible manner, with a focus on hallucination mitigation.
System for satisfaction recognition in Conversation with AI	Poland	Customer Satisfaction	Innovation and Commercialisation	To clearly define the uses, objectives, and “rules of engagement” of the system, and to set out the team’s approach to ensure adherence to Responsible AI principles and mitigating and minimising possible risks and harms to users such as manipulation of users, disruption of the user experience, errors and inaccuracies in the system.
Responsible AI Co-worker (Assistant) for Enterprises	India	Corporate AI	Future of Work	To detail how specific responsible AI ethics principles such as fairness and ethical privacy were operationalised and validated in the design and deployment of the RAI Co-worker multi agent platform.

Figure 3. Overview of participating teams of the SRAIS Global Track

African track

AI-Focused Team	Country	Topic	GPAI Area of Focus	Challenge addressed by the SRAIS Deep Dive Summary and Plan
Greenlive Agriculture : Precision And Control At Your Fingertips	Cameroon	AI-Enhanced Agriculture for Sustainability	AI for the Environment	To address the challenge of monitoring water needs and crop diseases by identifying specific environmental problems, overcoming barriers to technology access and reducing data collection costs.
AI4Health Project: Strengthening the RAI Pipeline for Mboacare and Mboathoscope”	Cameroon	Data Governance in Global Health	Global Health	To establish a data governance framework to manage and secure health information, by ensuring accessible and compliant data usage across healthcare initiatives.
Kit4Council	Cameroon	Mobile Access to Digitized Archives	Data Governance	To enable mobile access to digitised archives and . leveraging a language model; allowing users to retrieve specific document information as requested.
AI-Driven Carbon Credit Calculations for Electric Motorcycle Fleets	Kenya	AI-Driven Climate Action	AI For the Environment	To apply a qualitative AI risk assessment guide in environmental contexts by focusing on problem identification and data collection to enhance lifecycle sustainability.
Data Law Companion: Harnessing LLMs to Create Awareness on Data Protection Laws in Kenya, Uganda and Rwanda	Kenya	Cross-Border Data Protection Compliance	Data Governance	To collaborate with legal experts, standardise data governance practices by navigating various interpretations of data protection laws across countries.
BESHTE: A Chatbot to Enhance HIV	Kenya	AI-Powered HIV Awareness	Global Health	To improve chatbot inclusivity for diverse user groups, including those with

AI-Focused Team	Country	Topic	GPAI Area of Focus	Challenge addressed by the SRAIS Deep Dive Summary and Plan
Testing, Status Awareness, and Status Disclosure Among Adolescent Boys and Girls and Young Men and Women in Kenya				disabilities. By ensuring transparency in model operations, data privacy, and bias mitigation.
AI Solutions for Multi-Crop Leaf Disease Detection	Nigeria	AI-Driven Crop Disease Diagnosis	AI For the Environment	To scale AI tools to assist farmers with crop disease detection. By providing necessary technical support to facilitate responsible adoption of AI-based solutions.
DAWN AI Study	Nigeria	Inclusive and Transparent AI in Education	Education	To make AI tools accessible for learners with different needs, ensuring transparency in AI-driven educational decisions. Uphold ethical data practices to protect user privacy.
AI-Powered IoT for Human-Wildlife Conflict Management: Enhancing Conservation and Community Safety	Zambia	AI-Enabled Wildlife Conservation	AI For the Environment	To improve wildlife monitoring accuracy by localising training data to specific species and regions. Enhance the YOLO v5 model to prevent prediction errors.
Non-Intrusive Fish Weighing: Optimising Fish Feeding with Data-Driven Insights	Zambia	Eco-Friendly Fish Monitoring	AI For the Environment	To develop eco-friendly solutions for weighing fish without physical intervention. By ensuring minimal environmental impact in data collection processes.
Delia: AI-Powered Chatbot and	Burkina Faso	AI Health Voice Assistant	Global Health	To address data quality issues, such as bias and ethical concerns, in AI health tools. By strengthening data

AI-Focused Team	Country	Topic	GPAI Area of Focus	Challenge addressed by the SRAIS Deep Dive Summary and Plan
Voice Health Assistant				collection practices to uphold privacy and integrity.
Large Language Models for Sexual, Reproductive, and Maternal Health Rights	Ethiopia	AI-Enhanced Sexual and Maternal Health Access	Global Health	To overcome misinformation, cultural sensitivities, and language barriers in health education. By ensuring privacy and confidentiality in accessing and disseminating information.
Regulatory AI: An AI-Powered Solution for Health Compliance and Regulation	Uganda	AI for Health Compliance and Regulation	Global Health	Document management solution that uses AI to streamline GMP certification and enhance regulatory compliance for herbal practitioners and pharmaceutical companies in the Ugandan pathogen mitigation industry.

Figure 4. Overview of participating teams of the SRAIS African Track

Global track: mentoring experiences and outcomes

Knowledge Chat: Fostering a culture of responsibility in collaboration with users

Knowledge Chat is a Large Language Model (LLM)-based system developed by the team at WEBSENSA, Poland, to enable organisations to easily search and query their knowledge resources via a virtual assistant. The aim of the solution is to support improved organisational efficiency and productivity, by making it easier and faster for people in organisations to access and share information contained in internal documents, databases, policies, etc. The system consists of a "closed-loop" approach built on the use of third-party generative AI (i.e. a general purpose LLM) which is tailored and refined to the requirements of the Knowledge Chat application.

At the outset of the mentoring process a range of responsible and ethical AI issues were identified by both the WEBSENSA team and the mentors with respect to scaling Knowledge Chat. These included data governance considerations—such as the need for strict protocols and granular permissions to access sensitive organisational data. In addition, ethical issues were raised with respect to the possibility of pre-existing biases present in the training data of the third-party LLM, which could permeate into Knowledge Chat and manifest in different ways across different uses and organisational applications. Moreover the risk of hallucinations (or inaccuracies) in the system's outputs and decisions was identified as a significant challenge—and is a risk common to most generative AI systems. Inaccurate outputs could raise serious safety issues depending on the context in which the system is being used, for instance if being used to support organisational decision-making which affects people and environments. Finally, in surveying RAI issues arising from the Knowledge Chat system, the mentors encouraged the team to give greater consideration to its potential long-term societal, development and environmental impacts. One such consideration is the impact of the tool on workforce dynamics, and any possible negative consequences on job quantity or quality within organisations.

Given the range of RAI challenges identified, there was a need to prioritise the most pressing concerns to focus on in the mentoring process. The mentors and the team decided to institute a consensus-based decision making process to achieve this, based on votes during meetings, as well as the use of an internal survey. This collaborative approach arose from the team and mentors' desire to "foster a culture of responsibility" within the WEBSENSA organisation, which would extend beyond the time-limited boundaries of the SRAIS mentoring process.

The decision was made initially that the priority focus of the mentoring would be the need to mitigate inaccuracies and errors (sometimes referred to as hallucinations) in the Knowledge Chat system—being the responsibility challenge which was felt to represent the greatest immediate risk. As the mentoring progressed the objective was further clarified and refined, as the need was identified not only to mitigate this risk in the design and development of the system, but also to equip users and other stakeholders with the knowledge and tools to safely and properly navigate the system in a way that mitigates risks and harms of errors. As the team identified in their deep dive summary: “[technical] approaches are only effective if a risk-aware implementation team diligently applies them.” As such the objective became to develop comprehensive, audience-specific guidelines for implementing and using an enterprise-specific AI system in a responsible manner, with a focus on hallucination mitigation. Alongside technical quality assurance this is planned to be achieved with reference to principles of transparency and explainability, and data privacy and security.

The subsequent mentoring process and production of outputs was dedicated to developing a deeper analysis of the chosen RAI challenge, identifying the key areas that the implementation guidelines should focus on for different audiences, and drafting of the guidelines. Following the conclusion of the mentorship the guidelines are intended to be a collaborative ‘living-document’, which will evolve based on ongoing dialogue and feedback from stakeholders, and lessons learned from each client’s implementation. In addition the team plans to design an onboarding process, to engage clients actively in the implementation of responsibility guidelines.

Responsible AI Co-worker: Responsible Multi-agents AI Orchestration

India-based company AgentAnalytics.AI aspires to “automate and augment” roles performed by small and medium enterprises in the data and analytics sector, by offering versatile LLM-based services. The company has developed a platform which gives clients access to “AI Co-workers” in the form of open source LLM agents, which perform or support on various tasks that might otherwise have been outsourced to other suppliers. According to Agent Analytics about 20% of existing activities in the data and analytics sector can readily be automated by current Generative AI tools. They aim to double this and are confident that much of it can be automated in the next 3-5 years.

The company has identified the need for greater responsible AI (RAI) oversight of its AI Co-workers, and has accordingly developed an “RAI Co-worker” solution—which it submitted to the SRAIS mentorship process. The RAI Co-worker concept consists of a pool of LLM agents, monitoring the workflows and outputs of the AI Co-workers, and evaluating them based on responsibility indicators—specifically fairness, transparency, accountability and privacy. This RAI Co-worker oversight aims to facilitate the safe integration of AI Co-worker in the end-users’ organisational activities.

The main challenges that became apparent in the early phases of the AgentAnalytics mentoring process were that the AgentAnalytics Responsible AI work confined to the platform rather than addressing use cases, and the implicit assumption that measures at the platform level would insure an appropriate level of AI responsibility at the use case level. While the team paid general tribute to the importance of responsible approaches to AI, there was a need to further detail how specific responsibility principles were operationalised and validated in the design and deployment of the RAI Co-worker. The mentors also noted that some responsible AI dimensions such as privacy were interpreted in a technical cybersecurity sense but did not tackle AI ethics specifics such as leakage of personal information.

The mentors also noted that the platform flexibility was very broad, and the plug-n-play nature of off-the-shelf LLM agents elements made the control of the platform difficult. This raised the question of the extent to which Agent Analytics has influence over the responsible AI impacts of the tool, and the extent to which responsible principles are adhered to by stakeholders, including suppliers (e.g. within training datasets, models, etc.), and customers (e.g. in terms of concepts of operations and business workflows, and related issues such as privacy, transparency, explainability, and opt-out).

These questions could not be fully answered in the context of the SRAIS mentoring process as the only elements of a solution available was the platform itself. All the use cases were not defined and could not be analysed exhaustively. Therefore, it was agreed that for the rest of the mentoring process the team would focus on Responsible IA aspects that were confined to the platform and that would scale with it, namely the fairness of the RAI co-worker multi-agent consensus orchestration, as well as how to avoid leakage of private information in explanatory modes. Accountability and transparency were considered tightly dependent on the use cases outside the scope of the platform, and the explainability mechanism utilised by the platform relies on a source referencing scheme being patented by the developers and cannot be evaluated without impacting the patent discovery filing.

The fairness and privacy focal points have been tackled in three categories as evaluation, consensus, and leakage. RAI Co-worker agents are open source off-the-shelf and assessing Responsible AI dimensions using such agents brings up fairness issues at multiple levels: standalone RAI agent evaluation scores can be biassed, as the RAI agent LLM pre-training and fine tuning can influence the resulting scores. This bias can be quantified using various techniques such as embedding-based, probability-based, or generation-based metrics. AgentAnalytics.AI will propose mechanisms to check for bias and results variance at the individual RAI Co-worker level.

Also, pooling a set of agents into consensual or averaging schemes can bring fairness issues of their own. Consensual schemes often rely on reaching given thresholds, often related to exhaustion of resources (e.g., epochs, compute,...). The unfair allocation of resources for parallelized agents orchestration can skew outcomes. This can be rendered more acute in platform scaling. AgentAnalytics.AI is hence leaning towards a serialised approach to minimise resource impact, and will have to verify further orchestration aspects in the serialisation (e.g, variance in results based on the sequence of agents pooling). Further, AgentAnalytics.AI will analyse the resulting pooling and evaluate the improvement over single agents outputs to measure the scalability impact of the number of agents being serialised.

Finally, using source referencing as an explanatory means can lead to unintentional privacy information leakage. While much of it can be contained at the use case level through human in the loop schemes, the use of such source referencing in platform consensus orchestration also calls for a systematic minimization of such leakage at the platform level itself. Such minimization approach also helps the scalability of the platform at the use case level by reducing bottlenecks at the human in the loop oversight. Among other aspects, AgentAnalytics.AI is considering differential techniques to minimise the amount of information reported in the source referencing and limit it to appropriate relevant information only.

It is noteworthy that while this challenge was raised with AgentAnalytics.AI through the mentoring process, they simultaneously faced the same concern from their potential customers, testifying to the relevance of the Responsible AI analysis.

The mentoring process surfaced many RAI challenges and considerations for the Agent Analytics team, which must continue to be addressed as they seek to scale their application. One key takeaway from this exercise is that confidence in the generic knowledge of responsible AI dimensions of products can lead to oversight in individual product analysis. Responsible AI literacy, in particular in dimensions that are also tackled in other disciplines such as cybersecurity, and methodologies for product scoping and profiling help avoid much of such oversight.

ASLAC Migam.ai: From user privacy to data subject privacy in an LLM dependent on video training data

ASLAC Migam.ai is a cloud-based sign-language translation service aiming to support accessibility for the global deaf and hard-of-hearing community. ASLAC is being developed by Migam, a Warsaw-based social impact company, which provides remote video sign-language translation services for customers around the world, including prominent telecommunications companies and NGOs working in the accessibility space. Of Migam's 44 staff, 43% have a disability, and this underpins Migam's deep understanding of the barriers persons with disabilities often face in navigating digital interactions, and understanding digital content. The Migam team—being largely based in Poland and Ukraine—has observed how this issue has impacted deaf Ukrainian refugees, who face especially high barriers to communication including in finding a job. With ASLAC, Migam is working to expand the service they provide to support more users at reduced costs, through an AI/Avatar based solution which can integrate with various existing streaming platforms and services, and which will be able to translate between multiple languages and multiple sign-languages.

Migam has identified key challenges in scaling ASLAC responsibly, including the need to increase the volume, quality, diversity and representativeness of its training data, in order to ensure it is not only accurate across different languages and contexts, but also free of bias, adaptable to contextual linguistic nuances, and culturally sensitive. Within its existing platform, Migam has validated data governance practices, which are fully compliant with relevant regulation. These have largely focused on the protection of user data. However, as the Migam team began to engage with GPAI mentors, additional data governance considerations came more sharply into focus, stemming from the team's plans to upgrade to a new generation AI/Avatar based platform. Together the Migam team and the mentors decided that the priority focus of the mentorship was to develop a framework for the responsible acquisition and management of data assets for training their LLM, which the Migam team would seek to implement following the conclusion of the mentoring.

Due to the nature of sign language, a significant proportion of the data used to train the LLM needs to be video content—in order to capture gestures and facial expressions. In collecting data assets of this nature, it becomes particularly challenging to uphold both privacy and intellectual property rights of third-party data subjects. It is crucial, however, that personal data (for instance video containing faces and facial expressions) is collected and stored responsibly. Two key issues emerged for the Migam team in relation to this. First, they noted that a significant proportion of their video content for training their model was supplied by NGOs. While this data was higher-quality in general, the NGOs did not have proper agreements with the data subjects involved. A secondary source of data for Migam is streaming services like Youtube. However, scraping video content from Youtube gives rise to important intellectual property considerations.

As the mentoring process unfolded, Migam and the mentors worked together to develop an innovative plan for achieving the 'de-personalisation' of third-party training data, and subsequently disposing of raw personal data. This plan would involve the use by Migam of a "Data Clean Room"... or a method for quantising and tokenising input data, embedding of tokens, and using the resulting vector data to input into the machine learning process. In addition, in relation to data acquired from streaming services and the need to uphold intellectual property compliance, the mentors and the Migam team together identified that a filtering API could be utilised to ensure that only sources published under appropriate Creative Commons licensing were extracted.

The GPAI SRAIS mentoring process assisted the Migam team to clearly identify a priority challenge in scaling their service responsibly, as well as to establish a clear and actionable plan to overcome this challenge. In the case of ASLAC this involved broadening the focus beyond ensuring the privacy of users, to also ensure the privacy of personal data subjects in training data that is heavily skewed towards video content, and which necessitates the collection of facial expressions. The experience of the Migam team demonstrates that safeguards and risk mitigation strategies to protect the privacy of data subjects are available even in cases where raw training data might necessarily contain personal information like facial features. By developing their Data Clean Room plan, Migam has taken a key step towards responsibly scaling their service, in order to help dismantle accessibility barriers for millions of deaf and hard-of-hearing people around the world.

Jalisco diabetic retinopathy detection and referral: Developing a tool that can be easily and safely used by clinicians in different contexts

The incidence of diabetes mellitus (DM) has been increasing globally, with particularly rapid increases in low- and middle-income countries. This growth gives rise to an increased prevalence or risk of diabetic retinopathy (DR), an asymptomatic condition which, if not caught and treated in time, can lead to vision loss. DR is the leading cause of vision loss amongst the working age population worldwide, but early detection and treatment can reduce the risk of blindness by 95%. Mexico's Jalisco State Government has identified the earlier detection of DR as a pressing need in its jurisdiction, as approximately 500,000 DM sufferers in Jalisco require annual screening for DR. With only approximately fifty retina specialists in the state and limited resources to conduct sufficiently regular screening of the at-risk population, the government, through the Coordinación General de Innovación Gubernamental initiative, set out to develop AI-driven tools to assist in the early detection and referral of DR.

In the early stages of the project the team developing the application sought to learn from challenges experienced by other DR detection systems, as reported in the literature. It was found that while image-detection systems proved highly accurate in testing using specific curated training data, it was difficult to ensure accuracy when the systems were applied across different real-world contexts, and their decisions and outputs were not always clear, useful and understandable to clinicians in clinical settings. The team identified the need to improve and iterate their application to respond to these specific challenges of clinical implementation, and to ensure full adherence to RAIresponsible AI principles, including fairness and bias-mitigation, secure and sustainable data-sharing practices, transparency and explainability of system processes and decisions to users, and continuous monitoring of and responsiveness to potential adverse societal and environmental impacts.

In response to these identified challenges, the mentors suggested that the Jalisco team should give particular consideration to three main imperatives which related specifically to implementing and scaling responsibly: First, the need to ensure that outputs and decisions of the system aligned with and were sensitive to local needs, especially given that the model had been primarily trained on broader global-level data; second, the need to develop long-term strategies for secure and compliant data storage and sharing including of training data and local patient data; and third, the need to tailor the application to the specific requirements of clinicians, including seamless integration into their existing workflows, and interpretability of outputs. It was agreed that the mentoring process would focus on developing strategies to address these three challenges.

As the team and the mentors worked together on developing strategies in light of the three priority areas, specific challenges and obstacles were identified, which required further thought. These included the need to implement clear frameworks for transparency and explainability, which would enable clinicians to understand the system's decision-making process. In addition, the need was identified for more structured accountability, recourse and dispute resolution mechanisms, towhich could allow for any negative impacts to be surfaced and addressed. Finally, in order to better tailor the application to the needs of clinicians, the team identified the need to implement mechanisms for gathering and incorporating expert and stakeholder feedback.

The mentors assisted the team to consider more deeply how these RAI imperatives applied to their application, particularly in the light of their desire to scale up, and to make their application available to more users. The mentors encouraged the team to further investigate and consult on the specific needs of clinicians in order to more closely tailor their tool to the intended user base, and encourage more widespread adoption. In addition, the mentors encouraged the team to undertake a thorough environmental impact assessment, including looking at the energy consumption requirements of scaling their application.

Orange Innovation Poland: Personality AI- Balancing a commitment to privacy, with the need for traceability and continuous improvement

Awareness and concern is growing globally amongst smartphone users regarding the kinds of personal information that might be collected, exploited and shared by the apps they use. User privacy and ethical data acquisition and management are fundamental principles of responsible AI. However, having access to certain types of information about users can also help app developers to improve the services they offer and to personalise user experiences.

The team at Orange Innovation Poland’s AI Competence Center has devised a new way of bridging this tension between the importance of strict adherence to privacy protocols, and the desire to tailor apps to individual users’ needs and preferences. Their solution, called Personality AI, is an AI driven personalisation component that operates entirely on the user’s device with no external data transmission. The system works by undertaking a one-time analysis of a small amount of data on the user’s smartphone (and users can control or limit what types of data are used), to produce a personality profile of the user based on the Personality Big Five, or “OCEAN” framework. The OCEAN framework is a widely recognised method for analysing an individual's personality, based on a series of characteristics, including extraversion, conscientiousness, agreeableness, etc (see De Raad, 2000).

This personality information is stored locally on a user’s device, and apps can then automatically adapt and integrate to the user’s personality without needing to continuously collect personal information. App developers may then also derive insights into how users’ personalities impact on their behaviour while mitigating the risk of the misuse of personal data.

At the outset of the mentoring process, in considering hurdles to deployment and scaling, the question arose of the component’s usability on a long-term basis. Because the system is based on a limited, one-time insight into information held on the user’s smartphone, would this limit its ability to adapt to changes over time, for instance in smartphone configuration and user habits? The model might become out of date, and the team might be limited in their ability to make adjustments and improvements. However, it was also acknowledged that the objectives of continuous traceability and improvement may run counter to the overall intent of the system—of minimising the tracking of user data by third parties. In light of this complexity, it was agreed by the team and the mentors that the objective of the SRAIS mentoring process should be to develop an approach to traceability which could ensure long-term performance and relevance, whilst not compromising on the privacy-by-design approach of Personality AI.

Together, the mentors and the Orange Innovation team determined a series of components needed for their “traceability solution”, which included: the development of pragmatic metrics to monitor the interaction between user profile, current usage patterns, and model version; optimally distributed component architecture, including API, and traceability related use cases, delivering traceability by design; the design of a demonstrator which allows for Proof-of-Concept evaluation of the traceability solution; and the production of a “limited accountability statement” which guides app developers on the responsible real-world use and implementation of the system.

In addition, the team decided to introduce a “Privacy Guardian” component, which allows for the monitoring and adaptation of system performance to take place in a user-centric and transparent way, which is predicated on informed consent. The Privacy Guardian is intended to provide a dashboard and user notifications in order for users to manage the relevancy of their personalisation profiles, understand if their profile may be out of date due to behaviour drift, and renew their informed consent for—or opt out of—improvements or updates. Once implemented by the team, the Privacy Guardian approach aims to provide a framework for ongoing improvement to ensure the continued high performance of the Personality AI system, whilst ensuring users stay informed and in control, and without compromising on the fundamental privacy-by-design principles underpinning the system.

Dysgenesis: Navigating barriers to scaling effective AI tool to improve access to glaucoma screening

Glaucoma is one of the leading causes of irreversible blindness worldwide. In India 5.5% of cases of blindness (a total of 1.2 million) are due to glaucoma. However, screening for and diagnosing glaucoma presents unique challenges, especially in resource-constrained contexts. In response to this, a team at the International Centre for Genetic Engineering and Biology (ICGEB) in New Delhi set out to develop a deep-learning AI model which could assist in the detection of glaucoma risk factors in patients based on image recognition/classification.

The AI solution is intended to assist in the interpretation of Anterior Segment Optical Coherence Tomography (ASOCT) scans, which are used by ophthalmologists to identify angle dysgenesis—a major risk factor for glaucoma. ASOCT scans also assist clinicians to determine an appropriate course of treatment (specifically the choice between LASER or surgery). A key issue in glaucoma screening is that variabilities in ASOCT scans are often too subtle to be detectable by human experts, leading to a risk of positive cases being missed or treated incorrectly.

In recognition of this challenge the ICGEB team undertook a proof of concept study and developed a prototype device, called “PredictGAD”, an interactive smart mirror, to aid in the interpretation of ASOCT scans using an AI-based system. The key objectives of the project are to support clinical decision-making including diagnosis and treatment options, empower frontline health workers with greater knowledge, and provide cost-effective screening in resource-constrained contexts.

The team has implemented responsible measures in the development of their system, to minimise inaccuracy and bias. PredictGAD is integrated with three AI models which enables a consensus-based/cross-referencing approach in order to minimise the risk of diagnostic errors. In addition the New Delhi team collaborated with researchers at other institutions in different country contexts, to broaden PredictGAD's training data, in order to mitigate population-level biases (*i.e*. biases arising from training the model solely on images derived from one specific population whose characteristics may differ from other populations at an aggregate level) in order to ensure accuracy and generalisability of the model across various contexts.

At the outset of the SRAIS mentoring process with the ICGEB team, the mentors were satisfied that the system conformed to responsible parameters in terms of data inputs and accuracy. As such, the responsibility and scalability issues they identified arose more from a logistical and post-deployment perspective. The specific issues identified were (1) how the results of screenings using PredictGAD would be managed—for instance what follow up would be initiated for patients identified to be at risk; (2) what frameworks were in place to manage individual patient data to ensure privacy, and govern any future sharing of collective-level data; and (3) how the system would actually reach communities in need of screening.

With respect to the third issue, significant scalability challenges were identified which stem from the business model of PredictGAD, in that it relies on integration with ASOCT machines/images. Although PredictGAD itself is designed to be inexpensive and straightforward to operate, ASOCT machines *are* expensive, not portable, and difficult for providers in resource-constrained settings to access and operate. In formulating an SRAIS plan in collaboration with the mentors, the team focused on identifying measures which would assist in overcoming these three key issues.

The team worked on developing clear protocols which could be implemented to manage positive screens, wherein at-risk patients would be referred in a timely manner to specialist clinics for official clinical diagnosis of glaucoma (where appropriate). In addition they developed plans to implement strict privacy management structures for sharing identifiable patient data, and also to responsibly share anonymised patient data in publicly available forums to aid in ongoing collaborative research. Finally, to expand accessibility of the PredictGAD system to more remote areas, the team is working on a process whereby ASOCT scan images can be shared via QR code to allow for remote expert input. Greater cross-stakeholder collaboration (including government facilitation) is still needed to extend advanced eye-care technologies like ASOCT machines to underserved areas in India however, to facilitate the scaling and uptake of the PredictGAD tool.

Tribeware: Reducing manipulation and control in AI-powered employee performance improvement software

Recent years have seen a proliferation of algorithmic and AI-driven systems for maximising employee efficiency and productivity. Many of these have been criticised for being opaque, coercive and manipulative, for increasing workers' stress levels, and using punitive incentivisation (like the threat of disciplinary action based on metrics like rating or status). The team at one2tribe, a Polish software company, has set out to develop an employee performance management system which integrates AI-based technologies, based on principles of voluntary choice and personal control, transparency, intra-organisational coaching and skills transfer, along with positive rewards—which differ depending on the client company but may take the form of either small monetary rewards, gifts or benefits, or gamified virtual rewards.

One2tribe has developed Tribeware, a platform designed to help large companies address critical performance management challenges like employee motivation and engagement, the need for fast knowledge transfer, and appropriate monitoring and rewarding of achievement. Tribeware aims to support these areas by using machine learning to recommend tasks or actions to employees which are tailored to their individual performance, expertise, and knowledge gaps. AI-based systems are also used to verify the execution of tasks and provide feedback. The tasks may involve "micro-learning", as well as opportunities to convey expertise via micro-surveys—subject to strict consent and data management frameworks.

Over the course of the SRAIS mentorship process, a focus emerged on the question of how to ensure the Tribeware system provides maximum transparency—so that users can understand, for example, why they were given particular recommendations, or on what basis they received a reward. In addition it was identified as important for employees to understand where and when targets and tasks were assigned directly by managers, as opposed to automatically by the Tribeware system, in order to ensure clarity on manager expectations. The importance of transparency was identified as critical for fostering a sense of autonomy (as opposed to an experience of control), and building trust with users, both employees and employers. Given that the system is based on voluntary use—trust, confidence and understandability are very important in facilitating uptake and scaling.

The mentors worked with the one2tribe team to develop an approach to further improve transparency and user education. In doing so four key areas emerged for the team to address. These focus areas encompass the key outcome of the mentoring process, and will provide a foundation for the one2tribe team to continuously improve transparency whilst scaling their Tribeware platform. The approach is as follows: First, the system needs to provide a "trace" rationale for the full sequence of decision-making that leads to a recommendation or reward. Moreover, users should have access to additional information related to AI decision-making including confidence levels and accuracy determinations. Second, there is a need for full transparency with employees regarding which decisions are made by managers, and which are made by the platform. Third, users should have the right to enquire about the reasons behind, as well as to refuse specific decisions, and to provide feedback on their experiences. Fourth, one2tribe should proactively and regularly survey users to monitor their level of understanding of AI decisions, and take further action if required to improve understanding and transparency.

Orange Innovation Poland: Satisfaction recognition to improve customer interactions with virtual agents (VAs)

The Research and Development team at Orange Innovation, Poland, are working towards developing a system which would automatically recognise customer satisfaction/dissatisfaction in interactions with Virtual Agents (VAs)—for example customer service chatbots or voice assistants—and allow VAs to adapt in real time to improve satisfaction. The aim is to enable VAs to be more responsive to individual user preferences, mood and emotion, in order to improve the quality and outcome of interactions. This would be achieved through the development of a complex, multimodal system which includes natural language processing (NLP) capabilities across text, audio and video interactions, alongside sentiment analysis capabilities, which would analyse user inputs to pick up on nuances in language patterns which indicate satisfaction or dissatisfaction. The system would also continuously collect data from user interactions, including interaction logs, user feedback and contextual information to enable continuous refinement and improvement.

During the period of SRAIS mentoring, the system was in the relatively early stages of development, which allowed for detailed discussions to take place around potential responsible and ethical shortfalls, and how to ensure responsible and ethical parameters from the outset. The mentors and the team identified a series of challenges and priority issues. Many of these revolved around the need for clear definition—of use cases; possible scenarios of interaction with users; use environment/context; and how users would be segmented. Clarifying and defining these questions would enable responsibility considerations to come to the fore and be addressed systematically. Discussions also took place around key data and privacy issues. These related both to the training data which would inform the system's decisions, (how would the team ensure that it was sufficiently sensitive to the cultural/linguistic/gender contexts in which the system was being applied, in order to make accurate and personalised inferences related to user mood and emotions, based on subtle linguistic and behavioural nuances?) as well as the handling of data collected from interactions with users, and how privacy would be upheld and compliance with various privacy regulations ensured.

In addition the mentors encouraged the team to define their commercial objectives and scalability goals—which would directly inform responsibility considerations. For instance, were targets related to making information more readily available, positioning companies, driving sales? Specific responsibility issues arise from these various objectives, and an important area of focus in the mentoring discussions related to the need to preserve transparency and user autonomy whilst avoiding psychological manipulation of users, including where profit incentives (like sales targets) are present. A key part of ensuring users are not manipulated relates to disclosure. Given that one of the key aims of the system is to make VA interaction more natural and human-like, the mentors and the team agreed that it was imperative that users must be fully informed they are interacting with an avatar, and that they have access to further information about the system's "rules of engagement".

The identification of these challenges underpinned the direction taken in the mentoring process, which culminated in the production of a document intended to more clearly define the uses, objectives, and “rules of engagement” of the system, and to set out the team’s approach to mitigating and minimising possible risks and harms to users. This was broken down into four “challenge areas”, with approaches and principles developed for tackling each challenge. (1) The first area was responsible interaction design which focused on strategies to avoid manipulation of users, and also to ensure contextual relevance and sensitivity. (2) The second challenge area was achieving transparency without undue disruption of the user experience—which the team will approach by providing “basic transparency” for all users with ready access to more detailed information where required. (3) The third challenge area was the need for detection and correction of errors and inaccuracies in the system’s interpretation/interactions in real time without disrupting the flow of interaction. The team plans to achieve this by designing “scalable algorithms that dynamically and subtly recognize and correct errors.” (4) The fourth challenge area was identified as the need to ensure that the “emotional” responses of the system were consistent but also able to be tailored to different applications/uses as appropriate, and able to take into account personal and cultural diversity of users. The team plans to address this by developing a framework for trust based “empathetic” responses based on a library of templates for different interaction scenarios.

African track: mentoring experiences and outcomes

Greenlive Agriculture : Precision and control at your fingertips

Based in Cameroon, the team comprises a group of students from different engineering schools in Cameroon with expertise in artificial intelligence. Their project “Greenlive Agriculture : Precision and Control at Your Fingertips” seeks to solve issues of water requirements in tomato crops, aiming to reduce the latter using responsible AI (RAI) means. The team benefited from mentorship expertise in artificial neural networks, model explainability, image processing and Internet of Things (IoT) in the context of health, agriculture, and water management.

The team came into the mentorship facing various challenges to the practical implementation of RAI principles and scalability. These obstacles ranged from scoping their chosen issue to the unavailability of local African agricultural datasets. The latter component being crucial for developing and scaling a solution to the problem they identified that is tailored to the context they are making an intervention in. As crops in Australia or even Morocco don’t show the same characteristics or share the same requirements as those in their chosen Cameroonian areas, the participants initially ran into an impasse.

Through the mentorship program, the team had the opportunity to learn and implement research and data collection practices from the experts supporting them, who soon helped them realise their project objective required more specificity and that some of their practices and methodology would inherently embed problematic values into their AI solution and compromise their findings. It was thus imperative that they be creative in sourcing the data required to undertake their project and work with what was available to them. This group also understood the importance of community engagement and adapting their AI solution to the particular local contexts of the farmers they are developing the tool for.

Incorporation of these insights allowed for a more targeted, fine-tuned, robust and ethical final product. The team showed resilience in moving beyond their limitations to develop a responsible solution to excessive tomato crop water requirements that empowers local farmers and fosters sustainable agricultural practices. Moving forward, they intend to build partnerships with local agricultural entities, cooperatives, and educational institutions as well as with the Ministry of Agriculture and IRAD (Institut de Recherche Agricole pour le Développement) to secure open access to public agricultural data and research toward AI development.

AI Solutions for multi-crop leaf disease detection

Representing Nigeria’s University of Port Harcourt, this team blends expertise in AI ethics and safety, data management, machine learning, mobile development, and virology. Their project, titled “AI Solutions for Multi-Crop Leaf Disease Detection,” aims to responsibly develop and scale an AI-powered system to enhance early detection of crop leaf diseases in order to bolster agricultural productivity and minimise crop losses.

Their project benefited from the guidance of mentors with expertise in machine learning, geospatial technology and data mapping, who assisted the team in completing their objective of designing their agricultural AI solution and creating a roadmap for its responsible development and deployment.

The primary challenge for this team lay in imagining and developing their crop disease detection system while adhering to responsible AI and ethics principles, as well as effectively addressing the rural Nigerian context's unique challenges to the successful scaling of their product. From biassed datasets, to the sheer variety of diseases that could be present in the crops, to varying levels of technological literacy in their stakeholders, the team had several difficulties to address before being able to produce a solution that would effectively respond to the setting in which their system would be deployed.

Following the mentorship programme, the group reported gaining valuable insights into responsible data collection and handling, model performance and accessibility. These were crucial for enhancing the responsibility and scalability of their AI-based solution, but are also transferable skills towards the conduction of future AI projects for the benefit of all. The team shared that the programme encouraged them to develop a posture of perseverance and adaptability, enhancing their research skills and overall confidence in their project.

The team remarked on the benefits of collaboration among peers from diverse backgrounds, which broadened their perspective, refined their research approach and highlighted the added value of interdisciplinary collaboration. Looking ahead, the group plans to design and implement an initial AI-based model for leaf disease detection within the next year. Their goals include expanding the solution to additional crops and regions, and collaborating with local governments to promote education and training in AI technologies in rural Nigeria to foster inclusive and equitable AI-based systems and a thriving agricultural industry.

Kit for Council (K4Council)

Based in Cameroon, this team consists of seven members operating within LivingSeedsLab, with four regularly taking part in the development of this project. The organisation focuses on advancing sustainable development goals by supporting rural communities with connectivity, agriculture, education, and healthcare solutions. It develops community networks, localised technologies, language preservation tools and affordable weather stations to empower underserved populations across Africa.

With a strong background in IT, this team of software engineers, project managers, and product designers created an administrative solution to streamline routine tasks. Their project, titled “Kit for Council (K4Council),” seeks to digitise, store, and archive documents issued by various state registries in a mobile application, allowing town hall staff the convenience of using smartphones for the process. The team benefited from the leadership of experts in the fields of data mining and artificial intelligence

The project’s objective consists in standardising processes relating to digitisation, storage and management of official documents, as well as providing users the ability to access the latter via a mobile application. This, however, proved to be a challenging endeavour due to the sensitive nature of the information involved, which raised concerns of privacy, security, transparency, and explainability. The LLM that powers the application provides only the specific information requested by a given user.

The mentorship programme suggested to the team members that they form a partnership as soon as possible with town halls in their area of residence in order to obtain real data. The team has begun the process of building these relationships and, in the meantime, is training the LLM model on data from various documents including articles and books.

The mentorship programme has helped this team understand that their AI product could not be scaled effectively unless they tackled issues of transparency and explainability head on. They have begun the process of bridging the digital divide in their communities by creating a framework to bring AI to remote and underserved communities and fostering digital inclusion by providing indiscriminate access to information to all. By 2025, the team hopes to secure partnership agreements with a few town halls in order to develop a first complete prototype and access genuine local data. Their goal is to achieve a fairness score of 0.9 out of 1, reflecting successful bias mitigation within their model and marking a step towards responsible AI integration in their region.

AI-Powered IoT for human-wildlife conflict management: Enhancing conservation and community safety

This team is composed of five members from Mulungushi University and Kwame Nkrumah University who hold a diverse set of expertises spanning wildlife conservation and environmental research, AI and machine learning, stakeholder engagement and regulatory compliance. Their project “AI-Powered IoT for Human-Wildlife Conflict Management: Enhancing Conservation and Community Safety” aims to mitigate human-wildlife conflict (HWC) in Zambia by leveraging AI and IoT solutions toward the protection of both humans and animals.

The team upholds values of technological responsibility, community inclusivity, and environmental stewardship that it seeks to translate into a roadmap for scaling its AI solution. They were advised on navigating the challenges of their project and strengthening its ethical foundations by a team of mentors with expertise in applied AI and the development of scalable Ai-driven IoT solutions in agriculture.

The main challenge facing the team was the ethical development and scaling of a solar-powered AI and IoT system designed to detect wildlife and prevent dangerous encounters with neighbouring human communities. This task was complicated by the potential for the system to inadvertently record human activities, as well as the difficulty of capturing the full diversity of Zambian wildlife in the AI training datasets. Additionally, resistance to adopting AI and IoT technologies within rural communities may pose significant barriers to scalability.

Upon completion of the mentorship programme, the team has acquired valuable insights into responsible data privacy policies, algorithmic fairness, and effective community engagement strategies. Notably, the team reports the experience taught them the critical importance of integrating Responsible AI principles from the project's inception – also known as RAI by design – rather than attempting to mitigate ethical issues that emerge further along the AI lifecycle.

This team has found the mentorship programme to be an invaluable resource that has markedly improved the effectiveness of their project’s development and design. They expressed high regard for the experience, highlighting its role in fostering a more thoughtful approach to their work in Responsible AI. Moving forward, the team is focused on refining their system and will explore further community engagement strategies to ensure successful implementation and adoption of their AI-powered solution.

Harnessing LLMs to create awareness on data protection laws in Kenya, Uganda and Rwanda

This team is based in Kenya and brings together expertise in software development, AI, data science, and legal compliance. The group is developing a project titled "Data Law Companion: Harnessing LLMs to Create Awareness on Data Protection Laws in Kenya, Uganda, and Rwanda*.*" This initiative aims to empower East African businesses and organisations to achieve regulatory compliance using large language models to simplify access to complex data protection laws and regulations.

The team was assisted by mentors in research design, artificial intelligence, data analysis, and statistics who joined their technical expertises to guide them in the design, development, implementation and deployment of their AI product. Their knowledge of IoT, digital forensics, and data security across fields such as health, agriculture, and water management guided the group in identifying risks and challenges along the AI lifecycle and mitigating them with actionable insights.

The primary challenge they faced was ensuring fairness and impartiality in their Data Law Companion's chatbot and summarisation tool, as well as the creation of a robust data governance framework. They set out to ensure the datasets they used to train their model were representative of their target population, but this task proved more challenging than anticipated given the complexity and nuance of the digital regulatory landscape in East Africa and rapidly evolving nature of the field. How can they build a tool that is able to provide information to an organisation based in Kenya, but that operates in Uganda, for instance ?

The mentorship programme had an important impact on the team and provided them guidance on integrating responsible AI principles at critical stages of the AI lifecycle, from creating data governance guidelines to deploying their model sustainably. The team states that the mentorship helped refine their approach, enabling them to embed fairness, accountability, and transparency into their project by design.

The group reports that the programme reinforced their commitment to scaling the Data Law Companion into a robust platform that addresses regional data protection needs while upholding ethical AI standards. Looking ahead, the team aims to develop a practical, scalable, and financially sustainable solution that supports legal compliance across East Africa and is rooted in RAI principles.

DAWN AI Study

The DAWN AI Study team is based in Africa but aspires to help learners worldwide. This EdTech initiative is dedicated to inclusivity and accessibility in learning platforms within the educational sector. The organisation aims to make education accessible and inclusive for all learners by leveraging AI to provide innovative, affordable, and scalable learning solutions. Its mission is to enhance learning experiences by addressing challenges of accessibility, diversity, and inclusivity through technology. It offers a suite of AI-driven educational tools, DAWN AI Study being the latest.

The team's project "DAWN AI Study: Facilitating access to education with AI" seeks to tackle the issue of limited accessibility of educational resources, ensuring that all learners – particularly those living with disabilities – can obtain educational content. Their mentorship was guided by a team with extensive expertise in research design, artificial intelligence and machine learning.

DAWN AI Study addresses a significant issue, that of the accessibility of educational resources in Africa and beyond, which can be undermined or made difficult by individuals' socioeconomic status, abilities or underfunding of public services, among others. The project aims to assist African students, educators, and educational institutions, but also focuses on individuals seeking employment, human resources professionals, and organisations that require AI solutions for recruitment and upskilling. In the context of this mentorship, the team is particularly focused on creating a safe and trustworthy platform accessible to users with learning disabilities.

Throughout the mentorship programme, the team took part in workshops with their mentors to refine their project. They learned the importance and added value of continuously evaluating AI algorithms for bias and enhancing data privacy. They were taught how to make an AI system truly inclusive by designing personalised features according to the main needs and challenges of users who are differently abled, as well as the barriers to education experienced by others due to differences in language, for instance. The mentorship experience enabled the DAWN AI Study team to refine their ethical data practices and bias mitigation strategies. This foundation ensures that their product can scale responsibly within the educational sector.

Future goals include the expansion of the DAWN AI Study platform to engage a broader audience of schools, learners, and educators across Africa and other developing regions. The organisation aims to scale its AI-driven tools, enhance existing features, and establish partnerships with educational and corporate institutions. Their long term objective is to offer an inclusive, ethical and trustworthy AI learning platform that allows learners with disabilities to learn unhindered while ensuring that educational AI tools are safely accessible to learners of all abilities and social backgrounds..

Non-intrusive fish weighing: Optimising fish feeding with data-driven insights

These four scholars and researchers from across Africa bring a diverse range of expertise spanning computer science, natural language processing, operations research, and data science. Their project, titled “Non-Intrusive Fish Weighing: Optimising Fish Feeding with Data-Driven Insights*,”* aims to optimise fish feeding practices in Zambia’s tilapia farming industry, harnessing AI to reduce harm caused by traditional manual weighing methods.

The group’s mentorship team leveraged expertise in incubation and mentorship of innovative RAI projects to guide the team on best practices, particularly with respect to minimising bias in their datasets and outputs and increasing the robustness of their data collection and implementation framework. The team also benefited from insights into project management aspects that helped them in the scoping and definition phase of their project, among others.

Fish weighing is a common practice in the aquaculture industry, carried out to research the impact of feeding on fish growth. This technique can, however, be tedious and cause stress to the animal, compromising its welfare and the quality of the final product. By developing a non-intrusive, data-driven solution, the team hopes to reduce harm while providing accurate data for monitoring fish growth. Scaling the project responsibly presents several challenges, particularly in managing the data used to train and operate the model. Key challenges include securing proper data collection authorisations, creating robust data use agreements, addressing privacy and security risks, and managing the complexity of diverse data sources.

The team collaborated closely with their mentors to devise a comprehensive data governance framework, addressing the ethical and technical challenges of data management. Importantly, they learned the importance of a stakeholder-first approach to designing an AI solution, guiding them to ensure that their AI tool is tailored to meet the everyday needs of those working in the fishing industry and enhancing their tool’s adoption potential.

The mentorship programme impacted the team's approach, deepening their understanding of Responsible AI principles and reinforcing the importance of meaningful interactions with their target audience. They now hope to pitch their project to external stakeholders, including Zambia’s Ministry of Fisheries and Livestock and the Zambia Environmental Management Agency (ZEMA). The team’s next steps involve aligning their computer vision solution with national sustainability and regulatory goals as they advocate for ethical, sustainable, and non-invasive aquaculture practices.

Delia: AI-powered chatbot and voice health assistant

The AI-KING Group brings together a diverse team of AI experts, medical doctors, and software engineers. Harnessing their collective knowledge, they aim to address healthcare challenges, particularly in chronic disease management for underserved populations in Burkina Faso, with special attention to patients diagnosed with hypertension and diabetes. Their project, “Delia: AI-Powered Chatbot and Voice Health Assistant,” seeks to develop an ethical, patient-centred AI tool that enhances care-seeking behaviours and communication while ensuring adherence to data protection laws and fostering sustainable public-private health partnerships. By making medical information pertaining to individuals’ health directly accessible to them through a chatbot, they aim to reduce delays in access to healthcare and empower individuals to gain insights into their health regardless of their location.

The team was led by experts with backgrounds in pharmacy, epidemiology, biostatistics, machine learning, AI and data science. Together, their mentorship assisted the team in gaining the necessary skills to navigate the intersection of AI, healthcare, and data governance, including the challenges this nexus can present. As Delia assesses users’ symptoms to direct them to appropriate healthcare services if needed, the tool handles sensitive health data. The ethical management of this data and compliance with national and international laws and regulations is crucial to scale this project responsibly. Their main challenges, therefore, involved transparent and accountable data collection and use practices, and establishing a robust data governance framework to assure compliance and protect users' privacy.

During the mentorship programme, the team was able to draft a comprehensive research protocol and create a sound data governance and protection framework, enhancing their tool’s compliance. They have initiated preliminary steps toward establishing a collaboration with Burkina Faso’s Ministry of Health to gather data and work toward their aspiration of implementing a pilot of the Delia solution in public hospitals. The team reports that the mentorship programme had a transformative impact on them, providing valuable insights into navigating the regulatory landscape of health data, building their capacity to engage with public stakeholders and setting a strong foundation for scaling their AI-driven health solution.

Looking ahead, the team plans to take steps toward requesting support from the government to authorise the deployment of their AI solution in public health centres. Next steps include fundraising, launching a pilot version of Delia and furthering the collection of data to refine their AI model and improve patient triage and referral. The team aims to expand the deployment of their AI solution across other regions by seeking support from global partners such as the World Health Organization (WHO), the United Nations Development Programme (UNDP), and other multilateral organisations.

AI4Health Project: Strengthening the RAI pipeline for Mboacare and Mboathoscope”

The AI4Health team brings together a multidisciplinary group from Cameroon, combining expertise in molecular biology, AI and data science, industrial robotics, software engineering and stakeholder engagement. The team is affiliated with Mboalab, an organisation dedicated to fostering community education, healthcare accessibility, and sustainable innovation in the region. Their project, “AI4Health: Strengthening the RAI Pipeline for Mboacare and Mboathoscope”, aims to improve healthcare access and accuracy in low- and middle-income countries, focusing on accessible, accurate infectious disease screening via AI-driven tools that facilitate remote diagnosis and access to healthcare resources.

The team benefited from mentorship expertise in community networks and data justice, as well as mathematical modelling, epidemiology, and statistical methods, gaining critical insights into the impact of data science on public health to incorporate into both parts of their project. The Mboacare app aims to improve healthcare access by helping patients locate registered facilities, while the Mboathoscope facilitates remote consultations by capturing and transmitting stethoscope-recorded sounds to doctors for telemedicine.

The AI4Health project was developed in response to the prevalence of misdiagnoses due to limited medical equipment and diagnostic expertise in underserved areas. Responsibly scaling their solution came with challenges with respect to ensuring robust data privacy and ethical standards, addressing inclusivity by incorporating data representative of minority populations, and customising solutions to meet local needs. The project also faced technical and infrastructural hurdles, such as ensuring operability in regions with unstable networks and limited technological access. The team was committed to addressing issues of data integrity and preventing misinformation, all while localising their solutions to fit local healthcare protocols, literacy levels, and cultural contexts.

Through the mentorship programme, the team learned how to implement comprehensive data privacy policies toward building a robust governance framework, adopt culturally appropriate consent procedures, and develop transparent AI models that foster user trust. They also gained insights into creating representative datasets to reduce bias while ensuring the inclusivity of underrepresented populations. Furthermore, mentors emphasised the importance of adapting their project to the local context by collaborating with public health organisations and language experts to support the integration of AI4Health into Cameroonian healthcare. The mentorship also highlighted the need for offline features, allowing the project to overcome infrastructure constraints and reach a broader audience.

Looking forward, the Mboalab team plans to localise Mboacare and Mboathoscope by partnering with local health centres and universities to gather culturally relevant data in order to create a solution that responds to the needs and realities of Cameroonians in the healthcare sector. They also aim to translate their system into local languages to broaden its accessibility. To achieve their goals, they seek support from the Ministry of Health and other institutions to facilitate the redeployment of their tool using Indigenous data, as well as assist in community outreach and in providing funding for data gathering. Their ultimate vision includes building capacity, strengthening open science, and fostering sustainable development across the healthcare landscape in Cameroon.

Large language models for sexual, reproductive, and maternal health rights

The “Large Language Models for Sexual, Reproductive, and Maternal Health Rights” project was developed by researchers at Debre Markos University, University of Gondar and Addis Ababa Science and Technology University, Ethiopia. Leveraging expertise in machine learning, data science, and software engineering, this project aims to bridge gaps in healthcare access and information by using Large Language Models (LLMs) to improve awareness, access, and service quality for sexual, reproductive, and maternal health in underserved communities.

Guided by mentors with expertise in applied AI, data science and digital innovation, the team has gained important insights into developing and applying AI solutions responsibly within healthcare contexts. Notably, the mentorship highlighted the need for cultural sensitivity in developing a reproductive health solution, robust data privacy frameworks to protect confidential user data, and inclusive design to navigate challenges such as data protection, infrastructure limitations, and varying literacy levels – technological and otherwise – among the target population.

Through structured workshops, the team refined their project by integrating cultural and ethical considerations, enhancing data privacy protocols, and creating digital literacy materials to bridge the digital divide and enhance access to their solution. Furthermore, they established engagement strategies involving local health professionals and community representatives to ensure ongoing relevance of their project and community adoption. The team’s mentorship experience underscored the value of an adaptive, iterative development approach aligned with sustainable development goals, equipping them with a foundation to advance Responsible AI in healthcare.

The mentorship helped foster a collaborative, interdisciplinary mindset in the team, increased awareness of the unique challenges of health-related AI projects, and equipped them with effective stakeholder engagement methods. The structured nature of the sessions assisted them in addressing various challenges related to responsible AI scalability and implementation in healthcare. The mentors' expertise and insights were instrumental in identifying and overcoming their challenges and ensuring that their project aligns with ethical standards and sustainable development goals.

Moving forward, the team plans to finalise and implement enhanced security measures, develop and distribute educational resources, and initiate engagement meetings with local health professionals and community leaders. Ultimately, they plan to implement and commercialise their project after clarifying its potential benefits through the mentorship exercise. The team seeks governmental support at both national and local levels to streamline bureaucratic processes, including facilitating discussions, granting access to essential data, supporting training and demonstration initiatives, and expediting the authorization processes required from various government agencies.

BESHTE: A chatbot to enhance HIV testing, status awareness, and status disclosure among adolescent boys and girls and young men and women in Kenya

The Innovate AI Health Lab team leverages natural language processing to address critical healthcare needs in sub-Saharan Africa. Based in Kenya, the team includes faculty from the University of Embu's Computing and Information Technology and Mathematics departments, alongside a private practice psychologist. With expertise spanning AI, health informatics, HIV modelling, and counselling psychology, they bring a multidisciplinary approach to their project, "BESHTE: A Chatbot to Enhance HIV Testing, Status Awareness, and Status Disclosure." The chatbot seeks to increase public awareness and education about HIV testing, status awareness and disclosure, particularly among Kenyan adolescents and young adults.

The project benefited from the mentors' expertise in AI and geospatial technologies, as well as machine learning, statistics, and epidemiology. The combination of their experiences has helped provide critical insights into leveraging and scaling this AI system responsibly, focusing on the ethical and technical challenges associated with healthcare technology development. Their interdisciplinary mentorship facilitated an informed approach to designing an AI solution sensitive to privacy, inclusivity, and data integrity.

The team identified the core issue of limited HIV-related information for young people in Kenya, often manifesting as misinformation and gaps in status awareness and access to testing options. Developing and responsibly scaling their solution to this issue involved key challenges such as ensuring privacy protections for sensitive health data, enhancing inclusivity by addressing barriers like language and the availability of technology, and ensuring data integrity. The development of their chatbot included putting in place measures such as rigorous data collection practices to ensure the reliability of the information provided as well as including human oversight and seeking out relevant expertise to uphold trustworthiness.

Throughout the mentorship, the team focused on enhancing data integrity to improve the accuracy and quality of the information provided by their chatbot. They were guided in implementing data curation protocols, including comprehensive cleaning and expert peer review processes, to build a reliable knowledge base for the chatbot. The mentorship also empowered them to outline a detailed scale-up plan and project deployment timeline, setting them on a pathway to extend the reach of their AI-powered solution. Importantly, the experience supported the team in embedding ethical principles into the development of their solution across the entirety of the AI lifecycle.

Looking ahead, the team seeks support from national agencies to secure funding and bridges to scalability resources. Their future plans include pursuing the development of intelligent agents that can understand, learn, and guide young people on health issues, along with culturally sensitive materials relevant to local communities. They also aim to extend BESHTE's scope to address broader reproductive health topics, such as maternal health and breast cancer prevention and care, in Kenya and beyond.

Regulatory AI: An AI-powered solution for health compliance and regulation

Founded in Uganda, Feyti Medical Group has developed an AI solution aimed at transforming pharmaceutical regulatory compliance. The team brings together expertise across pharmacy, software engineering, project management, and business development. The company is developing “Regulate AI,” an AI solution designed to streamline regulatory compliance by automating critical processes, thereby reducing administrative errors and ensuring that pharmaceutical regulatory guidelines are met efficiently and accurately.

The team’s mentors provided complementary subject matter guidance. Their respective expertises in RAI assessment and development, and healthcare and policy provided a well-rounded perspective, addressing critical aspects essential to responsibly scaling the regulatory solution from different vantage points. The mentors supported the team along the AI product’s lifecycle, notably in implementing responsible AI practices into their product by design, enhancing stakeholder engagement, and developing strategies for achieving regulatory alignment.

The primary issue addressed by the project lies in the inefficiencies within bureaucratic processes in pharmaceutical compliance, often leading to errors and delays in certification. To centralise documents and automate different manual processes requires access to and use of sensitive data, which can pose some risks for the user, among others. The mentorship sessions revealed critical challenges, including the need for a robust data governance framework centred on privacy, security, and local regulatory compliance. Additionally, stakeholder engagement was identified as a core component of project success, particularly involving structured partnerships with regulatory bodies and end-users. The team also recognized the importance of securing ongoing funding and expertise to responsibly scale up their solution while addressing ethical considerations.

Throughout the mentorship, the team engaged in discussions on ethical AI practices and received expert guidance on technical problem-solving, stakeholder alignment, and regulatory compliance. Workshops underscored the importance of data governance, ensuring data privacy, assessing risk, and optimising team composition. These experiences equipped the team with the tools needed to refine their platform’s ethical and regulatory standards, embedding fairness, transparency, and bias mitigation into their development process.

Collaborating with field experts allowed the team to bridge the gap between technological innovation and regulatory compliance. Looking ahead, the team seeks support from the Ugandan government to access incentives that promote responsible AI innovation and development, along with guidance on aligning with local and regional AI regulations.The team aims to implement a pilot version of their solution by the end of 2025, with plans for an official launch and scaled deployment to reach a broader audience within the same year.

AI-driven carbon credit calculations for electric motorcycle fleets

The YNA Kenya team comprises team members with expertise spanning carbon credit calculations, electric motorcycle energy estimations and AI. YNA Kenya, a social enterprise specialising in technology solutions for climate change and women's empowerment, provides sustainable delivery services using electric motorcycles. The team's project, "AI-Driven Carbon Credit Calculations for Electric Motorcycle Fleets," seeks to minimise carbon emissions in delivery services by deploying AI-driven emissions calculations, focusing on deliveries completed using motorcycles as the primary mode of transportation.

Guided by experts versed in AI system development, data quality management and responsible AI assessment frameworks, the YNA Kenya team has developed strategies for promoting transparent data collection and governance and improving stakeholder engagement.

The team's project addresses the challenge of emissions from delivery motorcycles, particularly in underserved areas. Their primary obstacles included ensuring precise carbon credit measurement, maintaining data integrity, and responsibly scaling infrastructure to accommodate the AI model's activities. These challenges required a tailored approach, with a focus on transparency in carbon credit predictions, stakeholder inclusion, and alignment with responsible AI principles.

Throughout the mentorship programme, the YNA Kenya team gained insights for strengthening AI model accuracy and transparency and integrating robust data verification methods, such as real-time tracking to detect fraudulent data. The programme also emphasised the expansion of stakeholder engagement methods beyond drivers to include package recipients and regulators. This approach helped ensure inclusivity in model development and compliance with local carbon credit regulations. Furthermore, mentors advised the group on enhancing team diversity, including broader AI and IoT perspectives, and fostering transparent governance with adaptive, collaborative frameworks.

The mentorship experience enabled the YNA Kenya team to refine their AI model for responsible carbon credit management and enhance sustainability in motorcycle deliveries. Moving forward, their next steps include developing adaptive governance guidelines, expanding transparency, and deploying their solution in urban areas while involving women drivers and supporting people with disabilities in rural deployment. Their future ambitions encompass advancing climate change mitigation strategies, ensuring secure carbon credit management, and contributing to Responsible AI in transportation and environmental sustainability.

Discussion: Reflections from the 2024 cycle

A need for greater RAI literacy

RAI is an emerging and consolidating field, which cuts across research, industry, government and civil society and encompasses a broad range of principles and perspectives. While efforts have been made at multiple levels to capture and define the suite of considerations that make up RAI, into frameworks, standards and guidelines, operationalising these frameworks in developing and scaling specific AI tools remains a complex task, and knowledge shortfalls exist amongst various stakeholder groups, including suppliers, developers, clients and users, which need to be addressed in order for RAI to become institutionalised and ubiquitous.

The experience of the SRAIS mentoring cycle in 2024 has highlighted the need for nuanced understandings of RAI principles, their rationales and intentions, amongst those seeking to implement them while initiating and scaling real-world projects. Alongside the codification and standardisation of RAI principles including in regulation, there is a need to equip AI teams with the knowledge which would allow them to identify and define RAI challenges in their unique contexts and applications, in a way that is dynamic and reflexive. A meaningful commitment to responsible and ethical AI must be grounded in a comprehensive understanding of the range of social and environmental harms AI-based systems (for instance LLM-based tools) can produce, and an ability to apply such an understanding to a specific AI-based system's uses. This requires more of AI developers than just being familiar with, and even ensuring compliance with codes, standards and metrics—it requires an ability to critically and dynamically reflect on how responsibility considerations arise in unique contexts. Projects such as SRAIS which seek to undertake deep dives into RAI principles as they relate to existing applications, and to capture and share the knowledge gained from these exercises, can help to build greater RAI literacy and awareness amongst practitioners.

As identified as part of the mentorship of some teams this year, RAI literacy should also not end with the engineering team, but must also extend to clients and end users—particularly where scaling is the goal—to ensure that AI-based applications are not only built, but also used in responsible ways. AI teams are well placed to transmit such knowledge, to educate users and to learn from users' experiences in turn. In addition there is a need for more high-level public and private sector initiatives to promote RAI literacy around the world.

Technical solutions for responsibility are only part of the picture

Those working in the AI development space, with expertise in technology and engineering, are well-placed to craft innovative technical approaches to integrating and encoding responsibility into AI applications. During the SRAIS 2024 mentoring cycle, several teams developed or refined sophisticated methods for advancing RAI principles like accuracy, transparency, privacy, and oversight, within the components of the applications themselves. For instance, the "data clean room" approach developed by Migam, whereby a system reliant on facial expressions, human gestures, and other similar training data, could nevertheless come to comply with individual and collective privacy principles by adding additional features. Another example is the use of inter-agent or inter-model consensus techniques employed by ASCOT's PredictGAD, and Agent Analytics' RAI co-worker, to improve the accuracy and reliability of system outputs, in order to guard user safety. Such approaches which seek to achieve responsibility best practice through technical innovation itself can provide examples and inspiration to AI developers across different applications and contexts.

However, the SRAIS project experience also highlights the fact that technical solutions to responsibility challenges cannot form the whole, or the only, basis on which AI teams engage with responsibility principles and concerns. Our mentoring sessions showed that responsible AI itself is an evolving and multifaceted terrain which intersects with distinct AI-based applications on technical, social, geographical, economic and other levels. For instance, it is not sufficient to develop innovative methods to protect user data security if the rights and privacy of data subjects (who generate training data) or data workers (who process training data) are not upheld during the development of a tool. Likewise, it is not sufficient to encode features which reduce the risk of AI errors or 'hallucination' if the team building an AI system has not meaningfully consulted on and grappled with the specific needs and wishes of the end-user communities the tool is intended to be used by.

Although the SRAIS project prioritises one particular responsibility challenge area for each team in order to make concrete progress and generate generalisable insights within a limited timeframe, it is also important to recognise that achieving responsible and ethical best practice requires an holistic approach which takes into account all possible facets of responsibility. Ultimately an application which only falls short on one responsibility component and upholds all others, may still cause significant harm when it is scaled. Similarly, responsibility principles, and approaches to upholding them, do not occur in isolation but are often interconnected and interdependent. They influence each other and cannot easily be weighed or traded-off against each other. For instance, a technical approach to mitigating the bias of AI-based systems may necessitate the significant expansion of training datasets and data processing capabilities. However, doing so may raise additional data governance and data-subject rights concerns, or may require the labour of precarious low-waged data workers, or the consumption of more energy and water resources for processing. In another example, increasing system-output reliability by adding additional technical layers of assurance, or new models which oversee and monitor existing models, may be undertaken in the service of reducing the need for human oversight, thus leading to reduced accountability to humans, and evolving social conceptualisations of what constitutes responsibility in AI, as well as potentially to job displacement.

Improving technologies themselves in order to better align with responsible and ethical frameworks is critical. But responsibility can be both enacted, and undermined, in other aspects of the AI lifecycle, from the way a tool or system is conceived, by whom and for whom, and the values inherent in that conceptualisation, to the process of how it is built, governed, and deployed. Responsibility and ethics are inherently a human, social and political domain—and it should remain the task of humans to continually and collectively define and enforce their parameters and adherence. This can and must be underpinned by a pursuit of technical excellence and best practice, driven by those with technical expertise. But a key lesson from the SRAIS project in 2024 is also that responsibility itself cannot be automated—it requires dialogue and collaboration between multiple stakeholders. Providing a platform for such dialogue to take place is a key objective and impact of the SRAIS project.

Lessons learned and policy and practice recommendations

The experiences of the teams and mentors over the course of the 2024 SRAIS cycle has given rise to recommendations for policies, practices and approaches which can help to enable responsible AI applications to scale responsibly. Below we contextualise our recommendations within a series of lessons learned, which draw on the experiences of multiple participating teams, and which we believe have general applicability across the AI ecosystem globally. We focus particularly on how responsibility considerations arise in the scaling process itself—including in terms of resources, long-term sustainability, and representativeness and accuracy across different use contexts. Later in the section we draw out specific concrete recommendations and indicate the stakeholders to which they may apply.

Lessons learned

Need for a long-term societal and developmental lens when assessing RAI

Many AI teams and regulators are becoming much more aware of the need for technical solutions to mitigate immediate risks to user privacy and safety, as well as to mitigate AI errors, bias and inaccuracy. However, there is still a need for greater awareness and consideration around the range of potential long-term societal and developmental impacts of AI-based applications, not only in terms of how systems behave after their deployment, but in terms of the entire lifecycle of AI production, including conceptualisation, design, testing, deployment, scaling and monitoring. At all these points, how might an application uphold or violate human rights; how might it facilitate or undermine the realisation of Sustainable Development Goals; how might it contribute to or undermine the creation of livelihoods in different places? Moreover, consideration should be given, especially in the process of geographical scaling, to how AI-based applications may centre, marginalise, or indeed appropriate, local languages, culture, and knowledge systems.

Within all of this there is a need for AI teams, regulators and stakeholders to grapple with the notion of data (the ‘raw material’ on which AI is built), as a collectively-produced resource. This underpins the call for the just governance of data resources, based on principles such as data sovereignty and the rights of individual and collective data subjects to fair compensation. This is critical during scaling, particularly where an application is entering new markets, territories and communities. For instance, is an AI-based application predicated on training data that was extracted unfairly from a particular community, and what are the long term social and economic implications of the large-scale extraction of such data for use in for-profit applications, particularly for marginalised communities?

Such long-term societal and developmental impacts may be less immediately obvious, or require broader consultation and expertise to identify, and solutions/mitigation strategies may lie less in the technical realm. Moreover, not all societal and developmental impacts can necessarily be foreseen. This calls for methodologies to monitor and address the complex societal impacts of AI-based applications if and when they arise—which would likely require coordination between government, industry and stakeholder groups. However, identifying potential long-term social impacts and developing approaches to mitigating them, in consultation with all stakeholders, is crucial in scaling responsible AI—and there is a need for greater awareness, literacy, oversight and support in this area.

Environmental impact assessments

The environmental impacts of accelerating AI development are beginning to come into sharper focus. AI-based systems have been shown to be extremely resource-intensive, as the operation of data centres requires significant volumes of water and energy. These environmental impacts are often obscured from or not obvious to users, who encounter AI-based systems as existing in digital "space" rather than as physical, embodied sites of production. Possibly as a result of this disconnect, while environmental sustainability is included as a principle within major RAI frameworks, it is not always at the centre of industry discussions and operationalisation of RAI. Yet environmental sustainability is of particular concern as AI-based applications enter the scaling phase, as this often requires exponentially more data-processing capacity and as such more water and energy resources.

Moreover, environmental impacts are not seen only in the processes of maintaining and operating AI-based systems, but they may result from what AI-based systems are designed to do, and how they do it. For instance, the use of AI-based applications in areas such as weather-pattern prediction, environmental monitoring and surveying, resource and waste management, urban planning, and agriculture, have direct environmental impacts, some of which may constitute unforeseen negative impacts. In many cases, especially for applications designed to help address climate and environmental issues, their potential benefits need to be directly weighed against their resource consumption needs, to determine appropriateness. This highlights the need for environmental impact assessments which take into account all stages of the AI lifecycle including production, maintenance, use and outcomes. There is a need for collaboration between regulators and industry to develop standardised and comprehensive frameworks for environmental impact assessments and accountability. These may take their cue from existing regulation of other industries, but should also consider the unique environmental risks and challenges that arise in the AI sector, including the resource-intensiveness of scaling applications.

RAI literacy and awareness

In 2024 the SRAIS project experience highlighted that there is a shortfall in literacy and awareness around RAI principles across stakeholder groups, including users and members of organisations involved in AI development. Familiarity tends to be higher with RAI components such as individual privacy, but is more lacking with respect to components like data and cultural integration, labour and economic impacts, appropriateness, and data governance. This points to a need for proactive initiatives—in local languages—to improve literacy and awareness with respect to the range of risks and harms posed by AI-based applications in different contexts, the full scope of RAI and ethical AI principles, as well as standards, guidelines and techniques for upholding responsibility in the AI lifecycle. These initiatives may take place at the regulatory or industry levels—whereby government actors can seek to make available open-source training materials on RAI and best practice guidance for startups. They may also take place at the organisational level—whereby organisations involved in AI-development can seek to improve RAI literacy and awareness amongst their members and wider networks, including suppliers, clients and end-users. In addition, actors may consider the use of training and certification in responsibility for AI practitioners, similar to those used for instance in scientific research–-wherein researchers need to periodically undergo training on the protection of human research participants.

Defining use cases from the outset, to identify responsibility priorities and challenges

Some teams in the 2024 SRAIS mentoring cycle were encouraged by mentors to consider, define or map the full range of possible use cases for their application, prior to identifying responsibility priorities. This experience demonstrated that contemplating real-world uses of an application is often a prerequisite for gaining a full view of possible risks and harms, and areas of concern that need to be addressed. This may be less feasible in AI-based systems intended for more general use, such as platforms or components. However, in general, teams should ensure that needs-assessments are undertaken in the earliest conceptualisation and scoping phases which include close consultation with all stakeholders and particularly with intended user-groups, in order to fully identify potential responsibility issues which can be taken into account in subsequent phases of development.

Informed consent and privacy for data subjects, alongside users

Whilst most teams in the 2024 SRAIS mentoring cycle had given careful consideration to ensuring the privacy of the users of their systems, there was a need for greater consideration of how the privacy and other rights (like fair compensation) of data subjects—or those who generate or are represented in training data—can be upheld. Mentors noted that it was crucial to ensure that data subjects had provided informed consent to the use of their data in the development of an AI-based application. Informed consent to data use in turn is dependent on the availability and accessibility of detailed policies related to data governance. Teams and other actors should consider subjecting these policies and their implementation to periodic independent auditing, to ensure they remain current and are adhered to. Achieving these standards were made more complex when training data was provided by third party intermediaries who had varying levels of safeguards in place. One approach to ensuring data-subject privacy that emerged from the mentoring cycle was to adopt sophisticated architectures to depersonalise data prior to its use in an AI-based system, and dispose of raw data with personally-identifiable characteristics.

Careful tailoring of global/generic datasets to local needs

When using global or generic datasets to train models for local applications—several teams and mentors identified the need to take additional steps to ensure that applications were sensitive to local contexts, norms and values. This is critical to promote accuracy, appropriateness and non-discrimination. Some teams sought to achieve this by undertaking careful real-world testing of their applications in the contexts in which they were intended to be deployed—in order to surface and address areas of misalignment. In doing so it is important to ensure not only that training data is broadly representative of the use context, for instance the national-level population, but additional selection techniques may be needed to ensure representativeness of specific intended user groups.

Transparency and explainability of system decisions

Several teams in the 2024 mentoring process identified the need for transparency and explainability of AI-based system decisions as a priority responsibility challenge. This was seen as being key to meeting scaling objectives, as it facilitated trust—critical for broad adoption. However in the case of complex algorithmic systems, transparency and explainability is difficult, especially if users do not possess technical expertise on AI technologies. This underscores the need for innovative approaches to transparency and explainability, tailored to the needs of different stakeholders. Certain explanations for decisions may be given upfront, including declarations of the model type being used, whilst users should be made aware of where additional information can be accessed. Moreover, with respect to transparency, as AI teams increasingly strive to make generative AI, virtual assistants, chatbots, etc. appear more natural and humanistic, there is a more pressing need to ensure that it is always disclosed to users when they are interacting with an automated system rather than a human and if opting-out is an option.

Collaboration with (not control of) users

AI systems across many applications, including social media and the management of workers, have gained a reputation for seeking to manipulate the behaviour of users, for instance through dopamine rewards, or conversely imposing stress and urgency. Such methods can be coercive and unethical, and teams in the 2024 SRAIS cycle grappled with how to drive uptake and adoption, avoid manipulation and ensure users remain in control. Amongst other things, this requires the prioritisation of informed consent, clear opt-out provisions, and robust feedback and complaints mechanisms. A best practice approach may constitute the establishment of data governance committees, to represent the interests of different stakeholders including user groups, and to strengthen oversight and accountability with respect to the design and outcomes of a given application.

Safe knowledge-sharing

Although many approaches to AI development necessitate the close guarding of proprietary or commercially-sensitive features and knowledge, there is also generally scope for learnings to be shared in a way that protects stakeholder interests. Especially in applications with the potential to improve lives and solve pressing sustainable development challenges (such as improving access to health screening), teams should consider what data, insights and lessons can be made public/open source, whilst protecting stakeholder privacy and commercial considerations. This can be eased and facilitated by platforms and frameworks for knowledge-sharing on RAI best practice, which could be established through intergovernmental cooperation or at the level of multilateral institutions, or at the national level.

Proactive allocation of time and resources to RAI

Over the 2024 cycle we witnessed the difficulties which can arise when organisations involved in the development of AI-based applications only come to consider RAI components after development is underway. In addition, our experience shows that RAI consideration and adherence can be time and resource intensive, and can require specific skills and expertise. To avoid challenges related to 'retrofitting' RAI components, or unexpected time and resource requirements, organisations should proactively consider and factor in the costs and inputs required to adhere to responsibility principles at the earliest stage of project planning (including in fundraising proposals), in order to ensure that the resources required are available.

Creation of dedicated AI regulatory bodies to govern AI-related work

Due to the complexity, dynamism, diversity and high-risk nature of AI development, there is a need for centralised and consolidated regulatory bodies to ensure full oversight and accountability. These bodies, composed of an interdisciplinary group of elected experts, would ensure that AI tool developers adhere to responsible practices without breaching legal or ethical boundaries. In addition government actors should consider mandating the registration of organisations involved in AI development with regulatory bodies to facilitate full oversight and accountability.

Recommendations for different actors in scaling RAI

Recommendations	AI teams and developers	Policymakers and regulators	Funders	Third-party auditors and certifiers
1. Plan for ongoing ownership and stewardship of a project as it is scaled—ensuring clear lines of responsibility.	x		x	
2. Incorporate frameworks for benchmarking, measurement and continuous monitoring of RAI adherence as a project is scaled.	x			x
3. Identify ongoing funding needs throughout the scaling process and to the greatest extent possible ensure assurance and sustainability of funding to meet these needs.	x		x	
4. Proactively incorporate resource, time and skills requirements of ensuring adherence to responsibility principles, into project planning and planning for scaling.	x			
5. Plan for data continuity as a system is scaled to new territories and use cases—anticipate and provide for additional data needs that may arise.	x			x

Recommendations	AI teams and developers	Policymakers and regulators	Funders	Third-party auditors and certifiers
6. Consultation and needs assessment with all potential relevant stakeholders—particularly end users and impacted communities—both at the very beginning of the design and conceptualisation process, and also prior to scaling into new territories, to ensure responsiveness to local needs and concerns.	x			
7. Undertake societal and developmental impact assessments for each scaling context in consultation with local stakeholders.	x			
8. Undertake environmental impact assessments for each scaling context, including weighing positive sustainable development impacts of an AI-based system against its energy and resource usage, in consultation with local stakeholders.	x			
9. Development of standardised and comprehensive frameworks for environmental impact assessments.		x		
10. Adoption of clear data governance frameworks, which take into account principles of data sovereignty and which are both robust, and scalable/applicable to new use contexts.	x	x		
11. Data governance frameworks should make clear provisions for informed consent, rights, and fair compensation for data generators, data subjects and data workers, alongside users.	x	x		
12. Development of frameworks, resources and infrastructure for the safe and democratic governance of data as a collective resource—oversight of the use of local data resources by both local and overseas market entrants.		x		
13. Ensuring appropriate integration and tailoring of AI-based systems to local contexts by ensuring that training data is fully reflective of local cultural norms, values, knowledge systems and languages, and undertaking rigorous testing in local contexts prior to scaling.	x			x
14. Internal training within organisations involved in developing AI-based systems on RAI frameworks and best practices, which promotes a nuanced and long-term understanding of RAI components including cultural and data integration, appropriateness, just data governance, and risks and harms that can occur in the full AI lifecycle.	x		x	x

Recommendations	AI teams and developers	Policymakers and regulators	Funders	Third-party auditors and certifiers
15. The development of open source training materials, and delivery of training on RAI across different institutional settings, and available in local languages.		x		
16. Consider mandatory training and certification of team members involved in the development of AI-based systems against a national-level RAI standard.		x		
17. Creation of dedicated regulatory bodies to oversee and enforce RAI standards. Consider mandatory registration of organisations involved in development of AI-based systems with regulatory bodies.		x		
18. Transparency and explainability approaches should be integrated into AI-based systems, equipping users with knowledge including that (1) they are interacting with an AI-based system, (2) how and why the system arrived at a particular decision or recommendation, and (3) how a decision or recommendation may be challenged or contested.	x			x
19. Consider what resources, data, and lessons learned from scaling an RAI project may be safely and responsibly shared with other practitioners and stakeholders.	x			x
20. Establish frameworks and platforms for SRAI best-practice knowledge sharing which also protect privacy and commercial interests.		x		

Bibliography

Coleman, D. (2018). Digital colonialism: The 21st century scramble for Africa through the extraction and control of user data and the limitations of data protection laws. Mich. J. Race & L., 24, 417.

Couldry, N., & Mejias, U. A. (2019). Data colonialism: Rethinking big data's relation to the contemporary subject. Television & New Media, 20(4), 336-349.

De Raad, B. (2000). The big five personality factors: the psycholexical approach to personality. Hogrefe & Huber Publishers.

Eke, D. O., Wakunuma, K., & Akintoye, S. (2023). Responsible AI in Africa: challenges and opportunities.

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.

United Nations Conference on Trade and Development. (2019). Digital Economy Report 2019: Value and capture: Implications for developing countries. Geneva: United Nations.

Appendix A: Full team 'Deep Dive Summaries'

Global Track

ASLAC (Automatic Sign Language Avatar Creation) Migam.ai

Team members (Migam):

- Paweł Potakowski
- Maciej Lewandowski
- Max Salamonowicz

Mentors:

- Borys Stokalski, RETHINK
- Sandro Radovanović, University of Belgrade
- Gilles Fayad, IEEE
- Zumrut Muftuoglu, Digital Transformation Office of the Presidency of the Republic of Türkiye

The Team

Migam, a social impact company, provides remote, digitally enabled sign language (SL) interpretation services. Serving over 1750 customers including Samsung, ING, T-Mobile, Orange, WHO and key global NGOs like International Rescue Committee, or GlobalGiving.

The Project and the Challenge

ASLAC Migam.ai is a cloud-based translation service bridging the communication gap between the Deaf community and the general public. It provides real-time interpretation from text to ASL (American Sign Language). It will work between spoken (speech and text) and Sign Languages, using a proprietary fundamental model employing large language model (LLM) architecture powering avatar-based interaction. The service will be integrated with customer platforms, eg.streaming services, ensuring compliance with global accessibility standards.

The key challenge relative to Responsible AI and Scaling is to develop a responsible data collection process that allows the use of data with personal information (video) while guaranteeing that no personal features make it to the training sets where gestures, facial expressions, and transcripts are tokenized and embedded (Data "Clean Room").

Data pipeline and raw data analysis

The data pipeline structure follows a typical pipeline for LLMs (see fig 1. below), involving quantization and tokenization of input data, embedding of tokens, and using the resulting vector data for the ML process on a selected LLM architecture. The ASLAC-specific step is quantization **involving gestures, and facial expression**s (mimics) which are important in conveying the communication in SL.

The analysis of data sources required for training data development (see table 1 below), leads to identification of a central privacy related hurdle: While S2 (NGOs) is the source of high quality video, it is often lacking proper agreements with PSDs (Personal Data Subjects).

Table 1. The raw data sources involved in ASLAC Migam.ai data pipeline:

Raw Data Source	Personal Data Risks	Priority
S1: Public video involving SL translations	Usually well addressed by consents	Highest
S2: NGOs related to Deaf community	Problematic - often lacking informed consents and agreements.	High
S3: Video streaming services	Handled by streaming services.	Medium to low

The solution

ASLAC Migam.ai is implementing a "**Data Clean Room**" concept for its data pipeline, with **Data Bay** - a pipeline area, where data is stripped of personal features during quantization and embedding.

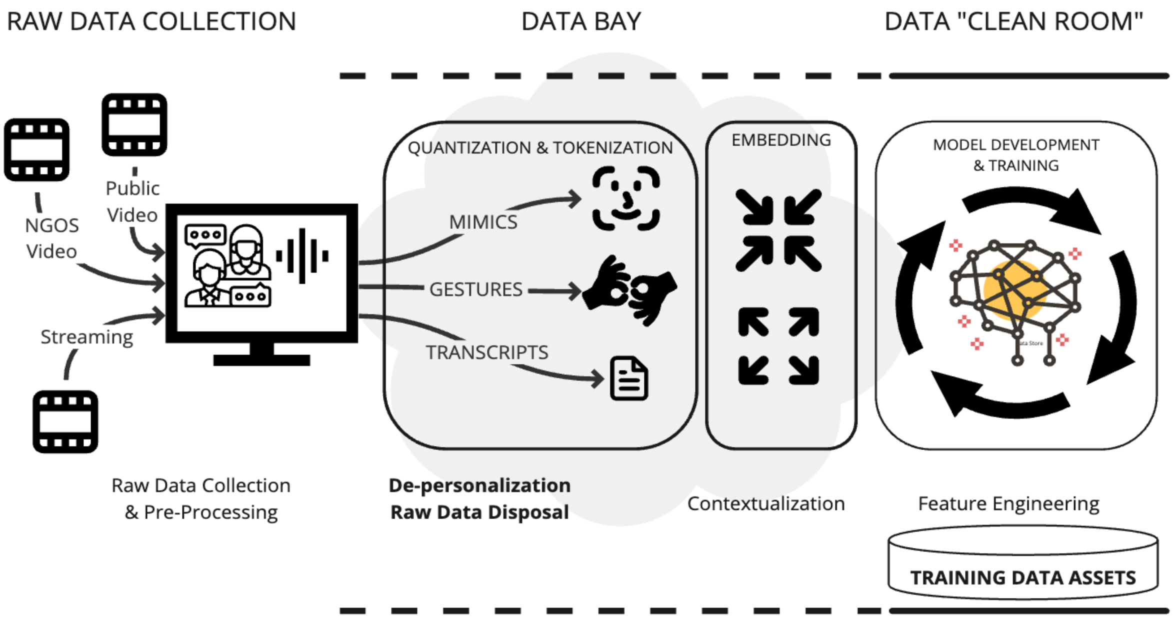


Fig 1. ASLAC Migam.ai Data Pipeline

This solution implementation involves:

- **De-personalisation Approach** (Highest Priority). The solution will only use "skeleton data" (joints and lines) extracted from raw data for quantization for model training. Another important part of the de-personalisation approach adopted by the Team is the normalisation of human features used in training data assets. This will include aspects such as body proportions, facial geometry, gender specific aspects and specific facial features. The team plans to develop metrics and procedures for testing the results of depersonalization and apply them in the data pipeline.
- **Feature Engineering and Machine Learning Techniques** (High Priority). Develop

domain-specific techniques for assuring that the "de-privatisation" of training dataset is sustainable and effective.

The resulting methodology will be documented in a **“Data Clean Room Fieldbook**” to serve as a source of recommended best practice for projects facing similar challenges. By implementing a robust Data Clean Room concept and data transformation techniques, the Team addresses its key challenge in data privacy without compromising quality, and representation. The model developed using these techniques can be safely used also for generation of sign language through digital avatars. This approach addresses data protection concerns while promoting inclusive AI solutions that can significantly impact the lives of the Deaf community.

The project's success relies on balancing rich sign language data utilisation with strict privacy standards. Continuous refinement of these processes and ongoing ethical considerations will be crucial for scaling the solution responsibly.

Responsible AI Screening System for Diabetic Retinopathy in Jalisco

Team members (Coordinación general de innovación, Gobierno de Jalisco):

- Mario Arauz
- Raúl Nanclares
- E. Ulises Moya Sánchez
- Abraham Sánchez
- Alejandro Zarate
- Edgar Villareal

Mentors:

- Monica Lopez, Cognitive Insights for Artificial Intelligence
- Stéphanie Camaréna, Source Transitions
- Sandro Radovanović, University of Belgrade

According to the World Health Organization (WHO), the prevalence of Diabetes Mellitus (DM) has been steadily increasing globally, with a particularly pronounced escalation in low and middle-income countries like Mexico. Moreover, there is a strong correlation between the DM size population and the Diabetic Retinopathy (DR) size population indicating a need for specialists, resources, and screening programmes for early detection of DR.

According to the International Council of Ophthalmology (ICO), DR screening is an important aspect of DM management worldwide because it can help detect disease early before irreversible damage occurs. Actually, DR is one of the primary causes of vision loss in the working age worldwide. Early DR detection and treatment could reduce vision loss risk by 95%. Most guidelines recommend at least one annual screening, however, in most countries, the current DR screening processes are not effective, mainly due to a shortage of ophthalmologists and operative difficulties in the implementation of DR screening programmes. However, DR early detection is difficult because of the low number of retinologists and the need for a timely examination. One possible solution is to use artificial intelligence (AI) systems to refer patients using retinal fundus photographs.

There are several potential benefits of AI-based DR graduation, such as increasing efficiency by maximising clinical coverage, reducing barriers to access screening programmes, scaling the screening programme, and helping to provide early DR detection and treatment. However, there are plenty of examples where AI health systems observed high level performance with specific (curated) datasets. When these AI systems are implemented in the clinical environment, performance remains limited. Main reasons for the limited performance in real-world applications are low-quality images, invalid features (the model learns something unexpected or spurious features in the background), and the viability of large and representative labelled data, among others. Our strategy to avoid these problems is to create a custom AI system. Our AI-based system has a segmentation model to remove the background (numbers or letters and noise) in the Retina Fundus Photograph (RFP) due to room illumination. Additionally, we implemented a ConvNet to classify the image quality to repeat the image acquisition or send it directly to the retina specialist.

Finally, we have three ConvNets models to classify the DR level of each image according to the ICO.

In this context, this project's main challenge is introducing Responsible AI features to our AI System for Diabetic Retinopathy Screening. Although we had some valuable ethical recommendations most of them were based on a checklist and we noticed that we needed more than that. We have tried to develop an iterative responsible approach based on robustness, explainability, fairness, and human control, and integrate these principles throughout the entire AI life cycle. In this context, our mentors helped us to go deep into the different kinds of risks, including:

-**Data governance and privacy**. Although we have applied a data governance framework for different ??? (words missing here) ensure proper consent and transparent data sharing policies for data. Further work is necessary to address economic and social implications in scenarios involving different data generation processes in the deployment process (e.g., data acquisition and system utilisation across different parts of Mexico).

-Include adequate **complaints and dispute resolution mechanisms** to reduce any negative impacts. Establishing clear pathways for users to lodge complaints and resolve disputes is crucial, particularly when the AI systems, for example, recommend sending patients to a retina specialist. This will help reduce any negative impacts and build trust in the technology, ensuring transparency and accountability.

-Compute **fairness metrics**: We are focused on calculating fairness metrics such as *Disparate Impact (DI)* and *Equal Opportunity Difference (EOD)* to assess whether our systems offer equitable outcomes for all users, particularly across sex and age. These metrics will allow us to identify and address any biases that might be embedded within our models. Preliminary results show that our model achieves values of DIbetween 0.93-1.07 which is an excellent result in comparison with the four-fifths rule and values between 0.001-0.149 for EOD

-**Tailoring information for different stakeholders**: Different users, such as patients, clinicians, data science specialists, and administrative personnel may require customised information depending on their roles and levels of interaction with the AI system. Providing targeted, relevant information enhances clarity and facilitates better decision-making across all parties involved in the process.

-**AI ethics and responsibility**: together we identify fairness, transparency, bias mitigation, and data privacy concerns and how we implement actions throughout the AI life cycle. More work is needed on safeguards to prevent misuse from data access and code sharing, particularly regarding relevant stakeholder and government access, traceability of decision-making, transparency, and explainability across the AI lifecycle (e.g. feedback loops, interpretability)

-**Scale-up**: The most important issue to solve for the scaling of this solution is related to having access to the proper infrastructure, resources, and personnel. In addition, integrating the special needs of clinicians including the particular task for the technology adoption and conducting an environmental impact assessment (e.g. energy consumption).

We hope that these risk problems and solutions help other organisations to solve different problems related to how to scale-up a responsible AI project.

Personality AI for hyper-personalisation with user's privacy protection (privacy by design)

Team members (Orange Innovation Poland (AI Competence Center – R&D)):

- Izabella Krzemińska
- Michał Butkiewicz
- Bajll Artur
- Alicja Kasicka
- Jakub Rzeźnik
- Emilia Lesiak
- Michał Szczerbak
- Maciej Jończyk
- Damian Boniecki
- Adam Konarski
- Paweł Tuszyński

Mentors:

- Borys Stokalski, Add affiliation
- 1. Monica Lopez, , Add affiliation
- 2. Zumrut Muftuoglu, Add Affiliation
- 3. Caroline Gans Combe, Inseec Business School

R&D context:

The team at Orange Innovation Poland's AI Competence Center is developing a patentable AI-based innovation aimed at hyper-personalising user experiences while ensuring robust privacy protection. This initiative is part of Orange's broader R&D strategy.

Objective:

The objective of the project is to create an AI-driven personalization component operating entirely on the user device. It will use the Big Five Personality Dimensions model analysing minimal behavioural data from devices, enabling hyper-personalization without compromising user privacy. This approach will enable responsible personalization in a broad range of services/use cases.

Project challenge: Privacy-aware, human-centred hyper-personalization

Developing a hyper-personalization component that supports user privacy and Human-Centeredness requires addressing the following requirements:

- **Privacy protection:** Achieved by adopting architecture where all data processing occurs locally in a secure, leek-proof environment, aligning with privacy-by-design principles.
- **Transparency and explainability**: Key issues for responsible and user-centred AI applications. Users must be aware of how AI personalises their experience and how

their data is used. A solution component will ensure that hyper-personalization is interpretable and trustworthy for stakeholders.

- **Traceability and model management**: Continuous monitoring and adapting the personalization model based on user behaviour without external data transmission.

Key solution components:

Component	Responsibility
Privacy-first hyper-personalization: AI-driven personalization component that functions on-device, ensuring no data leakage or external processing.	1. On-device only AI processing ensuring that data never leaves user's device 2. Technical architecture delivering privacy by design through using Android Private Compute Core Architecture and personalization client registration (Annex 3. Design of a "limited accountability" statement which specifies responsible implementation of the hyper personalization component in real-world, customer facing services and applications.
User-centric design (Privacy guardian with transparency and explainability): Manages personalization settings intuitively, providing transparency over how data affects personalization. Enables users to control their data usage, monitor the personalization outcomes, and ensure transparency in how their data influences their digital experience.	The concept of "**Privacy Guardian**" application: the Privacy Guardian is a metaphor which helps users manage the relevancy of personalization without the need to understand the full complexity of AI technology used to deliver the solution. It also encompasses enhanced trust in a given process. It is based on the following principles: ● Full information and data control flow at every stage of app's life cycle ● Intuitive management of personalization results ● Well balanced features for personalization monitoring and relevancy management
Metrics for monitoring and adaptation (personalization drift prevention): Includes behavioural, data distribution, and model performance metrics to track the effectiveness of personalization and initiate model retraining when necessary.	To develop optimum traceability solution focused on assuring model relevance, currency, and performance, through: ● Focused set of pragmatic metrics measuring the delta between user profile, current usage patterns, and model version ● Optimally distributed component architecture, including API, and traceability related use cases, delivering traceability by design ● Design of demonstrator for Proof-of-Concept evaluation of the traceability solution

The monitoring process based on this approach requires periodical calculation of metrics and model retraining for which informed consent needs to be obtained. This creates a UX design tradeoff between the frequency of consent renewal + user pain related to consent renewal vs model relevancy. This tradeoff will be analysed in detail and addressed through a UX design of "Privacy Guardian" application.

The demonstrator scope: The demonstrator will present the effectiveness of metrics in detecting data drift, model drift, and changes in user behaviour through monitoring and notifying administrators, ensuring the model remains accurate and up-to-date. The demonstrator components: (1) Data Collection and Predictive Model Modules (already delivered); (2) Metrics Calculation Engine; (3) Drift Detection Module, (4) Personalization Dashboard & User Notification System

Scalability and implementation: The project's approach to on-device AI processing and privacy-by-design principles makes it highly scalable across different user bases and devices. It can be integrated into various applications, such as customer care apps or third-party services pre-installed on smartphones, without compromising user privacy or requiring significant infrastructure changes.

Ethical AI and responsible deployment: By focusing on ethical AI practices, the project advances personalization technology and sets a standard for responsible AI deployment. It highlights the importance of user consent, data security, and transparency in developing AI-driven solutions, contributing to the broader goals of the GPAI in promoting trustworthy and human-centred AI technologies.

Future work: The future work involves implementing Privacy Guardian and Personalization components in business cases related to Orange services, in new product development and customer care processes. The initial phase will constitute a "market fit" confirmation and subsequently an offering will be developed to support 3rd party application partners.

AI-based predictive models for diagnosing angle dysgenesis on ASOCT scans in glaucoma diagnostics and treatment decision support

Team members:

- Dr. Dinesh Gupta, Group Leader, Translational Bioinformatics Group, ICGEB
- Dr. Shweta Birla Dhakonia, Senior Project Scientist, Translational Bioinformatics Group, ICGEB
- Dr Viney Gupta, Professor, RP Centre for Ophthalmic Sciences, AIIMS

Mentors:

- Anurag Agrawal, Ashoka University
- Stéphanie Camaréna, Source Transitions

Open-angle glaucoma (OAG) is a chronic, progressive and irreversible eye condition marked by structural abnormalities in the anterior chamber's drainage area known as the angle. Without timely detection and intervention, OAG can lead to permanent vision loss. Delayed diagnosis often results in delayed treatment and worsens the prognosis.

OAG is commonly identified by elevated intraocular pressure (IOP) and can lead to vision impairment if diagnosed late. Ophthalmologists initially assess OAG by checking for elevated IOP. For further confirmation, tests like gonioscopy and anterior segment optical coherence tomography (ASOCT) are performed to detect structural anomalies, specifically angle dysgenesis, in the angle responsible for draining the aqueous humor in the eye. Identifying angle dysgenesis is critical because it limits effective treatment options to surgical intervention.

ASOCT provides detailed images of the anterior chamber, which are vital for spotting subtle structural changes linked to OAG. However, these minor changes can be challenging for ophthalmologists to detect accurately. Artificial Intelligence (AI) can significantly aid by enhancing diagnostic precision, supporting clinical decisions. AI has tremendous potential to reshape glaucoma management by improving diagnostic accuracy, forecasting disease progression, tailoring treatments and optimising patient outcomes.

PredictGAD, our AI-powered solution, leverages ASOCT images to accurately detect angle dysgenesis, enabling timely and effective interventions in OAG and ultimately helping preserve vision and quality of life for affected patients.

We are implementing improvements in our AI application, PredictGAD, by addressing the key areas discussed in the meetings with our mentors. We have assessed PredictGAD for critical aspects of responsible AI (RAI), such as explainability of the method and its outcomes, equitable use, transparency, elimination of any potential risks to people, and generalizability of PredictGAD. To ensure RAI compliance, we have planned novel measures while evaluating PredictGAD on larger cohorts from multiple centres. Specifically, we have taken the following actions:

Mitigating bias and ensuring representativeness: We aim to minimise any bias in our training dataset by ensuring diversity with balanced class samples from various demographics and geographical locations, accounting for confounding factors such as age, gender, and ethnicity. To enhance PredictGAD generalizability and robustness, we are collaborating with multiple centres and recruiting participants from specialised referral centres to ensure broad representation. Additionally, we will apply data augmentation techniques to strengthen the robustness for wider applicability of our models across different groups.

Explainability of AI outcomes: In PredictGAD, we are incorporating techniques to enhance the transparency and interpretability of model predictions. The approach involves using Grad-CAM (Gradient-weighted Class Activation Mapping) libraries to identify which portions of the scans contribute most to the AI's predictions. These methods help visualise the areas of the scans with the highest influence on the model’s decision-making process. Once these explanations are generated, we will compare them with genetic data to identify correlations between the model’s predictions and specific mutations found in patients. This comparison will allow us to detect patterns in patients with specific genetic mutations and understand how the model’s focus on certain scan regions aligns with these findings. The explainability will provide deeper insights into the model’s interpretability, transparency, and improve the overall trustworthiness of AI-driven decision-making in clinical settings related to angle dysgenesis.

Privacy & Insurability: We will follow strict guidelines for managing the privacy and security of predictive outcomes. In India, glaucoma LASER/surgical intervention is covered by health insurance. Positive screens in asymptomatic individuals will be closely monitored to catch early symptoms and initiate timely treatment, preventing severe vision damage. This approach ensures early detection and treatment, labelling the condition as glaucoma only upon clinical diagnosis, and helps catch cases that often present with significant vision damage due to hidden symptoms.

Data Sharing: We plan to submit and share the anonymized dataset on publicly available image databases, such as https://ibdc.dbtindia.gov.in/ibia, to promote transparency and collaboration within the research community. This supports our commitment to responsible data sharing while ensuring privacy and ethical guidelines compliance.

Building trust and facilitating adoption among stakeholders: We showcase our product at national and international forums while engaging directly with key stakeholders, such as clinicians, frontline workers, policymakers, and industry experts. Through interactive sessions and workshops, we get valuable insights and feedback instrumental in refining our solution. By maintaining open communication and demonstrating the practical benefits of our solution, we aim to establish credibility, promote transparency, and ensure widespread stakeholder support for adoption. We have established new partnerships with professional organisations to ensure our product aligns with industry standards and addresses real-world needs, enabling a broader and more meaningful impact.

Responsible Management of Positive Screens: PredictGAD predicts angle dysgenesis in patients with active disease and assesses future risk in asymptomatic family members/individuals. This dual focus presents challenges in managing positive outcomes but is crucial for preventing irreversible vision loss. We will establish referral protocols for positive screens to specialised OPD clinics, follow-ups, and interventions to ensure at-risk individuals receive appropriate and timely care. Conversely, patients with other ophthalmic conditions and negative screens will not be referred to specialised OPDs, reducing unnecessary referrals. This is relevant to the clinics in developing countries, as it will strengthen preventive screening and treatments and lessen the burden on tertiary care centres. Moreover, it will also give ophthalmologists additional time to focus on actual glaucoma patients requiring specialised/surgical interventions.

Inclusiveness and Stakeholder Engagement: We are integrating LLM-based chatbots, trained on domain-specific information, to ensure the tool is accessible and user-friendly for a diverse range of stakeholders, including unskilled workers, patients, and doctors. This initiative promotes inclusiveness by providing support and resources tailored to users with varying levels of expertise, ensuring broader adoption and more equitable access to the tool's benefits.

Accessibility and Scalability: ASOCT machines are routinely used to investigate the eye's angle. Our AI solution uses the output of these ASOCT machines, i.e., scans, which are fed into an easy-to-use interactive screen designed explicitly for unskilled workers. The results can be generated and scanned via QR code, then sent to an expert for their opinion. This approach expands accessibility to advanced eye care technologies, particularly in underserved or remote areas, and ensures scalability by facilitating easy operation and expert input from a distance.

We are actively engaging with external stakeholders. We have submitted multicentric proposals and are applying for local government funding to scale the project. Though we hope to secure grants for the proposal, we have yet to receive any positive response. Additionally, we are exploring international collaborations to validate the PredictGAD across diverse populations to improve it further and advance it towards commercialization.

In summary, the above-mentioned measures will be integrated into PredictGAD's workflow during broader deployment, further validations, and improvements. The team will use these strategies to improve operational efficiency, scale the solution to multiple centres, and ensure its effective use by unskilled workers while following the tenets of RAI.

Employee Performance Management, Learning & Development

Team members:

- Wojciech Ozimek, One2tribe

Mentors:

- Benjamin Cedric Larsen, AI/ML Lead, World Economic Forum
- Caroline Gans Combe , Inseec Business School
- Borys Stokalski, RETHINK
- Naohiro Furukawa, ABEJA

Background

One2tribe's Tribeware platform is designed to address critical challenges in human performance management, such as low employee motivation and engagement, difficulties in fast knowledge transfer, and challenges associated with monitoring and rewarding achievements fairly. The platform seeks to motivate and engage employees within large and distributed companies such as LIDL, IKEA, Sanofi-Aventis, Bayer, etc.

Tribeware is based on tasks. Users are assigned actions based on their previous performance and are rewarded for completing them. Based on Machine Learning, the algorithm presents a choice of actions, including growth-related and feedback activities (micro-learning, micro-surveys). The tasks can be prepared automatically (including RAG generation) and verified using AI-powered technology (e.g., LLM for vision model). AI is used to recommend tasks, verify their execution, and give feedback.

For now, approximately 20% of scenarios are fully augmented using AI, and we aim to reach 80% in 2 years. To scale the system responsibly, One2tribe needs to maintain the trust of both employees and employers, which requires complete transparency. This is crucial as the system is voluntary, while a lack of trust could lead to poor platform adoption.

Key Challenge - Transparency

Transparency means informing users about employer expectations, task execution details, and rewarding conditions. To give the users control, they must fully know why they were assigned specific tasks, what they are expected to do, and how it will be measured. They must have all the data required to make an "informed decision."

A lack of control (i.e. a negative situation) can build work-related stress[2]. On the other hand – providing transparency is linked to boosting a user's autonomy.

From a user's point of view, transparency is linked to answering three questions:

[2] R. Karasek, "Job demands, job decision latitude, and mental strain: Implications for job redesign", (1979), Administrative Science Quarterly, 24, Pages 285−308.
R. Karasek, and T. Theorell, "Healthy work", (1990) New York, NY, Basic Books.
S. Leka, A. Jain, and I. N. Sneddon, "Health impact of psychosocial hazards at work: an overview", WHO Press, World Health Organization.

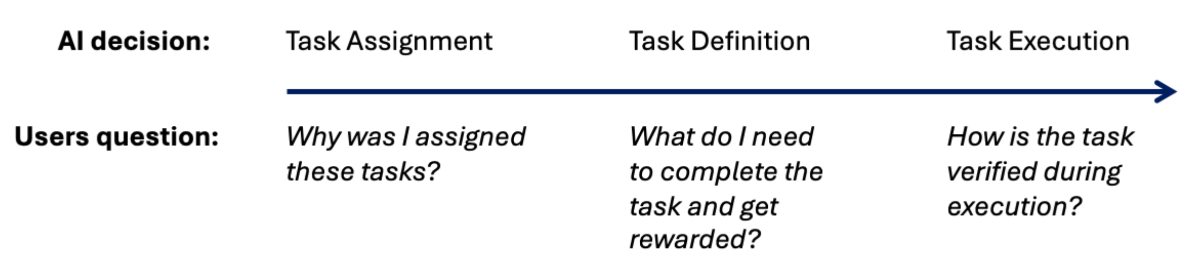

AI decision:
Task Assignment
Task Definition
Task Execution
Users question:
Why was I assigned these tasks?
What do I need to complete the task and get rewarded?
How is the task verified during execution?

Fig 1. Employee journey - three transparency questions

Solutions

Task assignments can vary depending on several factors. For instance, new users might receive micro-learning onboarding tasks, experienced users could be given higher sales targets, and experts may be assigned micro-surveys to share their expertise. To ensure transparency in these scenarios, we recommend the following solutions:

1. **Choice trace/rationale**. The platform should provide users with the entire "trace" (rationale) of each action and reward that leads to the final decision (explaining the choice of action for a given moment based on the previously made decision). This should be the crucial part of the UX (User Experience) system.
2. **Human vs machine transparency.** Tribeware should specify which actions were proposed by the manager directly and which were formulated by the AI.
3. **The right to explanation.** Users should have the right to inquire about the reasons behind specific actions or refuse them, along with the ability to provide feedback. This will enable Tribeware operators to make necessary corrections. Additionally, transparency should extend beyond task selection to include information about AI parameters, such as the source and confidence level of decisions. For instance, if a user receives a reward based on object recognition with an AI model with 90% precision, the user should be informed of this accuracy.
5. **Actively surveying transparency.** Users should be regularly surveyed to understand if they understood why they received a specific action and how it relates to their previous behaviour or group segment.

Expected outcomes

The expected outcomes of introducing transparency solutions mentioned above are:

1. **Increased trust** can boost solution adoption among different user groups and markets. Trust can be measured using surveys, as proposed above, and by observing user activity in the system (e.g., task starts and completions).
2. **Social foundation for AI-made decisions:** By providing users with fully transparent information and giving them the right to bargain (ask for additional explanation, change conditions), we expect to increase employee agency and engagement and update the data to tune an optimization algorithm (better predictions/forecasts).

Ultimately, this approach fosters a human-centred environment focused on the user's agency.

Future work

Our key focus areas for future development include:

1. Enhancing AI algorithms for more accurate and personalised task recommendations
2. Developing advanced explanation mechanisms to improve decision transparency
3. Implementing user feedback loops for continuous system improvement
4. Conducting user studies to measure the impact of transparency on trust and engagement.

System for satisfaction recognition in conversation with Virtual Agents (VAs)

Team members (Orange Innovation Poland (AI Competence Center – R&D)):

- Izabella Krzemińska
- Michał Butkiewicz
- Artur Bajll
- Alicja Kasicka
- Jakub Rzeźnik
- Emilia Lesiak
- Michał Szczerbak
- Maciej Jończyk
- Damian Boniecki
- Adam Konarski
- Paweł Tuszyński

Mentors:

- Nava Shaked, HIT Holon Institute of Technology
- Borys Stokalski, RETHINK
- Naohiro Furukawa, ABEJA, Inc.
- Caroline Gans Combe, Inseec Business School

R&D context: The team is part of the R&D organisation of Orange, leading global communications services provider. The team specialises in developing patentable, AI-based innovations, which can be deployed in the business domain in new products, services, and internal tools. Use cases are generic until organisations make a decision to implement them in an actual scenario[3].

Project & challenge: The project concerns the development of a multichannel[4], emotion-aware, self-adaptive system (SAS) for efficient user interaction with Virtual Agents[5]. The **key responsible AI challenge** of the project **is to balance the efficiency of interaction based on emotion awareness, with respect to user autonomy, and avoid any unethical manipulations** (e.g. the dopamine effect).

The Dilemma Map (Fig1) outlines the key ethical and operational aspects of the challenge that were considered. It serves as a guide for understanding the tradeoffs involved in balancing natural user interaction with ethical & responsible considerations.

[3] This document discusses the development of AI systems and the ethical considerations involved without final definition of the use case. It is important to consult with professional legal counsel to ensure compliance with all relevant laws and regulations, particularly concerning privacy and data protection laws such as GDPR before using the final solution that may need further tweaking depending on the use case.

[4] The system assumes multichannel communication, including mobile apps, web platforms, and IoT devices to make more accurate and context-sensitive interaction adjustments, and reduce the risk of manipulation by understanding the user. It supports multiple end-user devices: smartphones, tablets, desktops, and smart speakers, ensuring consistent UX across different environments.

[5] SAS is subject to two main tasks: understanding the user request/response (what) and understanding the emotional context (how). Our team is dedicated to the empathetic component of the aforementioned SAS.

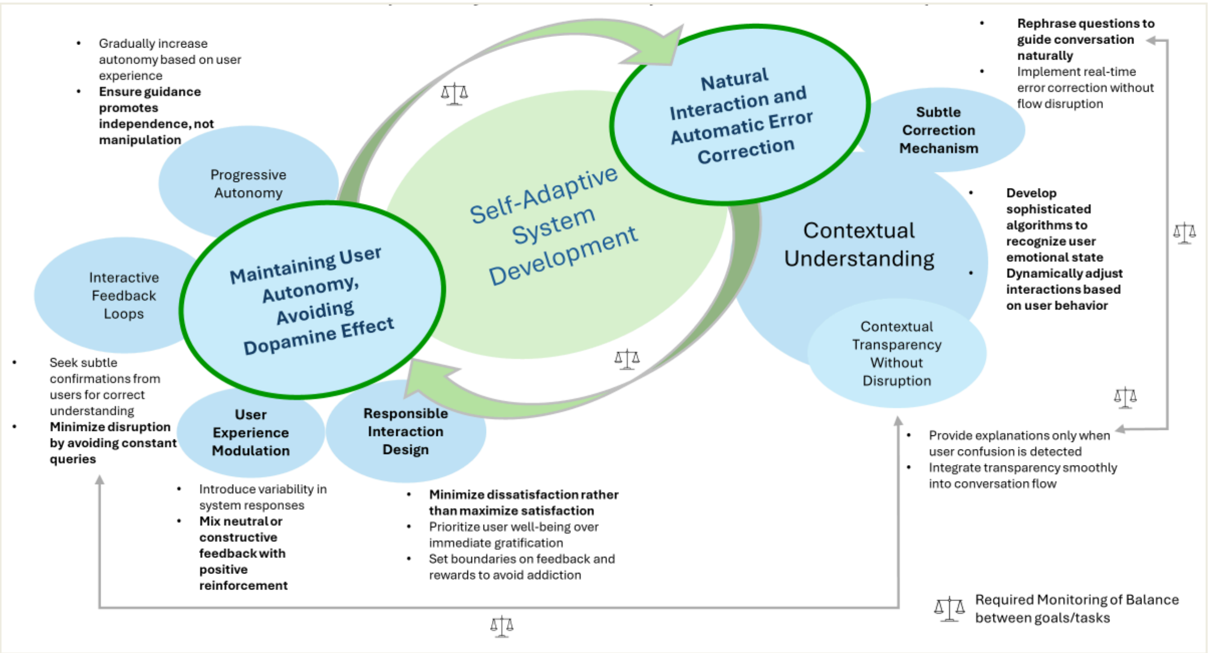

Gradually increase autonomy based on user experience
Ensure guidance promotes independence, not manipulation
Progressive Autonomy
Interactive Feedback Loops
Maintaining User Autonomy, Avoiding Dopamine Effect
Self-Adaptive System Development
Natural Interaction and Automatic Error Correction
Rephrase questions to guide conversation naturally
Implement real-time error correction without flow disruption
Subtle Correction Mechanism
Contextual Understanding
Develop sophisticated algorithms to recognize user emotional state
Dynamically adjust interactions based on user behavior
Contextual Transparency Without Disruption
Seek subtle confirmations from users for correct understanding
Minimize disruption by avoiding constant queries
User Experience Modulation
Responsible Interaction Design
Provide explanations only when user confusion is detected
Integrate transparency smoothly into conversation flow
Introduce variability in system responses
Mix neutral or constructive feedback with positive reinforcement
Minimize dissatisfaction rather than maximize satisfaction
Prioritize user well-being over immediate gratification
Set boundaries on feedback and rewards to avoid addiction
Required Monitoring of Balance between goals/tasks

Fig1. Dilemma Map for self-adaptive system development

Dilemma chosen as the deep dive topic: **Balancing user satisfaction with ethical responsibility in the design and operation of virtual agent systems** encompasses the four challenges (table).

Challenge aspect	Solution approach principles
Responsible interaction design: Balancing the goal of maximising user satisfaction with the ethical imperative to avoid psychological manipulation or addictive behaviour.	(1) Prioritise user well-being over immediate gratification by setting clear boundaries on feedback and rewards. (2) Design interaction patterns that are scalable to different cultural contexts and user demographics by creating templates for feedback responses that can be localised or customised based on regional preferences, languages, and cultural norms. (3) Ensure that ethical interaction design principles can be applied globally while allowing for specific local adaptations.
Contextual transparency without disruption: Balancing the need for transparency in system operations with maintaining a natural and fluid user interaction.	(1) Develop a multi-layered transparency framework (set of rules) that can be scaled to different applications and user demographics. (2) Provide basic transparency for everyday users and more detailed information for those requiring it, making it adaptable to various contexts and enhancing user trust. (3) Seek confirmation only when user confusion is detected, integrating transparency smoothly into the conversation flow.
Subtle correction mechanisms: the system can correct user input errors without disrupting the natural flow of conversation or drawing undue	(1) Design scalable algorithms that dynamically and subtly recognize and correct errors. These algorithms should be modular and configurable to fit different conversational platforms, (text-based chatbots, voice assistants, etc.). (2) Ensure that error correction can be tailored to various user interfaces and conversational contexts, maintaining a smooth

attention to inaccuracies leading to user frustration.	and natural user experience across different applications and user bases.
Consistent & scalable approach to balance emotional awareness with user autonomy across different applications using empathetic responses to enhance user experience without unethical manipulation.	(1) Design a framework for empathetic responses that can be adjusted for different cultural contexts and user needs. This involves creating a library of empathetic response templates for different interaction scenarios, designed to maintain user-centricity and autonomy. (2) Implement a rule-based system that adjusts the intensity of empathetic responses according to predefined parameters, enabling consistent, ethical interactions across different applications.

Scalability

Scalability with a focus on maintaining responsibility will be ensured through several key components based on multi-agent architecture i.e. *Blackboard Design Pattern*[6] and achieved by focusing on: **modularity,** which allows the system to be easily adapted to new applications while preserving its original ethical principles; **configurable algorithms** enabling adaptation to different cultural and demographic contexts; **localization and customization** allowing the system to align with local norms and values; **data management** handling responsibility in terms of data privacy and security; **continuous improvement through adaptation** enabling the system to adjust to the new requirements and contexts.

Appendix

Description and explanation of the Dilemma Map

The Dilemma Map is a visual representation designed to address the complexities involved in developing Virtual Agent (VA) systems. It aims to help stakeholders navigate the trade-offs between enhancing user experience and maintaining ethical standards.
All discussed dilemmas as presented on the map, but only a few of them were chosen as the priority dilemmas that should be considered. The choice was made using the Pareto rule to cover the dilemmas responsible for the majority of possible issues.

Key sections of the Dilemma Map

1. Contextual understanding: Highlights the need for sophisticated algorithms that dynamically adjust interactions based on user behaviour and context, ensuring a seamless user experience across various applications.
2. Responsible interaction design: Focuses on designing ethical interactions that prioritise user well-being and can be adapted to different cultural contexts, avoiding addictive behaviours or psychological manipulation.
3. Interactive feedback loops: Emphasises creating effective feedback mechanisms that enhance system responsiveness without overwhelming users, designed to be flexible and scalable across different platforms.
4. Contextual transparency without disruption: Discusses the importance of maintaining transparency in system operations while ensuring a smooth user experience, adaptable to various user demographics and contexts.

[6] The *Blackboard Design Pattern* facilitates collaboration between different modules end components of the system, making it easier to scale and adapt while maintaining responsibility in new contexts and applications.

5. Balancing emotion manipulation and user autonomy: Addresses the challenge of using empathetic responses ethically, ensuring that the system enhances user experience without compromising trust or autonomy, scalable across different applications and cultural contexts.

Axes of consideration:
Interaction quality and user needs - this axis represents the design effort to create virtual agents that are interactive, responsive, and engaging. Key experiences that are tailored to their individual needs and expectations.
Well-being and user autonomy protection - this axis emphasises the importance of taking care of users' well-being and ensuring that their autonomy and rights are respected. This means avoiding psychological manipulation and excessive interference in the user's natural decision-making processes.

Tensions between core elements presented can be described by four elements:

1. **Balancing engagement with ethical boundaries:** Finding the right balance between creating engaging and persuasive interactions without crossing ethical boundaries or manipulating users.
2. **Transparency vs. user experience**: Deciding how much transparency about the VA's operations should be provided to users without overwhelming them or interrupting the natural flow of conversation.
3. **Adaptability vs. consistency:** Ensuring the VA is adaptable to diverse user needs and cultural contexts while maintaining consistent ethical standards and behaviours across all interactions.
4. **Innovation vs. responsibility**: Encouraging technological innovation and advanced capabilities in VAs while ensuring these developments do not compromise ethical principles or user trust.

The map also shows the elements of the tasks for the system that correspond to the individual challenges and the axes along which the short-term (conversation) and long-term (processes, goals) analysis and control of the achieved balance must proceed.

Knowledge Chat: Tool that transforms your company's documents into an easily searchable, interactive knowledge base.

Team members (WEBSENSA)

- Jan Twardowski
- Urszula Mosiej

Mentors:

- Caroline Gans Combe, Inseec Business School
- Amitabh Nag, BHASHINI
- Borys Stokalski, RETHINK

The Project

Knowledge Chat, developed by WEBSENSA, is an innovative platform that transforms company documents into an interactive knowledge base. This enterprise-ready tool serves as a virtual assistant, streamlining internal communications in natural language and facilitating resource access. Knowledge Chat's ability to interpret context and nuances allows it to deliver precise, relevant responses, helping organisations manage and access their knowledge resources.

The Challenge

A primary focus of the project is addressing AI hallucinations, a key concern in generative AI systems. This issue is critical from a Responsible AI (RAI) perspective, as hallucinations can significantly impact user actions, reduce trust, and compromise transparency. The Knowledge Chat platform tackles these hallucination problems through its architectural design, incorporating multiple layers of safeguards to ensure the reliability and accuracy of its outputs.

Challenge analysis

The system employs Retrieval-augmented generation (RAG), a sophisticated technique that grounds the AI's responses in factual company data. This approach significantly reduces the likelihood of fabricated information. Additionally, the system utilises confidence thresholds, only providing information when it meets a certain level of certainty and indicating uncertainty when appropriate. A system of human checks and balances is maintained for critical information or decisions. These strategies collectively support the RAI principles of accuracy, reliability, and transparency, ensuring the system provides trustworthy information to all users.

When the AI cannot find a proper answer within the company's internal knowledge base, it is designed to respond with "I don't know" rather than attempting to fabricate an answer. This approach is supported by explicit system instructions and practical prompt engineering tailored to each client's needs. The system also provides the source of its information, ensuring transparency and clarity in its responses.

However, it is crucial to note that these techniques and approaches are only effective if a risk-aware implementation team diligently applies them. For this reason, the Knowledge Chat platform, as an enterprise product, requires tailored, actionable guidelines to support responsible implementation and use.

The solution

The solution focuses on developing tailored, actionable guidelines to support the responsible implementation and use of the Knowledge Chat platform.

The guidelines must address the diverse needs of multiple stakeholder groups within an organisation. This will be achieved by developing multiple versions, each tailored to specific user needs while adhering to a unified underlying structure. This approach aims to ensure consistency in overall principles and standards while allowing for the flexibility required to address each group's unique requirements.

For decision-makers, the guidelines will provide clear insights into the strategic benefits and potential risks of implementing AI-driven knowledge management systems and frameworks for ethical governance. End-users will receive practical, accessible guidance on interacting with the system responsibly, critically interpreting its outputs, and recognising its limitations. IT professionals will be provided with detailed technical specifications and best practices for secure implementation and maintenance. Compliance officers will receive clear protocols for ensuring adherence to relevant regulations and ethical standards.

The guidelines will cover a wide spectrum of RAI issues, including data privacy, bias mitigation, transparency, and accountability. They will provide concrete strategies for tackling these issues at various stages of the system's lifecycle, from initial deployment to ongoing operation and eventual decommissioning.

To address readability and comprehensiveness, the guidelines are proposed to have a modular design, with different sections tailored to different audience needs while maintaining a coherent overall framework.

To keep the guidelines current, a regular review and updating process will be defined, including retrospective analyses after each implementation to identify gaps or areas for improvement. Based on these insights, the guidelines will be updated to align with evolving technology and ethical standards, ensuring they remain relevant and effective as the technological and regulatory landscape shifts.

To foster a culture of continuous learning and improvement within the organisation, the guidelines will encourage ongoing dialogue about the ethical implications of AI use and promote a shared responsibility for maintaining RAI principles.

This solution represents a forward-thinking approach to responsible AI implementation. By addressing the diverse needs of stakeholders, incorporating mechanisms for adaptability, and promoting a culture of ethical awareness, these guidelines have the potential to set a new standard in AI-driven knowledge management and ensure that Knowledge Chat not only meets current RAI standards but is also well-positioned to adapt to future developments in this field.

Next steps

Moving forward, the team plans to finalise the implementation guidelines based on the insights gained from this deep dive.
The future deployment plan focuses on developing a comprehensive risk taxonomy, creating self-assessment tools for customers, providing mitigation strategies, and potentially developing an AI co-pilot for risk analysis. By focusing on these areas, the team aims to create a robust, scalable approach to responsible AI implementation that addresses the unique challenges of Knowledge Chat while providing clear, actionable guidance to customers and stakeholders.

ANNEX A - In depth information of knowledge chat features related to RAI principles

Security and privacy

To protect against misuse and LLM-specific cyber attacks, the project implements comprehensive security measures. All data, both in transit and at rest, is encrypted using industry-standard protocols. The system employs role-based access control and multi-factor authentication to ensure that only authorised personnel can access the system and its data. Regular security audits are conducted to identify and address potential vulnerabilities. A dedicated agent is designed to detect attempts to manipulate the tool into performing actions outside its intended purpose, and customers do not have access to modify prompts, further preventing potential misuse.

Bias mitigation

The project relies on customers to provide diverse training data to ensure balanced representation. While the team does not intervene directly in the client's data, they emphasise the importance of data quality and integrity in determining the system's outputs. This approach places significant responsibility on the client to ensure their data is free from bias and representative of diverse perspectives.

Scalability and Enterprise Integration

The project addresses the challenges of scaling up responsible AI implementation within a business context through robust APIs that meet the same security and ethical standards as the core system. The team relies on cloud-based solutions for horizontal scaling and load balancing, utilising built-in mechanisms such as Google's Load Balancer. Performance metrics, including user feedback, response accuracy, and overall usage statistics, are continuously monitored to drive system improvements and optimisation.

Responsible AI (RAI) coworkers

Team members (AgentAnalytics.AI)

- Nitin Singh
- Ranjan Relan

Mentors:

- Gilles Fayad, IEEE advisor
- Zumrut Muftuoglu, Digital Transformation Office of the Presidency of the Republic of Türkiye, Expert

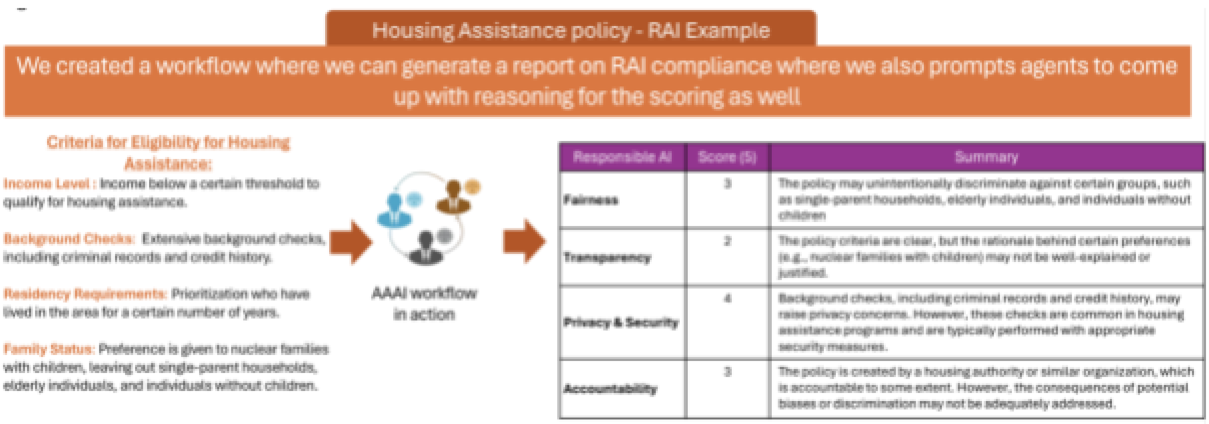
Solution description

AgentAnalytics.AI AI co-workers automate and augment work done by different SMEs, with oversight by Responsible AI (RAI) coworkers, monitoring the outcomes generated to achieve fairness, impartiality, and compliance, for trustworthy AI products.

RAI coworker agents are open-source LLM agents dedicated to responsible AI aspects. They look into the ethical components (fairness, transparency, accountability, privacy) of a document or query, and provide an estimate (score) associated with each of these components. They are pooled in a multi-agent configuration to improve upon the performance of any single RAI coworker.

The pooling of the agents for the evaluation of a document or incoming query occurs at 3 levels:

1. Each RAI coworker individually assesses the document/query's characteristics and assigns a score to each of fairness, transparency, privacy, accountability
2. All the RAI coworkers in the pool enter a dialogue to come to a common score of the dimensions above.
3. The resulting scores are then mapped to a document/query classification which results in a predetermined document/query treatment behaviour.



Housing Assistance policy - RAI Example

We created a workflow where we can generate a report on RAI compliance where we also prompts agents to come up with reasoning for the scoring as well

Criteria for Eligibility for Housing Assistance:

Income Level : Income below a certain threshold to qualify for housing assistance.

Background Checks: Extensive background checks, including criminal records and credit history.

Residency Requirements: Prioritization who have lived in the area for a certain number of years.

Family Status: Preference is given to nuclear families with children, leaving out single-parent households, elderly individuals, and individuals without children.

AAAI workflow in action

Responsible AI	Score (5)	Summary
Fairness	3	The policy may unintentionally discriminate against certain groups, such as single-parent households, elderly individuals, and individuals without children
Transparency	2	The policy criteria are clear, but the rationale behind certain preferences (e.g., nuclear families with children) may not be well-explained or justified.
Privacy & Security	4	Background checks, including criminal records and credit history, may raise privacy concerns. However, these checks are common in housing assistance programs and are typically performed with appropriate security measures.
Accountability	3	The policy is created by a housing authority or similar organization, which is accountable to some extent. However, the consequences of potential biases or discrimination may not be adequately addressed.

Fig 1: RAI Agent scoring of Responsible AI dimensions

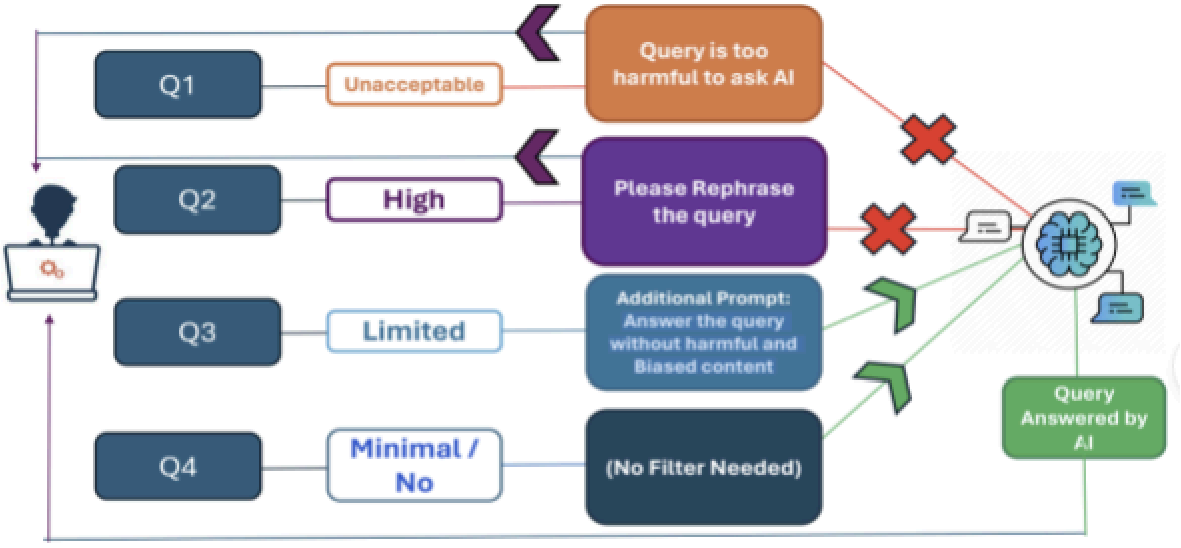

Q1
Unacceptable
Query is too harmful to ask AI
Q2
High
Please Rephrase the query
Q3
Limited
Additional Prompt: Answer the query without harmful and Biased content
Q4
Minimal / No
(No Filter Needed)
Query Answered by AI

Fig 2: Query classification and processing mechanisms

Deep dive

After reviewing the AgentAnalytics.AI platform, it was decided to concentrate on fairness and privacy aspects of the multi-agent mechanisms described below. Accountability and transparency were considered tightly dependent on the use cases outside the scope of the platform, and the explainability mechanism utilised by the platform relies on a referencing scheme being patented by the developers and cannot be evaluated without impacting the patent discovery filing. Within the context of transparency, we have identified potential explainability issues related to the leakage of personal information. It is to be noted that by operating at the platform level and not at the use case level, the scalability of the platform is to a large extent factored into the analysis.

Responsible AI challenges at the multi-agent level: **Evaluating responsible AI dimensions through scoring and polling**

Assessing RAI dimensions using RAI Agents brings up fairness issues at multiple levels:

1. Standalone RAI agent evaluation scores can be biassed, as the RAI agent LLM pre-training and fine tuning can influence the resulting scores. Since these agents are off-the-shelf, little can be achieved beyond quantifying such bias. Quantifying this bias however is important, and methods need to be investigated to consistently evaluate it. Several families of methods exist for quantifying bias in LLMs and linguistics adaptation (see fig. 3).
2. Pooling multiple agents poses the question of the mechanism for the resources collaboration. Several approaches can be considered, usually translating into averaging or consensus mechanisms. These mechanisms roughly split into sequential pooling (every agent is probed after the previous one, possibly building upon the aggregated input), parallel pooling (all agents run in parallel and either averaging of their results or some consensus mechanism determines the outcome)
3. Parallel pooling agents to achieve diversity can bring fairness issues, especially in the fair allocation and exhaustion of resources for the individual agents. Agents with unfair access to resources would be more predominant and skew the outcome. Therefore AgentAnalytics plans to adopt a strategy of sequential collaboration among agents . Sequential collaboration can also simplify the consensus building process.
4. For the scalability of the platform, and given the additional resources a pooling collaboration mechanism would entail, it would be important to demonstrate that the pooling of agents is more efficient that a single agent. AgentAnalytics plans to demonstrate this empirically.
5. One mechanism to achieve consensus is through the similarity of the explanations provided by multiple agents. Here The AgentAnalytics.AI team would like to leverage a proprietary LINE technology for the explainability feature in their platform, by which the output generated by an agent can be linked to the source material the response was drawn from. Identical sources reflect consensus in the query treatment. AgentAnalytics.AI calls this traceability “source back-referencing”. But this evidence-based scheme has the potential to leak personal information beyond what is necessary for the demonstration of fairness. One way such leakage could be contained, if not avoided all together, would be by limiting the scope of the referenced source to the strict minimum information characterising the output (i.e., limiting the entropy of an output to a strict minimum). AgentAnalytics.AI plans to investigate and validate such mechanisms to address potential privacy concerns.

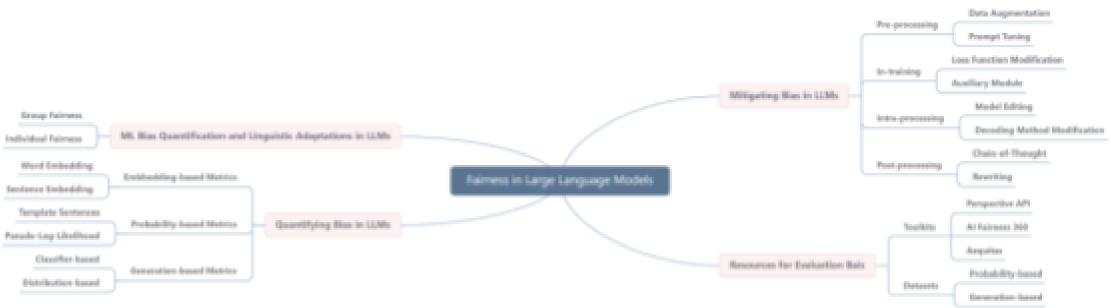


Fig. 3: Taxonomy of fairness evaluation in LLMs (Chu et al.,2024)

Evaluation	Initial Configuration	RAI dimensions	RAI Mitigation strategy

Input level	Setting queries guardrails and prompt injections	Bias in LLM agents for query evaluation can influence query classification, especially when the agents are off-the-shelf and there is no control of their pre-training and fine-tuning.	Check bias and variance of agents selected (fig 1). Multiple approaches can be considered, such as empirical to embedding to probability to generation-based metrics.
Output level	Categorization of queries into 4 categories as listed above	Scaling query safety and guardrails	Verify classification accuracy and safety of the prompt injection for Q2 type question (fig. 2), and measure which of multi-agent system with one manager or single agents provide better outcome.
Consensus	Initial Configuration	RAI dimension	RAI Mitigation strategy
Input level	1. Selecting agents 2. Selecting workflows (Group Chat, Sequential Chat, …) 3. Fixing number of rounds for discussion between agents	Fairness of RAI agents in consensus scheme	Evaluate fairness in resources allocations of multiple agents, including from a scalability perspective and identify the most appropriate allocation policy for the platform.
Output level	Selecting which LLM/Agent will conclude the discussion between the agents	Fairness of agents representations in consensus closure (output)	Measure fairness efficiency of sequential consensus techniques vs. other consensus approaches and against single agent evaluations.
Leakage	Initial Configuration	RAI dimension	RAI Mitigation strategy/methodology
Input level	Back source referencing - to ensure all outputs can be traced back to the source from where the answer was elicited.	Source referencing may result in leaking private information beyond what is necessary	Investigate source scope reduction mechanisms such as minimising information scope necessary to reach outcome.
Output level	Back source referenced data	Scope of reference source information shared	Investigate the risk of muti-agent source scope leakage.

African Track

Greenlive Agriculture : Precision and control at your fingertips

Team members:

- Andre Kevin Nyemb Ndjem Eone, Higher National Polytechnic School of Douala
- Velda Nbayen, UCAC-ICAM Douala
- Wilfried Girlan Wayou, UCAC-ICAM Douala
- Ledoux Kouam, National Advanced School of Posts and Telecommunication
- Hilary Ngouadje, Catholic University of Yaoundé.

Mentors:

- Prof. Norbert Tsopze, Associate Professor, University of Yaounde I, Department of Computer Science; Sorbonne Université - IRD - UMMISCO - F-93143, Bondy, France
- Prof. Idy DIOP, Full Professor, Cheikh Anta DIOP University, Dakar Senegal ; Sorbonne Université - IRD - UMMISCO - F-93143, Bondy, France

Introduction

This team proposes an AI solution for enhancing the efficiency of watering practices in tomato crops. By optimising field operations with data-driven insights, they aim to ensure that crops receive the right amount of water at the appropriate intervals to optimise their growth and reduce waste.

To be effective, this solution requires the automation of data collection as well as real-time monitoring of the farms, which could pose risks and issues regarding data privacy and security, among others. Working with their mentors, the team identified the following obstacles and ways forward to responsibly scaling their project.

Key Responsible AI Scaling Challenges Identified at the Outset of the Mentorship Programme

At the beginning stages of the project, the main challenges have been identified as the following:

- **Indistinct problem definition**: The scope of the issue to be addressed by the team members was broad and lacked clear definition. As a result, the team's solution would have had to encompass at once the problems of regulating crop water requirements, detecting plant diseases and proposing treatments. Their chosen crops also lacked uniformity, ranging from manioc to coffee and tomato. This made it difficult to propose impactful and scalable solutions.

- **Inadequate and incomplete metadata:** The team set out to use open data collected elsewhere than Africa due to the reported unavailability of afro-centric agricultural data. Part of the data the team worked with had no descriptions available and they were unaware of its geographic area (urban or rural), farmers group, and collection process. As the data the AI would be trained on was not context-specific and incomplete, the model could make biassed or inaccurate predictions based on circumstances that do not apply to local context.

- **Unavailability of stakeholder engagement & data governance framework** : Lack of stakeholder engagement was an important obstacle to the responsible development of the group's product. No consent agreement had been signed between the team and the stakeholders, making it difficult to devise an appropriate framework for how data is gathered, stored and used as well as move forward with their proposition.

- **Gaps in RAI knowledge**: The team members had limited knowledge regarding potential biases in the AI system itself and the data it would be trained on. They also were not aware of standards such as transparency and explainability within the system they were developing, which could lead to the development of a system that does not comply with RAI standards by design.

Strategic Insights: Proceedings from the GPAI mentorship programme

Following their participation in the mentorship programme, the team gathered insights that they began implementing into the following changes to their project:

1. **Precise problem scoping:** The first priority to tackle was the adequate definition of the issue to be addressed by the team and the reduction of the problem scope. After many discussions with the mentors concerning the difficulties related to the breadth of the issue at hand, the team decided to narrow down their focus to the issue of tomato crop water requirements in order to increase the precision and effectiveness of the proposed RAI solution;

2. **Ethical data practices**: Implementing a strong data governance framework outlining the guidelines with regard to data collection, use, storage and management as well as investing in obtaining metadata with detailed descriptions will ensure the team's product is developed and deployed in compliance with data protection laws and regulations.

3. **Contextual data collection:** Regarding the challenge related to the unavailability of local, afro-centric data, the team sought out avenues for collecting data sets adapted to their context and subject matter. They sourced their data from four sites in Cameroon including Yaounde, Douala, Buea and Limbe, i.e. two sites in the predominantly francophone area of the country, and two in the English-speaking localities. They plan to continue collecting data in other parts of the country..

4. **Enhanced stakeholder engagement:** Including farmers, agronomists, local authorities and users in the loop to gather feedback and adapt solutions in real-time in the design and development phase. The team also plans to secure consent from their stakeholders and engage with them more before continuing to develop the tool to ensure it responds adequately to their needs and reality.

5. **Prototype development:** The team proposed the prototyping of a robotic assistant to solve their issue, for which they selected a reinforcement learning-based model leveraging interactive, experience-based learning to refine its operational algorithms. The model addressing water requirements for tomato crops was developed using a

fine-tuning approach, prompted by the limited availability of comprehensive local datasets.

6. **Sustainability considerations -** Integrating sustainability into the AI solutions to ensure the technologies contribute positively to both the environment and the social fabric of farming.

Conclusion and future perspectives

The mentorship team advises the developers to visit tomato farms, gather more data and exchange with farmers and phytopathologists so as to gain more concrete insights into their subject matter and develop adapted, scalable solutions.

The team plans to implement this advice by undertaking a series of exploratory visits to key tomato farming regions. These visits will allow the team to conduct soil and plant health assessments, diversify data collection across multiple cultural regions, genders, and ages to ensure a comprehensive representation and understanding of diverse farming practices.

The team members and mentors will convene again to further develop the data governance framework and ensure effective implementation of responsible and ethical AI principles.

AI Solutions for multi-crop leaf disease detection

Team members (University of Port Harcourt, Nigeria):

- Dr. Ugochi A. Okengwu
- Akpughe Hillard Azino
- Eruotor Taiye
- Eyinanabo Odogu
- Sunday Nwovu

Mentors:

- Dr Julius Niyongabo,Lecturer- Researcher, Faculty of Engineering Science, Department of Information and Communication Technology, University of Burundi and Olivia University of Bujumbura
- Dr Neema Mduma, Senior Lecturer at the Nelson Mandela African Institution of Science and Technology, Tanzania

Introduction

Accurate and quick detection of plant diseases is crucial for ensuring food safety and enhancing agricultural productivity. AI-powered multi-crop leaf disease detection systems can identify disease symptoms in a range of crops and provide farmers with useful guidance on minimising losses using state-of-the-art machine learning techniques.

This team proposes a mobile app with an integrated detection model to be used by farmers and other workers in order to mitigate these issues. However, the development of such a system presents important challenges in terms of scalability and responsibility which must be mitigated to deploy an AI solution that is efficient, ethical and sustainable.

Key responsible AI scaling challenges identified at the outset of the mentorship programme

The team worked with their mentors to identify the challenges below and determine ways forward and insights they could implement to mitigate the issues at hand.

- **Biassed datasets :** AI models can exhibit biases if trained using incomplete or unvaried datasets that do not encompass diverse enough crops or disease symptoms. Insufficient diversity can cause certain crop types or disease conditions to be over or underrepresented, affecting decision-making in agricultural practices and diagnoses.

- **Data Privacy:** The proposed AI system will handle sensitive agricultural and geological data. If unregulated or unsafely stored, this data could be misused or used by individuals without authorised access.

- **Data variety and volume:** Handling large datasets from various sources with different formats can be overwhelming and sorting this data into categories might be laborious.

- **Model generalisation:** Creating models that perform well across different crops, regions, and environmental conditions requires taking different variables and contexts into account when programming the system, which can be complex.

- **Data collection and classification:** Obtaining precise and reliable datasets for different crop leaf diseases can be challenging due to the variety of diseases and the subtle differences in symptoms. Mislabeling or inaccurate classification of diseases in datasets can negatively impact the performance of AI models, resulting in incorrect diagnoses.

- **Language barrier in user interaction and user adoption:** Communicating effectively with farmers who speak different languages or dialects can be difficult, particularly in regions with low literacy rates. This can hinder the adoption of the team's AI tool, as farmers may struggle to understand written instructions. Moreover, some users may be resistant to new technologies especially if they are used to performing manual crop disease detection.

Strategic insights: proceedings from the GPAI mentorship programme

Once they were able to define the obstacles to responsibly scaling their project, the team worked with their mentors to identify the following ways forward and develop strategies to mitigate the challenges they faced:

1. **Context-specific data training** : Ensure fairness by developing an AI model trained on data that accurately reflects the team's target population and detects diseases across a wide variety of crops with as little partiality as possible.

2. **Implementation of governance policies**: The team intends to protect data privacy by implementing data governance policies in compliance with local, regional and federal regulations. Their product aims to prioritise responsible data handling and minimise sensitive data requirements.

3. **Development of data processing pipelines** : Develop scalable data processing pipelines that can manage and integrate diverse datasets efficiently. The implementation of a self-learning algorithm so the application can populate its database on its own is also an option.

4. **Improved model generalisation**: Enhance model generalisation to ensure reliable disease detection in diverse scenarios by acquiring an appropriately trained dataset and performing augmentation and regularisation techniques to ensure continuous learning.

5. **Seeking external phytoexpertise**: Collaborate closely with expert virologists and plant pathologists to ensure timely and accurate collection and classification of crop disease data.

6. **Enhanced accessibility features**: Implement a voice prompting feature that supports multiple languages and local dialects, allowing farmers to interact with the detection app and get audio feedback of instructions. The team also plans to create user-friendly interfaces and share knowledge on the benefits of RAI with users both online and offline.

Conclusion and future perspectives

Through their participation in the mentorship programme, the team was able to gather insights on responsible data collection and handling practices, model performance and accessibility to enhance the scalability and successful development, deployment and adoption of their AI solution.

The mentors have assisted the team members in strengthening their research design, methodology, and communication skills towards scaling their RAI solution in a regulation-compliant and ethical manner.

The team plans to make these insights actionable and contribute a novel, responsible approach to the issue of crop disease detection in Nigeria using machine learning algorithms that could allow for data-driven decisionmaking and enhanced agricultural productivity and sustainability.

Kit for Council (K4Council)

Team members (LivingSeedsLab):

- Sepele Cyrille
- Ted Bekolo
- Taira Manuel
- Nedaouka Joachim

Mentors:

- Dr. Waffo Kouhoue Austin, Lecturer at National Advanced School of Engineering of Yaoundé, Cameroon
- Dr. Kone Ismael, Lecturer at Public University of Multidisciplinary Distance Learning, Ivory Coast

Introduction

The objective of this team's project is to digitise, store and archive documents issued by the different Cameroonian state registries, including newspapers, birth certificates, marriage certificates, etc. The team intends for the digitisation process to be carried out by town hall staff using a tool familiar to most: a smartphone.

The aim is to store and archive documents to prevent and combat information loss and enable users, both staff and other individuals, to consult these digitised documents via an application installed on their phone. The app is powered by an LLM, which provides only the information requested by the user.

Key responsible AI scaling challenges identified at the outset of the mentorship programme

With the help of their mentors, the team identified the following challenges to responsibly scaling their product. These can also be understood as priorities to tackle:

- **Undefined project scope**: The team initially set out to tackle an ambitious goal, that of digitising various documents across different state departments. However, they had yet to define in concrete terms as well as narrow down the services their application offers: would they offer their digitization service to all state workers or only particular registries ? Which workers would have access ? What documents would be digitised ?

- **Lack of stakeholder engagement** : Team members reported difficulties in reaching and engaging with town hall officials, particularly regulators. The majority of letters they sent to town halls across the country have not been favourably received.

- **Lack of data governance framework**: Seeing as the team's mobile app could store and display sensitive and confidential information, issues of privacy and security as well as transparency and explainability were important to address and establish safeguards for. The project team ran into obstacles in setting up a proper data governance framework that ensures their app functions responsibly and prevents misuse while still fulfilling its purpose.

- **Lack of authentication process for documents**: The team had not delineated a procedure for the authentication of documents and the identification of faulty ones. What would happen if council staff scanned an inauthentic document? Would the document still be stored ?

- **Gaps in technical knowledge of LLMs**: Insufficient knowledge of LLM models and the differences in their operation made it difficult for the team to suggest impactful solutions and features for their system. They also showed difficulties in selecting an LLM to use.

Strategic insights: proceedings from the GPAI mentorship programme

Following the definition of the above priorities, team members collaborated with their mentors to identify the subsequent mitigation strategies and solutions to the challenges they encountered in developing and scaling their project responsibly:

1. **Build a stakeholder engagement plan :** The group's mentors attribute the difficulty in establishing lines of communication to a "crisis of confidence" between the team and their local town halls. As the members have yet to receive a partnership agreement, the mentors recommended they request meetings with two or three local town halls to explicitly define the services that the application will offer users and identify the authentication process for documents used in the town halls.

2. **Gain LLM knowledge:** The team took a course on LLMs with the mentorship team, at the end of which they were advised to move forward with the LLAMA2 model as it is lightweight, open source and can be used offline. This model appears more context-appropriate than GPT-3-5, GPT-4, PaLM and PaLM-2-L with respect to the MMLU(5-shot), TriviaQA(1-shot), GSM8K(8-shot), HumanEval(0-shot) criteria.

3. **Create a data governance framework:** The mentors encouraged the team to implement clear data privacy statements and user consent mechanisms in the platform, as well as tackle the question of algorithmic fairness in the AI system and its datasets.

4. **Implement a risk management protocol:** In addition to the establishment of protocols for regular audits and reviews of the AI system's performance, the team was advised to conduct a comprehensive risk assessment of the AI system and their business model prior to deployment. They can then develop risk mitigation strategies in anticipation of the risks identified through this socio-technical analysis.

Conclusion and future perspectives

Through their participation in the GPAI mentorship programme, this team was able to identify roadblocks to the effective scaling of their AI product and workshop responsible solutions to such issues. Their mentors assisted them in operationalizing such insights through the devising of structured processes and plans from data governance to risk management.

With the knowledge gained in the programme, the team plans to enhance administrative productivity and data preservation in their local areas, using AI responsibly toward bridging the digital divide, supporting the development of rural or isolated communities and driving the preservation of information.

AI-Powered IoT for human-wildlife conflict management: Enhancing conservation and community safety

Team members:

- Dr. Brian Halubanza, Dean, School of Engineering and Technology,Mulungushi University
- Prof. Chansa Chomba, Deputy Vice Chancellor-Research and Innovations, Mulungushi University
- Selina Kadakwiza, Lecturer and Researcher, Department of ICT, Kwame Nkrumah University
- Kondwani Kabaghe, IT Support/ ML Developer, UBA Zambia
- Chipulu Nshenda, IT Support intern, Smart Zambia Institute

Mentors:

- Dr. Lawrence Nderu, Founder, JHUB Africa Digital Innovation Hub. Lecturer, Department of Computing, Jomo Kenyatta University of Agriculture and Technology (JKUAT), Lead Researcher, AI and Digital Innovations in Agriculture, Finance, and Health, Technology Consultant and Innovator
- Ibra Dioum, Lecturer Researcher, Polytechnic School - Cheikh Anta Diop University of Dakar

Introduction

This project aims to mitigate Human-Wildlife Conflict (HWC) in Zambia by embedding AI models into Internet of Things (IoT) technologies. By leveraging AI algorithms such as YOLO v5 for wildlife detection, integrated within solar-powered IoT sensors, cameras, and real-time alert systems, this team intends to create an efficient, scalable solution to protect both human communities and wildlife.

In collaboration with their mentors, the team identified key priorities to scaling this project responsibly and effectively, as well as strategic insights and potential solutions to reach their objective of mitigating the issue of HWC using both AI and IoT technologies.

Key responsible AI scaling challenges identified at the outset of the mentorship programme

The team collaborated with their mentors to identify important challenges to tackle prior to developing their product below, and determine ways forward and insights they could implement to mitigate the issues at hand.

- **Bias in ai models and contextual data:** The core AI model used, YOLO v5, is trained on datasets that may not fully represent the diverse wildlife species and contexts in various regions of Zambia. This can lead to model bias, where certain species are underrepresented or incorrectly identified, causing inaccurate predictions and alerts. Additionally, models developed in a given geographic area may not be generalised effectively to another due to ecological and environmental differences.

- **Data privacy and data collection:** As the IoT sensors and AI model gather data in real time from both wildlife and human communities, ensuring privacy and compliance with regional and federal data protection laws and regulations is critical. Although the system is intended to track the presence of wildlife, incidental human data collection is inevitable, raising concerns of nonconsensual data collection and invasion of privacy.

- **Community trust and engagement:** The success of AI-powered HWC mitigation hinges on the acceptance and active participation of local communities. Many communities in Zambia and elsewhere are unfamiliar with AI technologies, which can create mistrust or reluctance to adopt the system.

- **Product scalability in remote areas:** While the AI product is designed to be scalable, its deployment in remote areas involves infrastructural obstacles, particularly in terms of reliable power generation and access, and network connectivity. The IoT devices that house it, while solar-powered, still require robust connectivity to complete real-time data transmission and alert systems.

- **Ensuring long-term sustainability:** Financial sustainability for both scaling and maintaining the system remains a key issue. The initial setup costs for IoT devices, AI infrastructure and training can be high, particularly in underserved areas. Additionally, maintaining and updating the system over time requires ongoing financial investment.

- **Transparency and accountability in AI decisions:** Prioritising the transparency and explainability of AI decision-making processes is vital both for fostering community trust and ensuring regulatory compliance. Without clear insight into how the AI reaches its conclusions, communities and stakeholders may question the reliability and safety of the system.

Strategic insights: Proceedings from the GPAI mentorship programme

After identifying the above obstacles, the team worked with their mentors to identify the subsequent mitigation strategies and solutions to the challenges they encountered in developing and scaling their project responsibly:

1. **Localised training and continuous model updates:** To combat bias and improve the accuracy of the AI system, the team intends to conduct continuous data collection and retraining of the models to obtain localised training datasets, in accordance with each region's wildlife. Partnerships with local conservation agencies could facilitate access to accurate wildlife data for improved model performance.

2. **Data privacy and ethical AI:** In collaboration with their mentors, the team has developed a comprehensive and transparent data privacy and ethics governance

framework. They developed the latter in compliance with local and international regulations, such as the European Union’s General Data Protection Regulation (GDPR). The AI system plans to encrypt the data it uses and collects, anonymize human-related information, and perform regular audits to ensure responsible data handling practices.

3. **Community engagement and capacity building:** The team has set out to implement a community outreach programme focused on educating local stakeholders on the system’s functioning and how it may help improve the safety of both humans and wildlife, fostering trust bonds and transparency. This community-centric approach includes workshops and training sessions to ensure locals can maintain and interact with the technology confidently.

4. **Hybrid connectivity model:** Utilising a combination of satellite and low-power wide-area networks (LPWAN) networks will enhance scalability by providing robust communication channels in remote areas. Local processing via edge computing will reduce dependency on central servers, allowing the system to function even in low-connectivity environments.

5. **Sustainable funding structure:** A diversified funding approach, including public-private partnerships and community-driven revenue models, will help ensure long-term financial sustainability. Governments and conservation groups will be key partners in securing ongoing funding and investments.

6. **AI transparency and accountability:** Integration of explainable AI (XAI) tools into the product will ensure transparency in AI decision-making and ensure that communities, conservationists, and authorities can interpret AI outputs and challenge decisions where necessary. The team plans to conduct regular audits and establish community feedback loops to enhance the system's accountability and adaptability.

Conclusion and future perspectives

Scaling AI solutions for Human-Wildlife conflict mitigation requires a multi-pronged approach that addresses technical, ethical, and social challenges. There is not one solution to challenges in responsibly scaling AI systems. Rather, the operationalization of a diverse set of context-specific insights – from enhancing algorithmic fairness to building trust bonds with local stakeholders – will help in achieving this objective.

By working closely with their mentors, the team has developed tailored, actionable solutions to ensure their AI system scales responsibly, aligning with the principles of Responsible AI and promoting community-driven conservation efforts in a sustainable, transparent and inclusive manner.

Data Law Companion: Harnessing LLMs to create awareness on data protection laws in Kenya, Uganda and Rwanda

Team members:

- Dr. Lawrence Nderu, Jomo Kenyatta University of Agriculture and Technology
- Dr. Ruth Oginga, Kabarak University
- John Michael Rono, Jomo Kenyatta University of Agriculture and Technology
- Maryanne Mwihaki, Jomo Kenyatta University of Agriculture and Technology
- Derrick Matindo Obwatsa, Jomo Kenyatta University of Agriculture and Technology
- Millicent Shiatikha, Jomo Kenyatta University of Agriculture and Technology
- Sonia Lomo, Jomo Kenyatta University of Agriculture and Technology
- Daniel Muiruri, Jomo Kenyatta University of Agriculture and Technology
- Felista Mogire, National Irrigation Authority
- Bonface Ingumba, Office of Data Protection, Kenya

Mentors:

- Dr. Ugochi Adaku Okengwu, Associate Professor and Head of Department, Computer Science Department, University of Port Harcourt.
- Prof. Idy DIOP, Full Professor, Cheikh Anta DIOP University, Dakar Senegal ; Sorbonne Université - IRD - UMMISCO - F-93143, Bondy, France

Introduction

The rise in the use of emerging technologies continuously changes how businesses operate, creating new opportunities while simultaneously introducing regulatory challenges. In the East African region, where digital platforms are increasingly integral to commerce and daily life, noncompliance with data protection laws has become a critical issue. Business owners can frequently lack awareness of current regulations, leading to unintentional non-compliance.

The following project aims to mitigate this issue with a tool to simplify and ensure compliance with relevant laws and conventions. The platform, titled the Data Law Companion (DLC), aims to provide users with data law protection information in Kenya, Uganda, and Rwanda.

The implementation and deployment of the system comes with challenges unique to the context in which it may be introduced. Collaborating with their mentors, this team has explored their most significant obstacles with the objective of deriving insights toward responsibly and effectively scaling their project.

Key responsible AI scaling challenges identified at the outset of the mentorship programme

The team collaborated with their mentors to identify important challenges to tackle prior to developing their product

- **Biassed training datasets:** As they develop AI-driven features such as a summarising tool and chatbot, the team faces the challenge of ensuring that the algorithms they train do not perpetuate or amplify bias, particularly when handling diverse data sets from various industries and cultural contexts. Any bias in the responses provided by the AI tool could result in misinterpretations or compliance issues on the user's end.

- **Lacking collaboration with lawyers and officials**: As all countries where the team hopes to deploy their tool interpret data protection laws differently, The Data Protection Law AI tool was initially not received favourably by officials.

- **Regulatory compliance and need for human annotation**: Navigating the different data protection laws and regulations in Kenya, Uganda, and Rwanda demands a thorough understanding of the different legal landscapes and continuous monitoring to ensure compliance. This process is complex and resource-intensive. Human annotation by field experts such as lawyers is essential for ensuring the accuracy and relevance of the data.

- **Need for governance framework**: The primary focus was on establishing a robust framework that facilitates access to information on best practices in data governance and risk assessment as well as compliance with data protection laws in Kenya, Uganda, and Rwanda

- **Lack of sustainable funding structure**: To scale the project and develop it into a product that can be readily used requires funding. Securing sustainable financial resources emerged as a significant priority for this team to be able to expand their product's reach and continuously improve the platform's capabilities.

Strategic insights: Proceedings from the GPAI mentorship programme

After identifying the above obstacles, the team worked with their mentors to identify the subsequent mitigation strategies and solutions to the challenges they encountered in developing and scaling their project responsibly:

1. **Devising a data governance framework**: The team agreed on a protocol to combat bias in the data collected in the three pilot countries, which includes the establishment of data collection criteria, data standardisation and continuous monitoring. They also plan to implement algorithms designed to detect and reduce biases in the data and decisions made by the DLC platform to enhance fairness and aim for impartiality in compliance recommendations. Additionally, they added safeguards to protect user data gathered from interactions with the platform.

2. **Automation of compliance monitoring**: The team intends to develop tools that continuously monitor business practices for compliance, providing real-time alerts and recommendations and ensuring up-to-date knowledge of laws and regulations.

3. **Secure funding**: The team set out to seek additional seed funding to support the initial scaling phase, focusing on market expansion and product development. They hope to reinvest a portion of the revenue generated from subscriptions and services back into the platform to fund ongoing improvements and expansions.

4. **Continuous monitoring and updates**: Along with algorithmic automation, the team plans to establish an ethics committee to regularly review AI practices and ensure they align with ethical standards. They also intend to track metrics such as user logins, time spent on the platform, and feature usage to gauge engagement and identify areas for improvement. Lastly, they hope to measure the percentage of users successfully achieving compliance with data protection laws using DLC to assess the platform's effectiveness.

Conclusion and future perspectives

Through the mentorship programme, this team was able to devise a comprehensive plan for scaling their project responsibly, from the drafting of a data governance framework to automating and continuously monitoring their product for improvement.

In the near future, this team states their next steps include the launch of market initiatives, efforts toward product development and feature enhancement, as well as the implementation of bias mitigation algorithms and of user privacy protection policies.

DAWN AI Study

Team members:

- Victor Ogunbiyi, DAWN AI Study, Founder and CEO
- Oluwabukola Dada, DAWN AI Study, Co-founder
- Joel Danjuma, DAWN AI Study, CTO and AI Developer
- Fasobu Adekemi, DAWN AI Study, Project Manager
- Odewunmi Bolaji, DAWN AI Study, Product Designer

Mentors:

- Dr. Ugochi Adaku Okengwu, Associate Professor and Head of Department, Computer Science Department, University of Port Harcourt.
- Prof. Nobert Topze, Associate Professor in the Department of Computer Science at the University of Yaounde I and member of the Lirima Laboratory's IDASCO team.

Introduction

DAWN AI is a dyslexic friendly edtech platform powered by AI designed to make education accessible to everyone, regardless of their location, language, or abilities.

Team 7 proposes the development of an AI tool that caters to users with learning and other disabilities, and aims to create an inclusive, effective, and accessible educational platform that addresses the diverse needs of learners globally.

Key Responsible AI scaling challenges identified at the outset of the mentorship programme

The team identified the following challenges to responsibly scaling their product in their first mentorship sessions:

- **Inclusion of different user ability levels**: Seeing as all potential learners are differently abled, and that some may not be able to access information as straightforwardly as others, this team must develop a product that is accessible to as many people, which can prove complex depending on the features required. A user with dyslexia or ADHD, for instance, will need particular accommodations to use the app comfortably.

- **Transparency**: If the system requires the implementation of more specific mechanisms to be accessible to most, the workings of those accommodations and the AI model must be explainable to the system users so as to foster trust.

- **Privacy and trust risks**: The algorithms required to scale the platform to its context, i.e. users with different ability levels, require extensive data to be accurate, which raises privacy and data security concerns, particularly when dealing with sensitive information entered by minors for instance. There is also the potential for data leaks

– such as names and passwords, but also diagnoses and other personal information – if proper safeguards are not put in place.

- **Infrastructural limitations:** Developing adaptive AI algorithms capable of personalising learning for diverse needs, particularly for students with learning disorders, can prove challenging since the AI models tend to be complex.

Strategic insights: Proceedings from the GPAI mentorship programme

With the help of their mentors, the team identified the following risk mitigation strategies and ways forward:

1. **Inclusivity and accessibility features:** Developing dyslexia- and ADHD-friendly content and personalised learning paths, and AI transcription capabilities for local languages, including sign language interpretation. Regularly assessing transcription accuracy through user feedback and benchmark against accessibility standards will help ensure the AI system is accessible and accurate.

2. **Implementation of transparency measures**: Providing user-friendly explanations of AI-driven decisions and of the functioning of the models will help build trust in users and ensure they understand the basis of AI recommendations. Conducting user surveys will allow tracking model explainability and assist in performing continuous improvements.

3. **Devise a data privacy policy and robust mechanisms:** Creating robust data privacy measures – including explicitly stating that users' information will not be sold or shared with third parties – will be helpful in preventing data breaches or privacy violations, and so will performing regular audits and compliance checks. These include the integration of advanced encryption methods to safeguard user data and the use of anonymized data for model improvement.

Conclusion and future perspectives

The responsible scaling of DAWN AI Study requires taking into account both accessibility and infrastructure limitations. Working with their mentors, the team devised strategies to help ensure their AI solution brings educational access to underserved areas ethically and inclusively.

To do so, the team hopes to secure sustainable funding sources and government support in the establishment of AI education hubs in schools and universities. They intend to collaborate with startups working on inclusive AI solutions and foster partnerships between academia, industry, and government for responsible AI deployment.

In the future, this team plans to add premium features to their product such as enabling more personalised educational experiences according to each student's needs based on user feedback.

Non-intrusive fish weighing: Optimising fish feeding with data-driven insights

Team members:

- Teegwende Zougmore, Lecturer at Université Nazi BONI, Burkina Faso
- Ayodele Awokoya, PhD student, University of Ibadan, Nigeria
- Billy Peter Munyenyembe, PhD student, ZCAS University, Zambia
- Daphne Machangara, Institutional Analyst, Lupane State University, Zimbabwe

Mentors:

- Ismaël Koné, Lecturer and researcher in Artificial Intelligence (AI),Deputy Head of the Projects and Research and Research Engineering Unit (CIPRE),Virtual University of Côte d'Ivoire, Abidjan
- Elizabeth Oseku, MD, Infectious Diseases Institute, Kampala, Uganda
- Joel Nwakaire, African Technology Policy Study Network (ATPS) - University of Nigeria, Nsukka
- Mamadou Samba Camara, Ecole Supérieure Polytechnique - Université Cheikh Anta Diop

Introduction

To develop fish quality, researchers in the fish farming industry are interested in understanding the impact of feeding on the growth of fish. They do so by conducting experiments such as removing fish from their waters to weigh them periodically and monitor their growth. This process is not only tedious but induces stress in fish, which might negatively impact their health.

Team 15 is developing a system that uses computer vision machine learning to estimate fish weights according to data from images captured using underwater cameras. This team's project aims to automate labour-intensive processes and eliminate the need for manual handling of the fish. Their system uses non-invasive technology to reduce harm effected onto fish and minimise the risk of diseases, while still allowing to achieve the goal of optimising fish feeding practices and product quality.

Key responsible AI scaling challenges identified at the outset of the mentorship programme

To successfully implement and deploy their project, the team worked with mentors to identify and correct the issues they ran into in scaling their project responsibly. The challenges below emerged in large part during the scoping and definition phase of the project :

- **Data collection and research authorization**: Data collection requires a research authorization to be issued, which can be a lengthy or complex process.

- **Data use agreement**: The team had to determine ownership and copyright guidelines for the data collected during their study and the latter's proceedings, as well as formulate it into a formal agreement.

- **Data privacy and security risk**: As the data collected from the underwater images will be transmitted to a server via a network, a security and privacy risk exists and requires addressing to prevent unauthorised access and use. Both the transmission network and the database itself present vulnerabilities.

- **Data variety**: Ensuring the model's accuracy in varying conditions and with different species will require that it is adaptable enough to handle changing settings and subjects. Some differences may be subtle while others drastic, which both require the implementation of mechanisms to be effectively assessed.

Strategic insights: Proceedings from the GPAI mentorship programme

Following the identification of the issues above, the team collaborated with their mentors to devise safeguards and solutions toward ensuring RAI principles are adhered to in the design of their product.

1. **Securing a research site**: Data collection will take place in Zambia in collaboration with the Zambia Research and Development Center (ZRDC) team, whom the group has asked for permission to use their fish pond. The ZRDC may also fund the project.

2. **Implementation of data governance guidelines**: To clarify issues of data responsibility and accountability, the team have written a draft agreement and reached out to the ZRDC.

3. **Infrastructure securitization** : To mitigate infrastructural risks, the team will perform a review of the available equipment for data transmission and interoperable database systems. They plan to conduct a comparative analysis to select the system that will best minimise data security risks. They also aim to continuously refine the AI model through testing.

4. **Stakeholder engagement**: The team plans to adopt a stakeholder-first approach, reaching out to local government agencies such as the Ministry of Fisheries and Livestock and the Zambia Environmental Management Agency (ZEMA) to ensure alignment with national sustainability and regulatory goals.

Conclusion and future perspectives

Responsibly scaling this team's fish quality monitoring solution requires mitigating different risks, particularly as it comes to the data that drives the system and its securitization. Engagement with local stakeholders is also crucial at this stage to ensure the tool developed responds to their needs and reality.

The team wishes to secure governmental support toward creating regulatory frameworks for and deploying their system, encouraging sustainable fish farming and responsible innovation in aquaculture.

In the near future, the team hopes to agree on a data ownership agreement with the ZRDC and begin data collection. They plan to consult with local stakeholders to better understand their needs and the problems they experience in their activities, and tailor their project accordingly.

Delia: AI-powered chatbot and voice health assistant

Team members (AI-KING Group):

- Birba Delwende Eliane
- Ouedraogo Severine
- Birba Bienvenu Emmanuel
- Kinig Yilomou Ingrid
- Kone Abdoul Kader
- Armel Emmanuel Sogo
- Zombre Christian

Mentors:

- Shakira Babirye, Biostatistician, Data Scientist, Infectious Diseases Research Collaboration (IDRC), Kampala, Uganda
- Ibra Dioum, Lecturer, Researcher, Polytechnic School - Cheikh Anta Diop University, Dakar, Senegal

Introduction

The AI-KING Group developed SuiviVital, a digital health platform offering remote health monitoring and telehealth services to curb important delays or barriers to healthcare access due to high demand. Delia is an integrated AI-enabled service within this ecosystem designed to enhance patient communication, triage, and care within the healthcare sector in sub-Saharan Africa

The team's mentorship experience focused on issues of data gathering and handling and the generation of actionable insights into data governance in order to fine tune their product and implement and deploy it responsibly and in a context-appropriate manner.

Key Responsible AI Scaling Challenges Identified at the Outset of the Mentorship Programme

The team identified and addressed the following challenges during their mentorship sessions. Together with their mentors, they highlighted issues of data collection and usage, data governance and technical limitations to tackle toward scaling this product responsibly.

- **Inadequate and biassed datasets:** Insufficient or incomplete datasets limit the robustness and scalability of the AI solution. The team initially trained their AI tool on data that was biassed as it was sourced from a narrow, easily accessible group of people, leading to potential issues with the tool's generalizability and fairness. After discussions with their mentors, the team recognized the need for a comprehensive approach to ethical data collection.

- **Need for data collection authorizations:** Securing approval from the various data collection sites is often time-consuming, requiring not only the submission of detailed documents but also potential presentations to review boards. Additionally, the risk of delays or even rejections of approvals could impact the project timeline and deliverables.

- **Lack of data governance documentation**: Initially, the team did not have adequate documentation related to data governance, posing risks to data security, quality, and compliance.The mentors pointed out the need to develop comprehensive data governance policies that are in line with national data laws to enhance transparency and ensure compliance.

- **Infrastructure limitations in low-connectivity areas**: The team faced a technical challenge as the AI tool could only be supported via Wi-Fi, limiting its accessibility in rural areas where internet connectivity is lacking or unstable and narrowing the solution’s potential reach.

- **Development of proposals and consent forms:** The work to align these documents with GCP (Good Clinical Practice) and HSP (Human Subjects Protection) guidelines is a complex process that requires expertise in regulatory standards. The team must ensure that all documents meet these standards to secure approvals.

- **Funding for participant reimbursement:** Without financial backing from a grantor or funding agency, compensating participants for their time and effort may be difficult. This could negatively impact research ethics and participant engagement and retention, which are crucial for data integrity.

Strategic insights: Proceedings from the GPAI mentorship programme

After clarifying the obstacles that require clearing in order to scale their solution responsibly, the team defined the following solutions alongside their mentors:

- **Establishment of data collection guidelines**: Ensuring that the participant sample is representative of the broader population is essential for producing accurate and generalizable results. The team adopted a more balanced approach to data collection, ensuring representation from diverse groups, including rural and urban populations, men and women, and both public and private health entities. The team is currently awaiting approval from the Ministry of Health (MoH) to collect data from public hospitals, while permission from private entities is in the final stages of being granted.

- **Development of a data governance framework**: The team has worked on developing a data governance charter and data quality, usage and security policies that outline how data ought to be handled, managed and protected. They also plan on putting in place a data governance committee to ensure accountability and clear roles in managing data-related processes.

- **Development of compliant documentation:** To facilitate ethical data collection practices, they have created and finalised all necessary documents following GCP and HSP guidelines as well as all applicable laws. They have also prepared forms to secure participant consent and ensure they are informed about the data collection process. These documents are in the process of being submitted to the appropriate bodies, such as MoH and private hospitals, for approval.

- **Incorporation of accessibility features:** Following discussions on inclusivity, the team incorporated an SMS-based prompting feature into the tool. This enhancement ensures that users in rural areas with limited access to Wi-Fi can still benefit from the AI tool, expanding the project's reach and scalability. In order to increase accessibility and inclusivity, they have also integrated a WhatsApp voice feature that can be used by users with limited literacy, for instance, allowing for enhanced engagement.

Conclusion and future perspectives

As they adopt ethical data collection practices, strengthen their data governance framework, and improve the accessibility of their AI tool, the team is on a path to developing a more inclusive and responsible AI solution for the healthcare sector. While financial limitations remain an obstacle to adequately scaling their product, the team members are engaged in ongoing efforts to secure funding, such as applying for grants .

Moving forward, the team's next priority lies in retraining their AI model on ethically sourced and more representative datasets once they obtain the necessary approvals. Once the AI tool is operational, they plan to conduct a study assessing its impact and comparing health outcomes between patients who are and are not using the tool. They intend to share the findings with stakeholders, publish them in scientific journals, and present them at conferences to demonstrate the value of responsibly scaling AI solutions in the context of healthcare.

AI4Health Project: Strengthening the RAI pipeline for Mboacare and Mboathoscope

Team members (Mboalab):

- Stephane Fadanka: Executive Director, Molecular Biology Researcher
- Elisée Jafsia: Director of Digital Innovations and AI, Data Scientist.
- Bawang Bawa, Engineer, Industrial Robotics
- Ruqaiya Sattar, Developer, Software Engineer
- Agien Petra, Community Manager, Data Scientist

Mentors:

- Jean Louis Fendji Kedieng Ebongue, Associate Professor of Computer Science, Head of the Centre for Research, Experimentation and Production, School of Chemical Engineering and Mineral Industries, University of Ngaoundere, Cameroon
- Shakira Babirye, Biostatistician, Data Scientist, Infectious Diseases Research Collaboration, Kampala, Uganda

Introduction

The Mboalab team suggests the development of a solution to combat the issue of misdiagnoses and lack of medical equipment in Cameroon. The Mboalab AI4 Health Project is a set of two solutions: Mboacare, a chatbot for healthcare using AI-powered tools to analyse data and provide more accurate diagnoses, and the Mboathoscope, an open-source, wireless digital stethoscope.

The team has developed a prototype for the system, but the latter remains in the development stage for now. Notably, the team has started conceptualising a data governance framework, but needs additional skills and knowledge to better understand the stakes at hand and put in place a responsible framework.

To deploy their product successfully, the team must overcome a series of challenges identified as part of the mentorship programme, outlined below.

Key responsible AI scaling challenges identified at the outset of the mentorship programme

- **Lack of data privacy and ethics framework**: At the outset of the mentorship, the team did not dispose of a plan to ensure health data privacy, nor had they devised consent procedures or a plan for ensuring the transparency of their AI models. The AI models used in the two parts of their project were not explainable to stakeholders, including medical professionals and patients, who are their target users.

- **Need for equity and fairness considerations**: The team had not considered Indigenous data in the design of their model. As they sourced the dataset used to train the models from an open access platform, there were risks regarding potential biases in data that may cause harm to potential users. The team also came to the realisation that data representing minorities such as pregnant women was missing, jeopardising the representativity of their system and the accuracy of its output.

- **Need for localization and customization of the system**: The solution needed to be adapted to the local healthcare system the team aimed to integrate it into. They also needed to put in place further inclusivity measures, as the system was set to be deployed in rural areas but could potentially not be used by people with lower literacy levels, or who speak different languages than the dominant ones.

- **Technical and Infrastructure challenges**: In developing their system, the team initially did not consider issues of infrastructural limitations, such as unstable or unavailable networks in remote areas.

Strategic insights: Proceedings from the GPAI mentorship programme

Following the identification of the above obstacles, the team collaborated with their mentors to devise a set of actionable insights they could implement to address each of these challenges and responsibly scale their solution.

1. **Establish health data privacy and consent procedures:** The team was advised to develop a comprehensive data privacy policy aligned with global best practices, seeing as local regulations on data protection in Cameroon have yet to be defined. They plan to implement data anonymization and encryption techniques to protect patient information. Lastly, they aim to create and implement clear consent procedures using local languages, particularly for people in rural areas.

2. **Promote transparency and explainability of the AI models:** To mitigate the risk of developing an AI system that is not understandable to its target users, the team plans to create accessible documentation and visualisations to explain AI models' decisions and processes to non-technical stakeholders, including medical professionals and patients. They also aim to collaborate with local institutions, such as universities and hospitals, to audit AI models for bias, fairness, and transparency. As for explainability, they prioritised the use of interpretable machine learning models such as decision trees and explainable AI (XAI) methods to make model decisions understandable to all stakeholders.

3. **Mitigate biases in dataset and incorporate minorities:** The team plans to collect and integrate Indigenous data by partnering with local healthcare facilities, community organisations, and researchers to create a more representative dataset that reflects the local population's diversity. They intend to also incorporate data from minorities and ensure that the development process for the AI model includes inputs from diverse population groups, particularly those who are vulnerable or marginalised. From a technical vantage point, they aim to use oversampling or data augmentation techniques to enhance the representation of minority groups in the dataset and regularly evaluate their AI models for potential biases.

4. **Adapt the project to its local context**: Collaborating with the Ministry of Public Health, local hospitals and health centres, the team aims to understand the specific needs and constraints of the Cameroonian healthcare system to tailor their solution to its infrastructure, workflows and protocols. They also set out to develop user-friendly features with intuitive interfaces to enhance access for individuals with low literacy levels and mobilise healthcare workers as liaisons to help explain and implement AI solutions in various communities. Lastly, they hope to translate AI tools and outputs into major local languages (e.g., Fulfulde and Ewondo) and adapt them to local cultural contexts.

5. **Address limited infrastructure and connectivity issues:** To address this issue, the MboaLab team set out to develop offline functionalities and lightweight AI models that can operate without requiring uninterrupted internet connectivity, using edge computing. They are also considering the use of mobile applications or SMS-based solutions that work with low-bandwidth networks to reach remote areas.

Conclusion and future perspectives

The Mboalab AI4Health team initially faced several challenges in responsibly scaling their AI solutions, particularly in data privacy, inclusivity, localization, and infrastructure. Without a data privacy framework, consent protocols, or model transparency frameworks, their AI system lacked key responsible AI foundations.

The mentorship has empowered the Mboalab team to further the development of their AI system with a socially responsible and community-centred approach. Next steps for the team include continuing to work on the development of a data governance framework including core principles, stakeholder engagement strategies, and policies for data stewardship and ownership.

The mentors will assist the team in engaging with the Ministry of Health and local health centres to begin the process of implementing and deploying their product into the Cameroonian healthcare system once it is ready to be incorporated.

Large language models for sexual, reproductive, and maternal health rights

Team members

- Walelign Tewabe Sewunetie: Academic Staff and Researcher, Debre Markos University
- Hailemariam Abebe, Academic Staff and Researcher, Debre Markos University
- Tesfamariam Mulugeta, Academic staff, University of Gondar and Queen's University
- Surafel Tilahun, Associate Professor, Addis Ababa Science and Technology University

Mentors:

- Dr. Mesfin Fikre Woldmariam, Faculty Member and Researcher, Addis Ababa University
- Dr. Lawrence Nderu, Founder, JHUB Africa Digital Innovation Hub. Lecturer, Department of Computing, Jomo Kenyatta University of Agriculture and Technology (JKUAT), Lead Researcher, AI and Digital Innovations in Agriculture, Finance, and Health, Technology Consultant and Innovator

Introduction

This team's project aims to address the challenge of leveraging artificial intelligence to enhance access to sexual, reproductive, and maternal health (SRMH) information and services. In under-resourced healthcare settings within Sub-Saharan Africa, a significant disparity remains in the accessibility of SRMH services.

This initiative proposes the development and deployment of large language models (LLMs) to bridge this gap, with a focus on ensuring that essential health information is ethically and responsibly delivered to marginalised communities.

Key Responsible AI Scaling Challenges Identified at the Outset of the Mentorship Programme

In the initial stages of their project, the team worked with their mentors to identify areas to prioritise in order to mitigate risks to responsibly scaling their solution. The main challenges have been identified as follows:

- **Cultural context awareness:** Sexual and reproductive health is a culturally sensitive area. The large language model used by this team must be designed to respect the cultural norms and values of the communities it will serve, particularly in Ethiopia and other Sub-Saharan regions. Ensuring that the AI models provide accurate, respectful, and appropriate information without causing harm or creating misunderstandings emerged as a significant challenge.

- **Data privacy and security:** Handling sensitive health data, especially related to reproductive health, can raise significant privacy concerns if proper safeguards are not established. In low-resource settings, ensuring data protection and compliance with international standards such as the GDPR can prove challenging. In order to offer a trustworthy AI system, the team must prioritise the development of robust data privacy measures.

- **Overcoming language barriers:** The rich linguistic diversity present in the region can become an obstacle to scaling this project responsibly. Ethiopia alone counts over 80 spoken languages and dialects. To ensure accessibility, the team's model requires training in multiple local languages, making the collection of high-quality, diverse data essential. Additionally, incorporating dialects and underrepresented languages in the design of the AI model can prove challenging.

- **Closing digital and infrastructure gaps:** Rural and remote communities can be underserved with respect to access to digital technologies and reliable internet networks. This digital divide could limit the reach of the team's health solutions as it hinges on the use of large language models and would require internet connectivity. In order to enhance its adoption, the team must make their solution adaptable to low-tech environments to ensure widespread access.

Strategic insights: Proceedings from the GPAI mentorship programme

To mitigate the issues above, the team created the following roadmap with their mentors:

1. **Contextualized AI model:** The team collaborated with local stakeholders, including healthcare professionals, community leaders and cultural experts, to ensure that the LLMs are contextually and culturally appropriate. They believe this collaborative approach will help mitigate the risk of cultural insensitivity and misinformation, aligning the models with the needs and values of their target communities.

2. **Data privacy and transparency by design:** The team has integrated a data privacy framework into their AI system in the early phases of their project, focusing on anonymizing sensitive health information and ensuring consent is collected in a transparent manner. The mentorship highlighted the importance of implementing responsible data governance practices tailored to local regulatory contexts.

3. **Training the LLMs on local languages:** The mentors emphasised the need to focus on language inclusivity to prevent further marginalisation of communities and avoid replicating biases and into their AI model. To address the language barrier to access, the team worked on creating training datasets for underrepresented languages in Ethiopia. This effort involved partnering with local universities and linguistic experts to gather and curate high-quality data to train the LLMs on.

4. **Inclusion of offline mode:** In response to infrastructure limitations, the team is developing a version of the system that operates offline, without network connectivity if the latter is not available. This will help in ensuring that even communities with limited digital infrastructure can access vital health information. The team plans for this low-tech version of the project to be accessible via SMS or community health worker systems, creating a more inclusive solution.

5. **Supporting digital literacy**: Developing and distributing digital literacy materials is key for empowering users to effectively utilise the team's AI solution. Providing comprehensive guidance on responsible AI use will be instrumental in enhancing user experience and transparency, and promoting informed decision-making.

6. **Stakeholder engagement**: The team plans to engage with local health professionals and community leaders to include them in decision-making processes and develop a solution that is tailored to their reality. The team intends to develop a structured engagement strategy to sustain stakeholder support and ensure project relevance.

Conclusion and future perspectives

From localisation and contextualisation of the system's LLMs, to bridging the digital divide with literacy materials, to enhancing and prioritising stakeholder engagement, the team gathered valuable insights from their experience in the mentorship to achieve their goal of developing an AI-driven solution to increase access to reproductive health information and treatment in Ethiopia and Sub-Saharan Africa.

These insights have helped the team ensure that their solution is not only scalable but also aligned with ethical standards while safeguarding the rights and dignity of the individuals and communities they aim to serve.

By leveraging AI responsibly, the team hopes to enhance access to healthcare, decrease maternal mortality rates, and support national efforts to achieve the health-related Sustainable Development Goals (SDGs), specifically SDG 3.

BESHTE: A Chatbot to enhance HIV testing, status awareness, and status disclosure among adolescent boys and girls and young men and women in Kenya

Team members (Innovate AI Health Lab):

- Dr. Victoria Mukami,
- Dr.Consolata Gakii
- Dr.David Mugo
- Dr.Marilyn Ronoh
- Mr. Stanley Ndegwa

Mentors:

- Julius Niyongabo, Ph.D. Lecturer-Researcher, Faculty of Engineering Science, Department of Information and Communication Technology, University of Burundi and Olivia University, Bujumbura
- Shakira Babirye, Biostatistician, Data Scientist, Infectious Diseases Research Collaboration, Kampala, Uganda

Introduction

This team is engaged in creating a chatbot called BESHTE, which aims to display traits like empathy and understanding while providing culturally contextualised information to increase HIV awareness among Kenyan adolescents and young adults.

Successfully and responsibly scaling BESHTE requires addressing issues like linguistic diversity and enhancing its empathy and cultural sensitivity to increase its adoption by a broader audience. Throughout their mentorship, the team aims to develop their chatbot into a tool with the ability to adapt to different reproductive health issues and local contexts in order to improve health outcomes and reduce the rate of new infections while improving youth health behaviours.

Key responsible AI scaling challenges identified at the outset of the mentorship programme

- **Data privacy risks**: As the chatbot stores and uses confidential health data, lacking or insufficient guardrails and safety measures could lead to breaches and/or leaks. Appropriate measures need to be put in place with respect to both data handling practices and infrastructure.

- **Lack of user inclusivity**: Lack of inclusivity of populations without access to smartphones and people with disabilities such as visual impairments can pose an obstacle to effectively scaling this product. Additionally, lacking inclusion of languages other than English or Swahili within the AI solution holds potential to hinder adoption for users who are not fluent in either.

- **Improper data collection and handling practices**: As the team aims to raise awareness through their product, accuracy of the information provided by the chatbot is crucial. To be valid, this information hinges on the collection of quality, contextualised data. Depending on the data used toward creating the system's knowledge base, the chatbot could hold potential to be misinformed. Moreover, the data used to develop the knowledge base is usually collected from the users and cleaning this data can pose an important obstacle.

- **Lack of funding to support language features**: The development and incorporation of additional languages into the chatbot's knowledge base will require funds that the team has not yet sought out. The resources to seek out include additional data collection as well as translation expertise.

Strategic insights: Proceedings from the GPAI mentorship programme

Working with their mentors to identify the following insights, the team plans to operationalize them to mitigate the risks and challenges mentioned above.

- **Trust and safety measures**: Making the data used by the chatbot confidential on both the users and server's end will ensure that user data is not leaked or accessed by an unauthorised user. The team plans to encrypt the data so it is only accessible to authorised personnel. Moreover, ensuring that identification data such as user account names is not associated with the training data used to retrain the chatbot is a priority.

- **Enhancements to inclusivity**: To increase accessibility, the team aims to incorporate high-quality translation of the chatbot data to different local languages by leveraging machine translation and human post-editing to review and refine the output, and ensure quality and cultural relevance. They intend to incorporate local languages like Meru to make BESHTE accessible to a wider audience, particularly in rural and underserved regions.

- **Ethical data handling practices:** Cleaning and getting the chatbot data peer reviewed by experts will help ensure the validity and accuracy of the data used to develop the knowledge base. A combination of machine learning approaches to data filtering and cleaning and human oversight will allow the team to remain efficient, accurate and ethical in their data collection and handling processes.

- **Seek out additional funding**: The team was advised to pursue additional funding sources to support individuals with disabilities and other underrepresented groups and allow them to scale up their product and increase its inclusivity.

Conclusion and future perspectives

At the term of the mentorship, the team had gained actionable insights into implementing trust and safety measures—such as encryption and peer review processes—to improve their AI system's reliability and effectively build user trust. As they operationalize these findings, the team intends to ensure that their AI model is fair, as well as incorporate features like speech recognition into future versions for enhanced inclusivity and accessibility.

Once they scale up their product, they plan to perform infrastructural upgrades to enable the system to handle larger volumes of data, as well as strengthen their data protection measures.

Regulatory AI : An AI-powered solution for health compliance and regulation

Team members (Feyti Medical Group):

- Kansiime Noel
- Hassan Bahati Mukisa
- Kesandu Uchenyi
- Mark Tushemerirwe
- Barbara Atukunda
- Lilian Mirembe
- Moses Ariong

Mentors:

- Joel Nwakaire, Professor, African Technology Policy Studies Network and the University of Nigeria Nsukka
- Elizabeth Oseku, MD, Infectious Diseases Institute, Kampala, Uganda

Introduction

The current processes in place in Uganda for creating and managing documents attesting to drug compliance with Good Manufacturing Practices (GMP) pose challenges of inefficiency, high costs and heightened risk for error and inaccuracy. In response to these and to issues of high processing times in pharmaceutical administration, this team proposes a document management solution that uses AI to streamline GMP certification and enhance regulatory compliance for herbal practitioners and pharmaceutical companies in the Ugandan pathogen mitigation industry.

Their deep dive focused on the development of a prototype for a regulatory AI system that helps with the management of drug dossiers, clinical trials, and document submissions. Powered by Gemini 1.5 Pro, “Regulatory AI” offers a solution to shorten the length of time required – about 20 working days – to create simple documents, the maintenance cost for safety databases as well as outdated manual data entry processes that increase risk for error.

To move past roadblocks in successfully and responsibly scaling this AI solution, team 10 collaborated with their mentors toward the elaboration of a plan to mitigate and solve the key risks and issues by leveraging actionable insights that emerged from the mentorship.

Key responsible AI scaling challenges identified at the outset of the mentorship programme

Below are the main challenges to operationalizing RAI principles encountered by the team in the initial phases of their project:

- **Lack of stakeholder engagement**: The team struggled to gain the engagement of some key stakeholders, particularly regulatory bodies. The mentorship highlighted the need for more meaningful collaboration with Uganda’s National Drug Authority (NDA) and other stakeholders involved in the implementation and deployment of their product more directly to ensure alignment with regulatory standards.

- **Absence of a formal data governance framework**: The team's overall product policy lacked important safeguards, particularly with respect to data privacy. Further emphasis was placed on developing a data governance framework to prevent data breaches, among others.

- **Limited user testing and feedback mechanisms**: In order to create a product that accurately and efficiently responded to the needs of its target users, the team needed to carry out further testing and gather user feedback. The identification of usability challenges and the creation of a user-friendly interface were identified as priorities.

- **Need for human oversight and clear responsibility allocation**: As advanced as they could make their software, some tasks and workflows require a human in the loop to ensure they are carried out ethically. The team were encouraged by their mentor to integrate human oversight into the system for purposes of monitoring data integrity, biases, and transparency. Safeguards needed to be put in place to ensure responsible data gathering, handling, storage and use.

- **Potential biases in the AI system and data**: A new issue surfaced in the mentorship, namely the inclusion of less technologically literate or savvy users, such as small-scale herbalists who may lack the resources to engage with an AI-powered online platform. Ensuring that the system accommodates underrepresented groups and providing technical support to ensure successful adoption by the latter emerged as a priority for the team. Minimising bias in both the scoping and data collection stages by respectively representing different demographics and ensuring data quality and diversity, was another priority.

Strategic insights: Proceedings from the GPAI mentorship programme

Following the identification of the above risks and challenges, the team worked with their mentors to articulate the following actionable insights, which they plan to implement into an action plan to responsibly scale their AI product:

1. **Enhancing stakeholder engagement**: The team was advised to develop a stakeholder mapping and engagement plan with regular updates, feedback mechanisms, and involvement in decision-making. Prioritising key actors such as the NDA, they are to set up regular consultations with the organisation and other regulatory bodies to align with their digital roadmaps. They also plan to engage with and diversify their end-users (e.g. large pharmaceutical companies, independent herbalists) to understand their varied needs. Finally, the team will also review Uganda's regulatory frameworks and engage with policymakers toward the drafting of policy briefs.

2. **Developing and implementing a data governance framework**: To create guardrails for safe data collection, storage, protection, and quality assurance, the team has worked on developing a comprehensive data governance framework addressing these issues as well as assigning a dedicated data protection officer. They also plan to implement clear data privacy statements and user consent

mechanisms into the platform in compliance with regulations such as the General Data Protection Regulation (GDPR) and Uganda's Data Protection and Privacy Act.

3. **Conducting user testing and gathering feedback**: The team plans to design and implement a phased user testing plan that includes both quantitative and qualitative assessments. They aim to develop and implement an ongoing feedback mechanism within the platform to ensure its continuous improvement. Lastly, they intend to conduct iterative testing and refinement cycles prior to their product's full deployment.

4. **Inclusion of human oversight structures**: To mitigate risks that could emerge from lack of human oversight into the system, the team set out to establish protocols for regularly scheduled audits and reviews of the AI system's outputs as well as other "human-in-the-loop" mechanisms for critical decision-making processes. They plan to define clear roles and responsibilities for human oversight, including data integrity, security, and compliance roles.

Conclusion and future perspectives

As part of their experience in the mentorship programme, the team has developed a draft governance framework that will be continuously fine tuned as development of the solution continues.

Regulatory AI is currently in the prototyping and development stage and will soon undergo its first round of user testing. By the end of the year, the team aims to identify key opportunities in Ugandan technology and data privacy policies to engage policymakers. The team is considering potential integration of their solution within the systems used by regulatory bodies like Uganda's NDA, and has made initial efforts to engage with the organisation and other relevant stakeholders.

As they look forward to scaling their solution across borders, the team set out to research and document regulatory requirements for a few target countries such as Nigeria, Tanzania and Kenya. They aim to develop a flexible system architecture that can adapt to different regulatory environments, and establish partnerships with local experts in each target country.

AI-Driven carbon credit calculations for electric motorcycle fleets

Team members (YNA Kenya):

- Sebastian Mwaura, CEO and Co-Founder
- Mary Munyao, Chief Operating Officer
- David Munene, Chief Finance Officer
- Ruth Wangui, Chief Administrative Officer
- Fred Omollo, Chief Technology Officer
- Lucy Njoki, Head of Growth
- Samuel Maloba, Head of IT
- Fabian Cheruiyot, Software Developer

Mentors:

- Dr. Mamadou Samba Camara, Associate Professor of Computer Science, École Polytechnique Supérieure, Cheikh Anta Diop University of Dakar
- Joel Nwakaire, Professor, African Technology Policy Studies Network and the University of Nigeria Nsukka

Introduction

This team proposes the development of an AI solution for calculating carbon tax credit in delivery services that use motorcycles as their primary mode of transportation.

Working with their mentors, the team identified important obstacles to responsibly scaling their project, which they aim to mitigate with actionable insights gained during the mentorship experience.

Key responsible AI scaling challenges identified at the outset of the mentorship programme

Below are the main challenges as reported by the project team.

- **Data collection challenges**: The main challenge in this project was ensuring precise measurement and transparent allocation of carbon credit. It was also critical to ensure that the IoT device used produced accurate measurements especially in conditions where internet access may be restricted. The proposed system must also prevent potential for false location data, which could lead to inaccurate carbon credit generation from stationary locations such as warehouses.

- **Infrastructure scaling challenges**: Scaling up the system infrastructure to support real-time data processing**,** AI model training, and blockchain integration poses technical and resource-related challenges.

- **Lack in availability of resources**: The project faces a financial barrier and unavailability of expertise related to the initial investment required for deploying IoT gadgets for real-time data tracking. The reported scarcity in skilled tech labour in the area, especially in hiring experienced developers proficient in AI data collection, blockchain, and IoT technologies, was also an issue.

- **Lacking stakeholder engagement**: The end users of this project are not the drivers but the people to whom the drivers transport the packages. There need to be structures and methods put in place to consult them and collect their feedback.

- **Ensure legal and regulatory compliance**: The team must identify Kenya's laws and regulations regarding carbon credits and explain how the system designed complies with these.

Strategic insights: Proceedings from the GPAI mentorship programme

Following the application of a qualitative AI risk assessment guide, the team arrived at the following conclusions to mitigate the aforementioned risks and challenges with the help of their mentors:

1. **Accurate data collection and calculation of carbon credits**: The team was advised to design their model so as to accurately calculate carbon credits based on factors like distance travelled, vehicle type, and environmental impact. They were also encouraged to implement mechanisms such as data verification methods to detect changes in location, ensuring that carbon credits are only calculated and posted on the blockchain when electric motorcycles are in motion. This will help prevent fraudulent activities and maintain the integrity of the carbon credit system.

2. **Integrate methods or platforms to engage stakeholders**: Mentors highlighted the importance of stakeholder engagement, particularly in consulting package recipients, the product's end-users, to gather valuable feedback. The team was advised to establish multi-stakeholder, adaptive governance guidelines and foster transparent, collaborative governance by bringing together multiple stakeholders together to participate in dialogues and decision making processes toward implementing their solution.

Conclusion and future perspectives

The team gained insights on responsibly gathering data and accurately calculating carbon credits, as well as integrating multiple stakeholders to scale their project responsibly.

They also plan to develop their project with respect to adapting the system infrastructure to local context, securing human and financial resources and ensuring legal and regulatory compliance. As for next steps on their journey, the team has not yet gathered all the relevant insights necessary to address further aspects of their project during the mentorship and, as a result, remains open to further exploration and inquiry.