GLOBAL TASK FORCE

PREDICTIVE ANALYTICS

FOR SECURITY AND DEVELOPMENT

TECHNICAL NOTES

Authorship Robert Muggah, Gabriella Seiler and Gordon LaForge

This project is partially funded by the Global Innovation Fund

IGARAPÉ INSTITUTE
a think and do tank

NEW AMERICA

Summary

Task Force Context

The development of new predictive technologies is accelerating and transforming societies worldwide. As governments, companies and nonprofits rush to deploy predictive analytics to “optimize” everything from the deployment of policing resources to preventing illegal deforestation or improving public transit and energy consumption, the potential risks are not receiving the attention they deserve. This is especially the case in developing countries that are still in the midst of their own digital revolutions and have their own baseline needs, concerns, and social inequities to consider when deploying these tools.

The Igarapé Institute and New America believe crafting the appropriate frameworks will require forging a consensus on the basic principles that should inform the design and use of predictive AI tools. Moreover, while there have been several initiatives from civil society and intergovernmental organizations on ethical AI standards, the discussion has been concentrated overwhelmingly in North America and Western Europe. Governments in lower- and middle-income countries continue to struggle with consequential decisions about trade-offs of deploying predictive analytics.

This new Global Task Force aims to bridge this gap by convening digital-rights advocates, public-sector partners, tech entrepreneurs, and social scientists from the Americas, Africa, Asia, and Europe, with the goal of **defining first principles for the use of predictive technologies in public safety and sustainable development in the Global South.**

Task Force Members

Reuben Abraham

CEO, Artha Global

Yolanda Martinez

GovStack Initiative Lead, International Telecommunications Union

Chinmayi Arun

Research Scholar in Law and Executive Director of the Information Society Project,Yale Law School

Nanjala Nyabola

Writer, researcher, political analyst, and activist

Wafa Ben-Hassine

Human rights lawyer, and Principal, Responsible Technology, Omidyar Network

Taylor Owen

Associate Professor and Founding Director of the Center of Media, Technology and Democracy, Max Bell School of Public Policy,at McGill University

Tiffiniy Cheng

Co-founder, Fight for Future and Open Congress

Maria A. Ressa

CEO, President, and Co-Founder, Rappler and Nobel Laureate

Rumman Chowdhury

Former Director for ML Ethics, Twitter and AI Fellow,Berkman Klein

Fabro Steibel

Executive Director, Institute for Technology and Society Rio

Haksoo Ko

Professor of Law, Seoul National University School of Law (observer)

Stefaan Verhulst

Co-Founder and Chief Research and Development Officer, Governance Lab (GovLab),New York University

Ronaldo Lemos

Professor, Rio de Janeiro State University Law School

Andrew Wyckoff

Director of the Directorate for Science, Technology and Innovation, Organization for Economic Cooperation and Development (OECD)

Vukosi Marivate

ABSA Chair of Data Science, University of Pretoria

Objectives

We aim to develop principles and practical recommendations for the ethical and inclusive design and use of predictive analytics for security and development, with a focus on lower and middle-income countries. How to best identify and mitigate potential risks while advancing the benefits associated with these tools?

For the purpose of our discussions and recommendations, we will organize the conversation in three phases:

March, 2023
Introductions and debate on major issues concerning design and deployment of algorithms in each task force member's field

June, 2023
Concerns and recommendations related to the **design** of predictive analytics tools

October, 2023
Guidelines and best practices for **deployment and evaluation** of predictive analytics tools

For each discussion we aim to consider (i) the existing regulation, experience and expectations of predictive analytics in the public safety and sustainable development sectors, (ii) risks and issues including accountability, transparency, fairness and discrimination, with a view of proposing minimum social impact measures, and (iii) best practices and processes from around the world to address (intended and unintended) outcomes in the use of predictive analytics tools in developing countries and cities.

The Igarapé Institute and New America will issue short briefs after each meeting and a public report at the beginning of 2024, in time for the Summit of the Future. The final list of recommendations will also be circulated in late 2023 for inputs from the task force. All of the virtual and written proceedings will be conducted according to Chatham House Rule unless explicitly requested by task force members and task force members will be credited in the final report.

Scope & Definitions

Predictive analytics refers to the use of real-time and historical data to define the probability of certain events occurring. For the purpose of our discussions, we will focus specifically on the development and deployment of algorithms that aim to monetize predictions by using a range of methods, including machine learning, to extract insights from existing data sets.[1]

Sustainable development predictive tools include predictive analytics designed to advance the Sustainable Development Goals ("SDG") as defined by the 2030 United Nations Agenda for Sustainable Development. A study published in Nature Communications Magazine has shown that the development and deployment of AI may enable 134 SDG targets, including uses related to public security, but also in education, water and sanitation, energy, sustainable cities and other sectors.

Public security predictive tools include surveillance and monitoring systems focused on crime prevention and the protection and safety of citizens. Common examples include surveillance cameras and facial recognition for predictive policing.

Some other examples of uses of predictive analytics in sustainable development and public security[2]:

- Crime Prediction: used to analyze crime patterns, predict future criminal activity, and inform police deployment.
- Emergency Response: models to forecast natural disasters and other emergencies, allowing public services to respond more effectively.
- Healthcare: predictive analytics used in public healthcare to forecast disease outbreaks, plan preventive measures, and optimize resource allocation.
- Traffic Management: predictive analytics helps cities predict traffic congestion and optimize traffic signal timings and road network design.
- Social Services: tools to predict and address social needs, such as homelessness and food insecurity, and allocate resources more effectively.
- Environmental Management: predictive models are used to forecast and respond to environmental hazards, such as air pollution, droughts, and natural disasters.
- Education: predictive analytics can be used to improve learning outcomes and optimize resource allocation in education.
- Energy Management: predictive analytics directed at optimizing energy production and distribution.

1 While concerns related to data collection, including privacy, also merit significant caution and attention from governments and policymakers, we will not be focusing on them due to our limited time and the vast amount of work already advancing guidelines on this important step of the process. We will instead focus on the systems and platforms employed to extract value from data once it has been collected.

2 Chat GPT3 https://chat.openai.com/chat

Discussion #1

March, 2023

Introductions and Major Concerns

Our first discussion will focus on each task force member's major current concerns with respect to the design and deployment of predictive analytics tools in their field of work, with a specific focus on lower and middle income countries. Below we provide some context of the current debate around predictive analytics to help kickoff the discussion.

Investments in AI

- According to Stanford University's Artificial Intelligence Index Report 2022 "private investment in AI in 2021 totaled around USD $93.5 billion". At the same time "research on fairness and transparency in AI has exploded since 2014, with a fivefold increase in related publications at ethics-related conferences"[3].
- Investments vary significantly by region and country with the US and China and Europe taking the largest shares.
- The recommendations of this task force become even more relevant as lower and middle income countries strive to catch up and not be left behind in the adoption of predictive analytics tools that can be used to leapfrog stages of economic development.

3 Daniel Zhang, Nestor Maslej, Erik Brynjolfsson, John Etchemendy, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Michael Sellitto, Ellie Sakhaee, Yoav Shoham, Jack Clark, and Raymond Perrault, "The AI Index 2022 Annual Report," AI Index Steering Committee, Stanford Institute for Human-Centered AI, Stanford University, March 2022.

International Standards and Regulation: Real and potential gaps related to regulation of predictive analytics in the Global South

- As a starting point to our discussion, we have mapped out the major international developments in responsible and ethical use of AI:

Over the past decade, public organizations, research institutions and companies from around the world have created several guidelines and principles for ethical AI. The first instruments to emerge were from Western liberal democracies, including the: Asilomar AI Principles (2017); Montreal Declaration for Responsible AI (2017); Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems (2018); General Principles of Ethically Aligned Design (2017); Five Overarching Principles for AI Code (2018); Tenets of the Partnership on AI (2018); and the European Commission's Ethical Guidelines for Trustworthy AI (2019).

Since then, efforts have been made to broaden the conversation on ethical AI to a more representative international audience to ensure equitable input from countries in the Global South. Accordingly, in 2019 the Organisation for Economic Development and Cooperation's (OECD) published its Principles on Artificial Intelligence, and in 2021 the United Nations Educational, Scientific and Cultural Organisation (UNESCO) adopted its Recommendation on Ethical AI (2021), which establishes the first global agreement on the ethics of AI for 193 member states. A core objective of UNESCO's Recommendation on Ethical AI is to focus on the practical realization of these ethical principles by creating a framework that leverages the knowledge and experiences of different contexts. The UNESCO protocols specifically target nations in the Global South, including Low to Middle Income Countries (LMICs), which have not enjoyed the same level of influence in developing normative frameworks as countries in the Global North.

In addition, a recent initiative by Data for Development Network (D4D) and Research ICT Africa will advance responsible AI by drafting a set of benchmarks that will measure a country's adherence to human rights principles in the development and implementation of AI systems. The Global Index on Responsible AI will establish a set of indicators that rank countries according to their capacities and commitments to: (1) use AI systems to advance human rights agendas; and (2) implement risk mitigation measures to respect and promote civil and political rights. The Global Index will establish regional hubs and capacitate researchers in more than 100 countries to conduct independent research using inclusive and participatory methods to measure country commitments to responsible and ethical use of AI. A central focus of its research will be on the experiences of historically marginalized communities to assess whether they enjoy equal opportunity to benefit from the promises of AI-driven technologies.

Some Key Practical and Ethical Challenges and Risks

The normative framework for ethical AI includes a set of six overlapping principles, including: explicability, accountability, fairness, oversight, privacy, and human-centered. We propose to structure our mapping and recommendations around each of these principles.

With regards to each of these principles, which are the most relevant risks to lower and middle income countries and how can these be mitigated?

- Explicability: people must be able to understand what it does, how it works and the risks involved,
- Accountability: ensures the proper functioning of systems and takes responsibility when things go wrong,
- Fairness: does not perpetuate bias or impose unfair discriminatory outcomesagainst particular categories of persons,
- Oversight: the power to decide whether to act ultimately rests with humans,
- Privacy: privacy, security,
- Human-centered: promoting well-being, preserving dignity and sustaining the planet.

Proposed Best Practices and Processes: emerging lessons from around the world

To help guide our recommendations we have mapped out a few examples of emerging best practices and processes that contribute towards responsible and ethical AI and ways to ensure that AI-driven technologies do not impose unintended harms.

Best Practices

- Independent Fairness Testing: used to detect forms of algorithmic bias that may create or reinforce disadvantages or discriminatory practices against disadvantaged and underrepresented groups
- Third-party auditing: users of AI should invite independent and experienced third parties to understand and review their algorithmic decision systems, which requires disclosing sufficient information to allow accurate testing, monitoring and feedback
- Social Impact Assessments: measure the impact of AI-driven technologies on the social elements of life on affected groups of stakeholders

Processes

- Fairness-by-Design: involves examining different parts of the machine learning process from different vantage points, using an interdisciplinary team of experts with different theoretical lenses
- Privacy-by-Design: methodology to ensure that privacy principles are embedded into the products from their conception through the development process.
- Ethics-by-Design: This methodology provides guidance for embedding ethical principles in the design, development, and deployment of AI based solutions.

Assessing risks and best practices in the design and development of AI systems

June, 2023

There are multiple concerns when it comes to the design and development of AI systems to improve public safety and economic development, not least in low-to-middle income countries. Among these are fears of existential threats, concerns with job losses, deepening of inequalities, and the spread of misinformation and disinformation. There are also more practical worries related to bias and discrimination. Drawing on open and public research generated by non-profits, companies, and intergovernmental initiatives, the following background paper highlights a sample of salient issues and remedies to drive the Global Task Forces' discussions. Divided into two parts - key concerns and best practices - it is intended to provide a cursory summary of salient debates with respect to the design and development of algorithms.

Key Concerns in Algorithmic Design and Development

Several concerns arise when considering the design and development of algorithms to drive security and development outcomes. There are particularly risks associated with biases and discrimination reinforced by the algorithms, the data they are trained on and the developers who are creating the underlying systems. It is important, then, to consider "who" is designing the tool, "how" are algorithms being developed, and "what" the algorithm is intended to achieve. Several of these issues are treated in more detail below.

Who is designing and developing the algorithms?

The identity, expertise, and institutional affiliations of AI developers can have a bearing on the values, goals, and priorities imbued in an AI system. Yet, to date, the design and development of AI systems are highly concentrated in very specific regions and companies. Indeed, with some exceptions, most AI development is occurring in companies, laboratories and universities in North America, Western Europe and China.

Specifically, AI research and development is concentrated geographically. One study of six major American AI developers – Amazon, Apple, Facebook, Google, IBM, and Microsoft – found that these firms' AI labs were concentrated in major cities primarily in the US, France, the UK, China, and Israel, with only three labs in Africa and none in Latin America; similarly, 68 percent of these companies' AI staff are located in the United States.

What is more, AI journal publications and citations are also highly concentrated. Recent literature reviews indicate that 31% of journal publications come from China, 19% from the EU and UK, and 14% from the US. Yet just 3.5% come from all of Latin America and the Caribbean and 1% from Sub-Saharan Africa. The skewed concentration of researchers and research articles underlines the priorities attached to specific types of algorithmic research and underlying data on which it relies.

As private investment in AI skyrockets – doubling from 2020 to a total of $93.5 billion in 2021 – the number of newly funded AI companies fell from 1051 companies in 2019 to 746 companies in 2021. One analysis at the end of 2022 found that more than 50% of all new AI venture capital investments went to companies based in the San Francisco Bay Area ($6.8 billion), followed by New York City ($1.1 billion), London ($500 million), and Tel Aviv ($388 million).

How is the algorithm being designed and developed?

If the datasets on which algorithms are trained are not sufficiently representative, this can result in unintentional stereotyping, bias, and discrimination. Moreover, a lack of engagement with a diverse group of stakeholders and representatives of affected populations during the design and development process can mean the system might reproduce adverse real-world consequences.

Of course, designers of AI systems routinely consider potential impacts of algorithms. But they are often narrowly focused on a specific context or intended use case. Once deployed tools are applied to new contexts, this can generate unintended consequences. Thorough due diligence must consider the responsibility of designers and developers in frequently adapting systems to consider newly identified stakeholders and use cases, including in the Global South.

What are the features of the AI system?

The design features and parameters that are baked into the development of algorithmic models (and that may emerge from them) can be sources of risk and harm. While a growing number of developers are introducing fairness testing and promoting algorithmic transparency, it is also the case that increasingly complex AI are still poorly understood and communicated. New laws and regulations are seeing to require more "explainability", but these are challenging, particularly for lay audiences.

As AI models become larger and more complex, they are also becoming more biased, according to Stanford University's AI Index Report 2022, which cited data showing large language models are showing greater propensity to reflect biases from their training data. This is particularly the case when algorithms are trained on data that is not representative of a wider population (over- or under-sampling from particular groups). The result can include bias and discrimination against protected classes, in particular.

Many developers make their AI systems black boxes either by design or default. In many cases, training data, inputs, and operations are opaque to users and researchers. And some AI models, including deep neural networks, are so complex that it is impossible to make sense of what the machine is doing – even for the developers who created it. Meghan O'Gieblyn notes that in cutting-edge AI systems, "If you were to print out everything the networks do between input and output, it would amount to billions of arithmetic operations, an 'explanation' that would be impossible to understand."

AI alignment refers to ensuring these systems pursue goals that are desirable and beneficial to human and societal goals. In the contexts of public safety and development, it is not enough for developers to consider the direct alignment problem – i.e. whether a system accomplishes the goals of the entity operating it – but they must consider social alignment, the effects of the AI system on society overall. Some AI systems may fulfill the goals of its operator, while at the same time generate harmful externalities for other groups in society.

Best Practices to Improve Design and Development of AI

A shortlist of principles are emerging to help address several of the challenges identified above, including issues of algorithmic bias, discrimination, opacity and alignment. A key is to proactively promote equal access and address adverse effects in advance. Igarape Institute and New America have identified over 100 distinct sets of standards, guidance and directives from across the public, private, and non-profit sectors that set out first principles. A representative sample of these will be visualized for the Global Task Force and wider public later in 2023.[4] A number of basic best practices are included below.

- **Fair.** Algorithms should be designed with fairness as an explicit goal. A recent paper from a team of DeepMind researchers argued that a way to ensure fairness was to adopt a "veil of ignorance" approach to selecting the principles that should govern an AI system. Coined by the philosopher John Rawls, the veil of ignorance refers to a situation where a person makes choices about the principles that should govern a system or society without knowing their relative position in that system or society beforehand – i.e. they make their choices from behind a "veil of ignorance". The researchers' study found that when a veil of ignorance was imposed, participants chose principles governing an AI assistant that prioritized the worst-off, and thus maximized for fairness.
- **Participatory.** The design of algorithmic models should be, to the extent possible, participatory, with input from diverse and interdisciplinary groups of experts. A particular focus must be on including those from the societies and sectors in which a particular AI tool will be used, and including policy makers, researchers, and civil society. This might require civic engagement, involving elements of the public and of public institutions that will use and be affected by the AI system.

4 See prototype at https://kumu.io/Igarape/ai-network-visualization#untitled-map/introducing-unitys-guiding-principles-for-ethical-ai-unityblog.

- **Representative.** The datasets used to train and test the model should be representative of the societies in which they will be used. What is more a diverse group of testers should test an AI system prior to its deployment. Finally, developers must evaluate the performance of an AI system in real-world scenarios across different subgroups, use cases, and geographic, cultural, and socioeconomic contexts. Monitoring and evaluation should be done regularly to identify emergent biases, risks, and unintended consequences.
- **Interpretable.** Communicating and enabling users to understand how an AI system generates its predictive output should be a priority for developers. Systems should be transparent about their capabilities, their inputs, and whose interests they represent – whether those of the user, the company that created it, or the government or public entity that commissioned it. Wherever possible, companies ought to provide clear explanations on how its technologies make decisions and generate predictions to users. This includes explaining what the technology is designed to do, how it was trained, and how it makes decisions and predictions. This is especially critical in public safety and law enforcement use cases, where system output can trigger life-and-death legal and judicial actions and processes.

  According to Google AI's responsible AI best practices, this means using the simplest model and smallest set of inputs necessary for performance goals; learning causal relationships instead of correlations; and designing the training objective to match the goal of the system. For highly complex models, disclosing the training data, mathematical model, and implementation won't amount to a sufficient explanation of the system's behavior. Instead, the system's decisions need to be observed across a variety of actual cases and in response to modifications – what-if explorations referred to as local interpretability, which systematically explores model output given changes in model input.

- **Accountable**. Steps should be taken to hold systems, their creators, developers, and users responsible for associated outcomes and actions. That implies that structures should be in place that enable society to hold those who design and develop AI technologies answerable for ethical, legal, and social implications of these systems. AI systems should be developed such that they are accountable to human oversight and control. Human judgment and ethical oversight should be obtained over the development of AI, such that responsibility for these systems is clearly in the hands of those who develop and oversee them.

  AI systems should be developed and designed in compliance with existing regulations, laws, standards, and ethical guidelines. Many jurisdictions and sectors have developed such codes – even if only voluntary at this point. AI developers should also adopt internal procedures to close the accountability gap – such as algorithmic auditing throughout the development life-cycle – but where possible trusted third-parties should be empowered empowered to review, red-team, and assess new systems.

Selected Resources

Chinmayi Arun, "AI and the Global South: Designing for Other Worlds," in Markus D. Dubber, Frank Pasquale, and Sunit Das (eds), The Oxford Handbook of Ethics of AI (2020; online edn, Oxford Academic, July 9, 2020).

Aspen Institute, "Building and Distributing Artificial Intelligence for Equitable Outcomes: A Blueprint for Equitable AI". Aspen Institute Science & Society Program. https://www.aspeninstitute.org/wp-content/uploads/2023/01/Equitable-AI-Aspen-Institute.pdf

IEE, "Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems," (version 2) The IEEE Global Initiative of Autonomous and Intelligent Systems, 2018.

Google, "Responsible AI Practices," Google AI, accessed May 21, 2023.

Anton Korinek and Avital Balwit, "Aligned with whom? Direct and social goals for AI systems," Brookings Institution Center on Regulation and Markets Working Paper, May 2022.

Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, TimnitGebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes, "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing," In Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27–30, 2020, Barcelona, Spain. ACM, New York, NY.

Daniel Zhang, Nestor Maslej, Erik Brynjolfsson, John Etchemendy, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Michael Sellitto, Ellie Sakhaee, Yoav Shoham, Jack Clark, and Raymond Perrault, "The AI Index 2022 Annual Report," AI Index Steering Committee, Stanford Institute for Human-Centered AI, Stanford University, March 2022.

Deployment and monitoring of predictive AI systems; risks and best practices

September, 2023

Due in part to the unprecedented spread of generative AI, governments, industry and digital rights groups are calling for greater investment in safety and alignment. While such issues are top of mind when it comes to designing and developing algorithms and large language models (LLMs), there is comparatively less focus on how to ensure greater accountability, explainability, equity, and non-discrimination when it comes to procuring, applying and evaluating AI systems. This is particularly so in parts of the Americas, Africa and Asia - the Global South - where the absence of regulations, low awareness and high costs are potentially larger factors driving decision-making. This technical note prepared for the Global Task Force considers a range of mechanisms to help ensure that responsible principles are considered at all stages of the AI procurement-deployment-monitoring cycle.[5]

Drawing from publicly available research, this note issues a number of reflections on how to improve safety and alignment at the deployment stage of AI. The first section considers several concerns and the second explores emerging best practices. It is important to stress that the note includes several biases, not least a reliance on studies produced in North America and Western Europe. Moreover, there is relatively limited evidence of outcomes of predictive AI systems in the Global South, a lacuna frequently noted in the literature. A key priority for the Task Force, then, is reviewing the relevance and applicability of these insights in ostensibly lower- and middle-income settings. Indeed, one recommendation for the final report may well be the need for additional empirical research on AI procurement, deployment and impacts in lower- and middle-income settings.

5 The Igarape Institute and New America produced a technical note on the design and development of predictive AI algorithms issuing a series of basic principles for good practice. See Igarape Institute and New America Foundation (2023) "Background Paper 2: Assessing risks and best practices in the design and development of AI systems". June.

Key findings include:

- **Information asymmetries related to the costs and consequences of predictive analytics can affect procurement decisions:** Transparent and inclusive procurement processes and robust testing are essential, especially in settings with limited data availability;
- **There is a need to anticipate and manage evolving data across settings, including planning for lower-income settings deploying predictive analytics systems:** More robust and evidence-based testing and auditing before and after deployment is urgently needed to avoid unintended outcomes;
- **Safeguards are required to limit the intentional and unintentional misuse of predictive analytics systems:** Reliable evaluation mechanisms and vendor data are lacking and continuous oversight, including auditing, independent certification, user training, accountability, and governance mechanisms could help curb or minimize misuse.
- **Rapidly evolving norms require continuous update of predictive tools to ensure compliance:** Legal framework evolving must faster in the US and EU without considering specific needs of other regions. Weather governments need to update and simplify regulation and vendors must catch up and adapt to frameworks for different realities and use cases.

Section I. Concerns related to algorithmic deployment and monitoring

Information asymmetries during the procurement process

The process by which predictive analytics tools are identified and procured by governments, companies, and nonprofits can become an important risk factor. This is because most purchasers are unable to assess and foresee the full potential costs and consequences of novel technologies, indeed many lack any formal or credible impact evaluations on which to base decisions. These risks are paramount in lower- and middle-income settings, with more limited access to information on technology suppliers and the outcomes of specific predictive analytics systems.

To minimize risks related information asymmetries, data sharing mechanisms and due diligence processes could be developed. These could at a minimum enhance transparency and embed AI ethics guidelines and principles during the tendering and procurement stages. Providing third party verification processes made up of a representative group of subject matter experts and potentially affected populations can also improve accountability, transparency and legitimacy of decision-making. Moreover, robust evidence-based testing of AI-powered analytic systems post deployment is essential, especially when deploying more high risk applications such as crime prediction.

Access to credible datasets in lower-income settings

Predictive AI systems typically reflect the underlying data and rules they are trained on. Even if developed according to the most robust recommendations highlighted in the previous technical note - including peer-reviewed algorithms, representative training data, rules that are designed under fair and participatory principles and allow for interpretability and accountability[6] - once deployed in the real-world, predictive systems will interact with evolving data from use cases and scenarios which may or may not have been envisioned during design and training phases.

The challenges of managing evolving data are particularly relevant when considering the needs of low income populations and vulnerable groups. For example, a credit model trained for a representative historical population of clients of a financial institution may be neither applicable nor fair for a different client base. What is more, AI systems developed in the US or Western Europe and trained on a limited amount of data and use cases are often also deployed in the Global South. These may however not always be appropriately trained and vetted for the specific reality and needs on the ground.

What is more, models may develop biases once deployed depending on the feedback data they receive while in use. The same may be true for a digital recruitment agent trained to screen resumes in a particular context and to receive and continuous feedback data on new hires to input into training its predictive model. Continuous oversight is essential as once the model is fed new data it may learn new patterns and behave in ways that were not observed during testing and training.

Risks of unintended or intended misuse and non-intended purposes

A recurring concern relates to controlling how predictive AI systems are used by users after they are procured. As the explosive debate around GPT has shown, a recommendation algorithm designed to help find the cure of a disease by developing a new drug or a new application for an existing drug, can also be used to recommend a drug (or genetic/synthetic sequence) for harmful purposes. Considering that AI design concerns and recommendations are context dependent, new applications of the same algorithm would require revisiting and addressing all the potential design issues from the lens of that specific application prior to deployment.

There are also concerns about the ways in which certain types of predictive algorithms are applied to intentionally or unintentionally discriminate against certain populations and protected categories. For example, a furore emerged in the US and Europe over the application of facial recognition and other biometric tools that facilitate racial profiling. These tools have the potential to improve agility and fairness by reducing human bias and inefficiencies[7], but they can reinforce inequality and structural discrimination in certain settings. Likewise, there is push back in North America and Western Europe against spatial and individual-based predictive analytics tools used for policing and criminal justice, and how they may intentionally or unintentionally reinforce biases. These challenges are potentially even more dramatic in non-democratic contexts.

---

6 See Igarape Institute and New America Foundation (2023).

7 See Aguirre, K., Badran, E. and R. Muggah (2019). Future crime: assessing twenty first century crime prediction", Igarape Institute, https://igarape.org.br/wp-content/uploads/2019/07/2019-07-12-NE_33_Future_Crime.pdf

The Igarapé Institute has accumulated significant experience in the use of predictive tools for crime prediction since developing its CrimeRadar app - a public-facing crime forecasting platform that evaluated relative crime frequencies in different locations and times of metropolitan Rio de Janeiro - in 2016. In a Strategic note published in 2019[8] to assess the state of crime prediction technologies noted the lack of clear evidence and called for more robust evaluations: "*There is still comparatively mixed evidence of the accuracy of crime prediction, its impact on clearance rates, whether it improves response times or even leads to significant reductions in crime. The only way to really gauge the impacts of crime forecasting is to conduct statistical evaluations that isolate the effects of the measure, including randomized control trials (RCTs).*"

Indeed, in recent years there has been significant negative reaction against predictive policing technologies due in large part to the way it is perceived to reinforce biases.[9] While many new crime forecasting technologies and solutions have been developed, robust evaluation mechanisms and regulation are still lacking.[10] There is lively debate about how efficient these solutions can be at improving safety and reducing crime rates while ensuring accuracy, fairness, and transparency of decisions. Given this context, policymakers and program managers can demand more reliable impact evaluation mechanisms, including RCTs to better assess the risks and benefits of each technology.

Governments can also introduce requirements that AI vendors provide better evidence of outcomes.[11] Regulation can help close this gap by requiring the disclosure of more detailed and transparent data from vendors, and demand independent certifications for specific use cases.

Rapidly evolving norms and rules

A predictive analytics system is typically developed for a particular use case for which it is expected to be compliant with regulation and privacy norms. Yet when applied in different contexts, these same platforms may be infringing laws. Indeed, the norms, rules and principles themselves can change as legal opinion and societal views evolve.
One example of this is an AI system developed to assist doctors with identifying and recommending a specific drug. Depending on regulatory regimes, the same system could be considered legal and safe or illegal and unsafe for direct patient use without a doctor in the loop. In most democratic countries, it is increasingly expected that AI-enabled systems are routinely monitored to avoid potentially exposing sensitive data or infringing data protection laws, while these issues are essentially ignored in autocratic regimes.

Nevertheless, regulations are constantly evolving to meet the needs of society and predictive tools should also be updated to comply with new rules. As shown in a new visualization[12] developed by the Igarape Institute and New America, over 1,000 AI ethics guidelines, agreements and voluntary commitments have been issued over the past decade to address both the design and development of predictive analytics and the deployment and monitoring stages of these tools.

8 Ibid

9 See Cumming-Bruce, N. (2020) "U.N. Panel: Technology in Policing Can Reinforce Racial Bias". New York Times. November, https://www.nytimes.com/2020/11/26/us/un-panel-technology-in-policing-can-reinforce-racial-bias.html

10 See Verma, P (2022) "The never-ending quest to predict crime using AI". Washington Post. July, https://www.washingtonpost.com/technology/2022/07/15/predictive-policing-algorithms-fail/

11 See Aguirre, K., Badran, E. and R. Muggah (2019) "Future crime: assessing twenty first century crime prediction", July 2019.

12 https://kumu.io/Igarape/ai-network-visualization#initiatives/ethics-guidelines-for-trustworthy-ai

However, the vast majority of initiatives and guidelines issued to date originated from the US and EU most likely do not consider the particularities of the impacts and risks of predictive analytics for the rest of the world.[13] Moreover, the fact the laws are evolving slower and less rigorously enforced in many lower- and middle-income settings can lead to challenges for competition and compliance. Vendors will try to adapt applications to different regulatory settings and may potentially choose to take advantage of having less oversight in countries where regulation is still pending. Alternatively some vendors may decide to stay clear from countries where regulatory risk is unclear.

Section II. Best practices to improve monitoring and evaluation of AI

There is a general consensus that the deployment and monitoring of predictive analytics systems should be accompanied by fundamental principles of fairness, transparency, participatory design, representation, accountability and interpretability[14]. Decisions about the acquisition, application and evaluation of predictive analytic tools should be accompanied with ongoing oversight to ensure adherence to basic safety and alignment principles. For example, AI systems that are ranked as more "high risk" in terms of intended or unintended consequences should be subject to periodic internal and external assessments by a representative group of stakeholders - and even by external auditors - to ensure data and outcomes remain fair, interpretable, representative and that decisions have clear accountability. Several recommendations stand out when it comes to promoting oversight over predictive analytics systems after they are deployed.

Apply transparent and inclusive procurement processes

A basic condition of procurement and deployment should be that ethical AI principles are adopted across all stages of the process. For example, there should be individuals in decision-making authority that are literate in AI in the selection, vetting and decision-making activities associated with AI acquisition. There should be a call for minimum ethical AI principles in the request for proposals and any technology that does not account for these basic standards will not advance in the bidding process.

When it comes to procurement processes in lower- and middle-income settings, knowledge and capacity in AI and emerging technologies are more limited. Specific recommendations include: (I) including AI impact assessments in the procurement process and (II) ensuring the AI procurement process is transparent to the public; and (III) allowing for the opportunity for civic review. These requirements may require additional funding and capacity considerations, particularly in lower- and middle-income settings. To help fill these gaps, it may be advisable to establish an AI procurement knowledge-sharing network or hub to help governments better identify safe and high-quality tools to procure. Additionally governments must plan for training professionals in AI domains and partnering with institutes and foundations that can help assess AI tools.

13 See Muggah, R. (2023) AI will entrench global inequality, Foreign Policy, May 29, https://foreignpolicy.com/2023/05/29/ai-regulation-global-south-artificial-intelligence//

14 See Igarape Institute and New America Foundation (2023).

Iterative and representative testing and training

The most common best practice for responsible AI post deployment is to test and train systems continuously and frequently in many different ways and with representative feedback loops. Google's recommended practices for AI include a specific item labeled "test, test, test", recommending testing components in isolation and in their interaction, frequently testing with different use cases, users and datasets, incorporating diverse sets of user needs, and building quality checks.[15] Adversarial testing is also recommended to systematically evaluate models and expose potential failures or inaccuracies. Experts recommend having mechanisms to facilitate third party testing of vulnerability and external feedback.

Another common recommendation is to have humans in the decision loops of AI systems, especially for applications considered more risky. The EU AI Act recommends human oversight for high-risk AI Systems.[16] Likewise, Google's recommended practices for AI suggest "incorporating human feedback before and throughout project development".[17] Public authorities in the Global South can potentially develop a similar risk-adjusted framework but adapted for regional particularities. Depending on local data literacy and availability, the same use case may be considered higher-risk in a particular country and should therefore be subject to greater oversight.

Responsible data training is a consideration, though subject to limitations advance by Bender, Gebru, et al (2021)[18] which may act as counter incentives including the high compute costs of training systems for several frames and scenarios."[19] The computing costs involved in responsibly training and developing AI can become prohibitive in some settings, especially in the Global South where energy and infrastructure costs are already high. In such cases, governments must be prepared to weigh risks and benefits of each AI application in determining the magnitude of training that will be mandatory (how many scenarios/frames to plan for).

Expand third party audits

Experts recommend regular reviews such as third party audits to monitor several AI systems for discrimination and biases and other potential harms. Ethical reviews and impact assessments should be conducted by interdisciplinary committees involving experts in the social sciences, ethics, law, technology, and relevant domains to guide responsible deployment and monitoring. More importantly for the purposes of the Task Force, oversight should be conducted by a diverse group of stakeholders, and representative of all affected populations.

---

15 See Google (2023) "Responsible AI Practices," Google AI, accessed May 21, 2023.

16 See European Commission (2023) "Regulatory framework proposal on artificial intelligence". Accessed August 31 2023.

17 See Google (2023).

18 See Bender, E., Gebru, T.,McMillan-Major, M., and S. Mitchell (2021) "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?", https://dl.acm.org/doi/pdf/10.1145/3442188.3445922.

19 "Developing and shifting frames stand to be learned in incomplete ways or lost in the big-ness of data used to train large LMs — particularly if the training data isn't continually updated. Given the compute costs alone of training large LMs, it likely isn't feasible for even large corporations to fully retrain them frequently enough. See Bender et al (2021).

Given shortages of expertise and capacity, it is recommended that governments find ways to mandate and incentivize third party audits. One way to start is to develop a ranking of more risky applications for which external certifications are always required. Models that require above a certain threshold or quantity of computing power to train can be subject to additional transparency obligations, as was proposed by California lawmakers recently.[20]

Indeed, public authorities and or private actors should also conduct minimum auditing that is aligned with financial capabilities. Alternatively, financing could be pursued from multilateral and bilateral sources to facilitate required or voluntary audits. Of course, policy, legal, economic and ethical concerns must be carefully balanced.

Improve transparency and interpretability

Users should be able to understand how a predictive analytics AI system generates outputs. Interpretability should be considered a core part of user experience, a factor that many technology and social media platforms now recognize (though do not necessarily practice).[21] This will allow for better user feedback and testing. In this case, public authorities will need to introduce regulations and standards that require vendors to ensure products include minimum transparency and interpretability, including disclose the potential use of generative AI in any application. The EU AI Act is expected to provide a range of recommendations in this regard.[22] However, in lower- and middle-income contexts, this may require additional layers of action, including training for low literate populations. Vendors must be responsible for introducing and enforcing transparency mechanisms that meet their users' capacity to fully understand the consequences of AI tools thtat impact them.

Another important aspect of transparency in the deployment stage is the disclosure of capabilities and limitations of systems. Stakeholders should be informed of the extent to which systems have been tested for each use case and can be employed responsibly and be expected to function as designed for a specific task. Additionally, both the draft EU AI Act[23] and the White House briefing on voluntary AI commitments[24], require appropriate disclosure to users of when any content is generated by AI so that they know they are dealing with an AI system (and not with another human).

Ongoing education and user training

Ongoing awareness and education with users is essential as predictive analytic systems evolve. Graham (2023) observes that "transparency is not a quick fix" and "will not necessarily lead to safer products and services. Indeed, “consumers do not always have the ability to monitor and understand information that becomes available and often do not have a real choice to decline a service or migrate to a different provider". Transparency will only limit risks of AI systems if users are able to easily understand potential intended and unintended consequences, choose to opt out when they wish, and make decisions about the risks they are willing to incur.

---

20 See Perigo, P. (2023). "Exclusive: California Bill Proposes Regulating AI at State Level". https://time.com/6313588/california-ai-regulation-bill/

21 See Google (2023).

22 See EU (2023) https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

23 See European Commission (2023) "Regulatory framework proposal on artificial intelligence". https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

24 See White House briefing (2023) "Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI", July.

Many recent ethical AI guidelines prioritize training users and the general public on potential harms of AI systems.[25] Once users have access to transparent information and better understand predictive systems, they will in principle be better equipped to monitor, question and provide greater oversight. To wit, the Aspen Institute's Blueprint for Equitable AI[26] considers AI education a "prerequisite to equitable AI" and recommends that educational undertakings begin "with an assessment of current AI literacy and the areas where citizens could most benefit from greater understanding or education". This is especially important in countries with lower levels of AI literacy and those face basic challenges in their education systems. In addition to adapting school systems to incorporate appropriate knowledge to prepare students to interact and assess AI systems, countries will need to consider new retraining programs for citizens who are no longer in school and with limited access to education.

Strengthen AI governance, bolster norms and reduce externalities

Social alignment includes systems that are consistent not only with the objectives of the operator, but with broader, societal goals. It requires continuous monitoring of positive and negative externalities that are generated by AI in society writ large. Such externalities "exist whenever an AI system affects others without their agreement and without the beneficiary compensating others for it"[27]. It can be a challenge to address social alignment when social preferences are not well defined and agreed upon. Well crafted impact assessments can help users be more aware of the risks and benefits to society. Policy-makers can then assess the need to develop new legislation or social norms to address new forms of social misalignment that may emerge as technology develops.

Given the challenges of updating regulation to keep up with evolving technology, the IEEE Global Initiative recommends the development of multi-stakeholder ecosystems to create norms and eventually best practices and/or laws where they do not yet exist because a specific AI technology and its impact is too new.[28] Vendors in particular must bear some of the burden of ensuring comprehensive mechanisms are in place to ensure continued compliance with existing and new norms in each society. AI systems should also have accountability embedded in a way that it is always clear to users and regulators who is legally responsible for potential unforeseen harms.

A challenge in the Global South is how to ensure complex and evolving legal systems do not deter vendors from making products and services available in countries where navigating regulation is too costly. Governments in lower- and middle-income settings must in turn invest and potentially partner with multilaterals and other countries in the region to make AI regulation and potential consequences as clear and predictable as possible.

---

25 Including, for example, UNESCO's Recommendation on the Ethics of Artificial Intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000380455 and Aspen Institute's Blueprint for Equitable AI https://www.aspeninstitute.org/wp-content/uploads/2023/01/Equitable-AI-Aspen-Institute.pdf

26 See Aspen Institute, "Building and Distributing Artificial Intelligence for Equitable Outcomes: A Blueprint for Equitable AI". Aspen Institute Science & Society Program.

27 See Korinek, A. and Avital Balwit, A. (2022) "Aligned with whom? Direct and social goals for AI systems," Brookings Institution Center on Regulation and Markets Working Paper, May.

28 See IEEE (2018) "Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems," (version 2) The IEEE Global Initiative of Autonomous and Intelligent Systems.

Closing Reflections

There are a host of concerns when considering the deployment and monitoring of predictive AI systems. These include (i) information asymmetries that may impact procurement decisions, (ii) evolving data and use cases that may not applicable across settings leading to unforeseen outcomes, (iii) intentional and unintentional misuse of AI Systems and (iv) AI norms that are being updated rapidly in the EU and US, but still lagging in the Global South.

Several recommendations are emerging, largely originating from the EU and US, but that can be tailored to the Global South. These include transparent and inclusive procurement processes, robust evidence-based testing and evaluations, and continuous oversight - including ongoing auditing, independent certifications and user training. Wealthier governments need to update and simplify regulation and train citizens to be able to identify potential risks. Vendors must be held accountable for keeping up with regulation and local norms, but also for ensuring target users are able to interpret and understand risks associated with their systems.

Selected Resources

Aspen Institute (2023) "Building and Distributing Artificial Intelligence for Equitable Outcomes: A Blueprint for Equitable AI". Aspen Institute Science & Society Program.

Association for Computing Machinery (2023) "World's Largest Association of Computing Professionals Issues Principles for Generative AI Technologies", July.

Kathy Baxter and Yoav Schlesinger (2023) "Managing the Risks of Generative AI", Harvard Business Review, June.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, Shmargaret Shmitchell (2021) "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" March.

Rumman Chowdhury (2023) "A New Era in AI Governance". August.

Nick Cumming-Bruce (2023) "U.N. Panel: Technology in Policing Can Reinforce Racial Bias". New York Times. November, https://www.nytimes.com/2020/11/26/us/un-panel-technology-in-policing-can-reinforce-racial-bias.html

European Commission (2023) "Regulatory framework proposal on artificial intelligence". Accessed August.

Google (2023) "Responsible AI Practices," Google AI.

Mary Graham (2023) "Disclosure Dilemmas: AI Transparency is No Quick Fix", Ash Center for Democratic Governance, August.

IEEE (2018) "Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems," (version 2) The IEEE Global Initiative of Autonomous and Intelligent Systems.

Anton Korinek and Avital Balwit (2022) "Aligned with whom? Direct and social goals for AI systems," Brookings Institution Center on Regulation and Markets Working Paper, May.

Billy Perigo (2023) "Exclusive: California Bill Proposes Regulating AI at State Level". https://time.com/6313588/california-ai-regulation-bill/

Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, TimnitGebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes (2020) "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing," In Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27–30, Barcelona, Spain. ACM, New York, NY.

Pranshu Verma (2022) "The never-ending quest to predict crime using AI". Washington Post. July. https://www.washingtonpost.com/technology/2022/07/15/predictive-policing-algorithms-fail/

White House briefing (2023) "Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI", July.

IGARAPÉ INSTITUTE
a think and do tank

The Igarapé Institute is an independent think and do tank committed to citizen, digital, and climate security. Founded in 2010 and based in Brazil, the Institute designs and delivers data-driven and evidence-based solutions to address global challenges. The Institute oversees activities in over 20 countries across the Americas, Africa and Europe leveraging a combination of primary research, technology innovation, strategic partnerships, and global advocacy and communications. Ranked among the top social policy and environment think tanks in the world and a leading non-profit in Brazil, the Institute works with governments, the private sector, and civil society to deliver lasting social impact.

See www.igarape.org.br

NEW AMERICA

New America is a think tank and civic enterprise dedicated to renewing, reimagining and realizing the Promise of America in an era of rapid technological and social change. Since 1999, New America has nurtured a new generation of policy experts and intellectuals and pioneered bold and successful policy initiatives in areas ranging from education to open technology to political reform. As the nation's premier idea incubator and accelerator, New America has the unique capacity to quickly innovate new programs and scale existing programs that address the most relevant problems in our society — today, and in decades to come.

See www.newamerica.org

www.igarape.org.br

IGARAPÉ INSTITUTE
a think and do tank