INFOCUS - RISK-BASED APPROACHES TO ARTIFICIAL INTELLIGENCE (AI) GOVERNANCE

Recent advancements in AI technologies leading to new commercial applications with potentially adverse social implications: the way forward.

Over the last five years, 117 initiatives worldwide have published Artificial Intelligence (AI) ethics principles. Despite a skewed geographical scope (91 of these initiatives come from Europe and North America), the proliferation of such initiatives on AI ethics principles is paving the way for building global consensus on AI governance. Notably, the 38 OECD Member States have adopted the OECD AI Recommendation, the G20 has endorsed these principles, and the Global Partnership on AI is operationalising them. UNESCO is furthermore developing a Recommendation on the Ethics of AI that 193 countries may adopt in 2021.

An analysis of different principles revealed a high-level consensus around eight themes: (1) privacy, (2) accountability, (3) safety and security, (4) transparency and explainability, (5) fairness and non-discrimination, (6) human control of technology, (7) professional responsibility, and (8) the promotion of human values. However, these ethical principles are criticised for lacking enforcement mechanisms. Companies often commit to AI ethics principles to improve their public image yet give little follow-up on implementing them, an exercise termed as «ethics washing". Evidence also suggests that knowledge of ethical tenets has little or no effect on whether software engineers factor ethics into the development of their products or services. Defining principles is certainly essential, but it is only a first step to developing ethical AI governance. There is a need for mid-level norms, standards and guidelines at the international level that may inform regional or national regulation to translate principles into practice. This article discusses the need for AI governance to evolve past the "ethics formation" stage by implementing concrete and tangible steps, such as developing technical benchmarks and adopting risk-based regulation.

Recent Advances in AI Technologies

Artificial Intelligence is developing rapidly. The 2021 AI Index report notes four crucial technical advances that hastened the commercialisation of AI technologies:

AI-Generated Content: AI systems can generate high-quality text, audio and visual content to a level that it is

difficult for humans to distinguish between synthetic and non-synthetic content.
Image Processing: Computer vision has seen immense progress in the past decade and is fast industrialising in applications that include autonomous vehicles.
Language Processing: Natural Language Processing (NLP) has advanced such that AI systems with language capabilities now have economic value through live translations, captioning, and virtual voice assistants.
Healthcare and biology: DeepMind's AlphaFold solved the decades-old protein folding problem using machine learning techniques.
These technological advances have social implications as well as economic value. For instance, the technology generating synthetic faces has rapidly improved. As shown in Figure 1, in 2014, AI systems produced grainy faces, but by 2017, they were generating realistic synthetic faces. Such AI systems have led to the proliferation of 'deepfake' pornography that overwhelmingly targets women and has the potential to erode people's trust in the information and videos they encounter online. Some actors misuse the deepfake technology to spread online disinformation, resulting in adverse implications for democracy and political stability. Such developments have made AI governance a pressing matter.

Challenges of AI Governance

These rapid advancements in the field of AI technologies have brought the need for better governance to the forefront. In thinking about AI governance, any governments worldwide are concerned with enacting regulation that does not stifle innovation yet also provides adequate safeguards to protect human rights and fundamental freedoms.
Technology regulation is complicated because, until a technology has been extensively developed and widely used, its impact on society is difficult to predict. However, once a technology is deeply entrenched and its effect on society is understood better, it becomes more challenging to regulate. This tension between providing free and unimpeded technology development while regulating adverse implications is termed the "Collingridge dilemma".
David Collingridge, the author of the Social Control of Technologies, notes that when regulatory decisions have to be made before a technology's social impact is known, continuous monitoring can help mitigate unexpected consequences. Collingridge's guidelines for decision-making under ignorance can inform AI governance as well. These include choosing technology options with (1) low costs of failure, (2) short response times for responding to unanticipated problems, (3) low costs of remedying unintended errors, and (4) cost-effective and efficient monitoring.

Technical benchmarks for evaluating AI systems

Quantitative benchmarks are also necessary to address the ethical problems related to bias, discrimination, lack of transparency, and accountability in algorithmic decision-making. The Institute of Electrical and Electronics Engineers (IEEE), through its Global Initiative on Ethics of Autonomous and Intelligent Systems, is developing technical standards to address bias in AI systems. Similarly, in the United States, the National Institute of Standards and Technology (NIST) is developing standards for explainable AI based on principles that call for AI systems to provide reasons for their outputs in a manner that is understandable to individual users, explain the process used for generating the output, and deliver their decision only when the AI system is fully confident.
Going back to our previous example, there is already significant progress in introducing benchmarks for the regulation of facial recognition technology. Facial recognition systems have a large commercial market. They are used for various tasks, including law enforcement and border controls. These tasks involve detecting visa photos, matching photos in criminal databases, and detecting and removing child abuse images online. However, facial recognition systems have been the cause of significant concern due to high error rates in detecting faces and impinging on human rights. Biases in such systems have adverse consequences for individuals, such as being denied entry at borders or being wrongfully incarcerated. In the United States, the NIST's Face Recognition Vendor Test provides a benchmark to compare different commercially available facial recognition systems› performances by operating their algorithms on different image datasets.
Defining benchmarks for ethical principles is an important step, however, in line with the Collingridge

Dilemma, it needs to be complemented by risk assessments to mitigate adverse social impacts. Risk assessments would allow for the application of risk-proportionate AI regulation instead of a reliance on blanket rules that may hinder technological development with unnecessary compliance burdens. The next blog post in this two-part series will engage with some potential risk-based approaches to AI regulation.

AI Risk Assessment Frameworks

Risk assessments can help identify which AI systems need to be regulated. Risk is determined by the severity of the impact of a problem and the probability of its occurrence. For example, the risk profile of a facial recognition system to unlock a personal mobile phone would differ from a facial recognition system used by law enforcement. The former may be overall beneficial as it adds a privacy-protecting security feature. In contrast, the latter could have chilling implications on the freedom of expression and privacy. Therefore, the risk score for facial recognition systems is relative to their use and deployment context. The following are some of the approaches followed by various bodies in developing risk assessment frameworks for AI systems.

The European Union (EU)

The European Commission's legislative proposal on Artificial Intelligence classifies AI systems by four levels of risk and outlines risk proportionate regulatory requirements. The categories proposed by the EU include:

- Unacceptable Risk: The EC has proposed a ban on applications like social credit scoring systems and real-time remote facial recognition systems in public spaces.
- High Risk: AI systems that harm the safety or fundamental rights of people are categorised as high-risk. The proposal prescribes some mandatory requirements for high-risk AI systems.
- Limited Risk: When the risks associated with the AI systems are limited, only transparency requirements are prescribed.
- Minimal Risk: When the risk level is identified as minimal, there are no mandatory requirements, but the developers of such AI systems may voluntarily choose to follow industry standards.

Germany

In Germany, the Data Ethics Commission has proposed a five-layer criticality pyramid that requires no regulation at a low-risk level to a complete ban at high-risk levels (see Figure 2). The EU approach is similar to the German approach but differs in the number of levels.

The UK

The AI Barometer Report of the Centre for Data Ethics and Innovation identifies some common risks associated with AI systems and some sector-specific risks.

The common risks include:

- Algorithmic bias and discrimination
- Lack of explainability of AI systems

- Regulatory capacity of the State
- Breach in data privacy due to failure in user consent
- Loss of public trust in institutions due to problematic AI and data use

The report identified that the severity of common risks varies across different sectors like criminal justice, financial services, health and social care; digital and social media; and energy and utilities. For example, algorithmic bias leading to discrimination is considered high-risk in criminal justice, financial services, health and social media but medium risk in energy and utilities. The risk assignment, in this case, was done through expert discussions. The UK's approach has a strong sector specific focus. The overall sector level risk is ascertained based on a combination of multiple AI risk criteria.

The Organisation for Economic Co-operation and Development (OECD)

The preliminary classification of AI systems developed by the OECD Network of Experts' working group on AI classification has four dimensions:

-
- Context includes stakeholders that deploy an AI system, the stakeholders impacted by its use and the sector in which an AI system is deployed.
- Data and inputs to an AI system influence the system›s outputs based on the data classifiers used, the source of the data, its structure, scale, and how it was collected.
- The type of algorithms used in AI systems has implications for transparency, explainability, autonomy and privacy.
- The kind of task to be performed and the type of output expected range from forecasting, content personalisation to detection and recognition of voice or images.

Applying this classification framework to different cases, from facial recognition systems and medical devices to autonomous vehicles, allows us to understand the risks under each dimension and design appropriate regulation. In autonomous vehicles, the context of transportation and its significant risk of accidents increase the risk associated with its AI systems, and they are therefore considered a high-risk category requiring robust regulatory oversight.

Next steps in Risk-Adaptive Regulation for AI

The four approaches to risk assessment discussed above are systematic attempts to understand AI-related risks and develop a foundation for downstream regulation that can address risks without being overly prescriptive. With these examples in mind, national level initiatives could improve their AI governance by focusing on the following:

- AI Benchmarking: AI systems need continuous development and updating of technical benchmarks to assess their performance under different contexts with respect to AI ethics principles.
- Risk Assessments of AI applications: Risk assessments of AI systems require development of use cases of different AI applications under different combinations of contexts, data and inputs, AI models and outputs.
- Systemic Risk Assessments: There is a need for systemic risk assessment in contexts where AI systems interact with one another. For example, in financial markets, different AI algorithms interact with each other, and in certain situations, their interactions could cascade into a market crash.

Once AI risks are better understood, proportional regulatory approaches should be developed and subjected to Regulatory Impact Analysis (RIA). The OECD defines RIA as a "systemic approach to critically assessing the positive and negative effects of proposed and existing regulations and non-regulatory alternatives". RIAs can guide governments in understanding if the proposed regulations are effective and efficient in achieving the desired objective. Such impact assessments are good regulatory practice and will become increasingly relevant as more countries work towards developing their own national AI legislations.

Given the globalised nature of different AI services and products, countries should also develop their national level regulatory approaches to AI in conversation with one another. Importantly, these dialogues at the global and national level must be multistakeholder driven to ensure that different perspectives inform any ensuing regulation. Collectivised knowledge and coordination will lead to overall benefits by ensuring AI develops in a manner that is both ethically aligned and provides a stable environment for innovation and interoperability.

Prateek Sibal is a PhD Candidate in Governance and studies patterns of misinformation and disinformation on social media platforms in the context of political polarization. He is affiliated with the Centre for Digital Governance. He worked at the United Nations Educational, Scientific and Cultural Organisation's (UNESCO) on artificial intelligence and public policies for more than three years and is a lecturer on digital governance at the School of Public Affairs, Sciences Po, Paris. In the past he worked as a legislative analyst to a Member of Parliament in India. He has a Bachelor in Mechanical Engineering from VIT University, India and a Master in Public Policy (summa cum laude) from Sciences Po, Paris. His research interests include online communication, platform regulation, political polarization, media economics and digital governance.
This is a revised version of a post first published on the Centre for Communication Governance, National Law University Delhi's blog. Its content is an outcome of ongoing research at the Centre for Communication Governance on AI and emerging tech.
The author would like to thank Jhalak Kakkar, Nidhi Singh and Moana Packo for their helpful feedback.
This is a revised version of a post first published on the Centre for Communication Governance, National Law University Delhi's blog. Its content is an outcome of ongoing research at the Centre for Communication Governance on AI and emerging tech. The author would like to thank Jhalak Kakkar, Nidhi Singh and Moana Packo for their helpful feedback.
UNESCO