ALLEN & OVERY

June 2019

AI-Principles and Practice

The OECD Recommendation Continues a Theme

EDITORIAL

At the end of May, the Organisation for Economic Co-operation and Development (**OECD**) member countries achieved a first – signing up to a **Recommendation on Artificial Intelligence (AI)** (the **Recommendation**) that, it is hoped, will promote innovative and trustworthy AI, respecting human rights and democratic values. As AI is such a hot topic, non-member countries such as Argentina, Brazil, Colombia, Costa Rica, Peru and Romania have already agreed to adhere to the Recommendation, with others expected and welcomed.

Emma Keeling
PSL
Corporate

Contact
Tel +44 20 3088 2182
emma.keeling@allenovery.com

The Principles

Those adhering countries are encouraged to promote and implement ***a set of Principles for trustworthy AI***, with organisations and individuals active in deploying and operating AI called on to do the same.

The Principles themselves are high level, intended to provide practical and flexible standards that can cope, as far as possible, with the ever changing world of AI. Aiming for trustworthy AI, they can be summarised as:

1. AI should benefit people and the planet by driving inclusive growth, sustainable development and well-being.

2. AI systems should be designed in a way that respects the rule of law, human rights, democratic values and diversity, and they should include appropriate safeguards – for example, enabling human intervention where necessary – to ensure a fair and just society.

3. There should be transparency and responsible disclosure around AI systems to ensure that people understand AI-based outcomes and can challenge them.

4. AI systems must function in a robust, secure and safe way throughout their life cycles and potential risks should be continually assessed and managed.

5. Organisations and individuals developing, deploying or operating AI systems should be held accountable for their proper functioning in line with the above principles.

Context and alignment with other guidelines

These Principles are not formulated in a vacuum and are intended to complement the OECD's existing standards regarding privacy, digital security risk management, and responsible business conduct.

The Principles also reflect a similar approach to that of the **Ethics Guidelines for Trustworthy AI** produced by the European Commission established Independent High-Level Expert Group on Artificial Intelligence in April 2019 (the **Ethics Guidelines**). The non-binding and voluntary Ethics Guidelines describe the desired "trustworthy AI" as being: lawful (see also Principle 2 above); ethical – ie respecting principles of respect for human autonomy, prevention of harm, fairness and explicability, based on the Charter of Fundamental Rights of the EU (see also Principle 1 and 2 above); and robust (see also Principle 3 above).

The more specific requirements of the Ethics Guidelines to aid "realisation" of Trustworthy AI also align with the OECD Principles, for example: ensure human agency and oversight (see also Principle 2 above); ensure technical robustness and safety (see also Principle 4 above); ensure privacy and data governance (see also Principle 2 above); ensure transparency (including traceability and clear communication – see also Principle 3 above); ensure diversity, non-discrimination and fairness (see also Principle 2 above); ensure

environmental and societal well-being (see also Principle 1 above); and ensure accountability (including auditability – see also Principles 4 and 5 above).

It is also clear to see how these ***principles chime with the increasing focus on the ethical, socially responsible approach to business models and technology***. Indeed, when looking to hit ESG (Environmental, Social and Governance) targets or live by value statements, organisations will need to be comfortable that their operations (AI-related or otherwise) are in-step with, and further, those aims.

Practical implementation

However, the ***integration of such high-level values into day to day practice will take further work***. The realities of the technology and the extent of AI expertise means that investment is required to reach a state where AI use can be widespread yet bias is avoided and systems are fully transparent, for example. How one determines whether AI drives inclusive growth, sustainable development and well-being (Principle 1) in a way that organisations can be held accountable (Principle 5) is certainly not clear at this stage.

That said, the Recommendation instructs the OECD's Committee on Digital Economy Policy to ***develop and iterate further practical guidance on implementation*** and report by the end of the year.

In addition, the **OECD's AI Policy Observatory**, an online hub for AI information, evidence and policy, is expected to be launched later this year. A particular focus will be to provide evidence and guidance on AI metrics, policies and practice to help implement the Principles, including the development of metrics to measure AI research, development and deployment and assess progress. Cross-border and sector cooperation, of increasing importance in this digital, data-rich age of IOT and connectivity, will also be facilitated by the Observatory, enabling international benchmarking, access to policy communities and a wide range of stakeholders (technical, private sector, academia, civil society etc).

Whilst the Ethics Guidelines reference similar goals to those of the Recommendation, they also provide input on how to operationalise principles in practice and so could act as a useful reference point and first step to more formalised guidance. The Ethics Guidelines encourage use of both technical and non-technical methods to meet Trustworthy AI principles for example: setting up internal and external governance frameworks including mechanisms such as human-in-the-loop, human-on-the-loop, or human-in-command approach; enabling public enforcers to exercise oversight; incorporating Trustworthy AI goals into organisation codes of practice; carrying out a fundamental rights impact assessment before AI system development; ensuring data collection is lawful and not discriminatory; implementing data access protocols; assessing energy consumption and resource usage of the system; using AI systems that meet industry standards, accreditation schemes or provisional codes; engaging stakeholders in panel discussions; ensuring diversity of developers; creating "white lists" and "black lists" of acceptable behaviours and states of the AI systems; using privacy and security by-design measures; implementing mechanisms to shut down or automatically re-start the AI system following an attack; monitoring and testing the AI models before, during and after deployment for stability, robustness and operation; using a diverse group of testers, "red teams" to "break" the AI system and "bug bounties" to reveal vulnerabilities; and using "explainable AI" research to explain behaviours of the AI systems.

Recommendations for governments

The OECD Recommendation also addresses ***National policies and international cooperation with suggestions for governments*** in light of the Principles that they:

1. Facilitate public and private investment in research & development to spur innovation in trustworthy AI.

2. Foster accessible AI ecosystems with digital infrastructure, technologies and mechanisms to share data and knowledge.

3. Ensure a policy environment that will open the way to deployment of trustworthy AI systems.

4. Empower people with the skills for AI and support workers for a fair transition (ie as jobs are replaced with technology).

5. Co-operate across borders and sectors to progress on responsible stewardship of trustworthy AI.

These mirror the European Commission's three pillars for its AI vision ie: increasing public and private investment in AI to boost its uptake and the EU's technological and industrial capacity; preparing for socio-economic changes; and ensuring an appropriate ethical and legal framework to strengthen European values (as included in the European Commission's Communication on Artificial Intelligence for Europe (COM(2018)237) and further in the Coordinated Plan (COM(2018)795)).

The Ethics Guidelines similarly encourage stakeholders to foster research and innovation to assess AI systems (see also recommendation 1 above); disseminate results and open questions to the wider public (see also recommendation 2 above); systematically train a new generation of experts (see also recommendation 2 and 3 above); involve stakeholders throughout AI lifecycle, fostering training and education so that all are aware of and trained in Trustworthy AI (see also recommendation 4 above).

UK actions

The UK government is certainly taking steps to address some of these recommendations already, pursuing its **AI Sector Deal** by: forming the **Office for Artificial Intelligence** to deliver the AI Sector Deal; appointing the new **AI Council** (with leaders from business, academia and data privacy organisations to promote adoption and ethical use of AI in organisations); establishing the **Centre for Data Ethics and Innovation** (expert advice on the measures for safe, ethical and innovative uses of AI and data-driven technologies); announcing an AI skills and talent package (including AI Fellowships and funding to attract and retain top AI talent); agreeing new centres of excellence for digital pathology and imaging, including using AI medical advances; announcing new research projects considering AI application in the legal and accountancy sectors; and partnering with the Open Data Institute to explore data trusts, tackling illegal wildlife trade and reducing food waste.

Against the backdrop of Brexit and with international conflict playing out in the development and roll out of new technology, the UK has clearly decided to pursue a proactive approach to AI with a view to becoming a "global leader in this technology that will change all our lives".

Next steps and actions to watch

However, the UK is not alone in its ambitions and governments and international organisations are all continuing to formulate positions and approaches, guides and frameworks in relation to AI.

Indeed, the OECD AI principles will be discussed at the G20's June summit in Japan, the OECD's Committee on Digital Economy Policy will develop and iterate further practical guidance on implementation and report by the end of the year and the Ethics Guidelines' pilot Trustworthy AI assessment list is open for a two-limb consultation and review before revision in early 2020.

The World Economic Forum (**WEF**) is also looking to build a more international, cross-cultural consensus around the use and governance of AI. At an event on the topic at the end of May, the WEF announced the creation of an AI Council, co-chaired by Microsoft president, Brad Smith, and Sinovation Ventures CEO, Dr Kai-Fu Lee, and formed of members as diverse as UK government ministers, IBM and Future of Life Institute. The Council is intended to identify the key issues in AI and how to "bridge governance gaps".

On a national level, in light of its aims to become the world leader in AI innovation and in the context of its New Generation AI Development Plan (2017), China is currently working on its own governance principles with AI experts for example.

Businesses will be hoping that input on practical application of ethical, best-practice goals will be forthcoming and that cross-border, cross-sector discussions can arrive at a consistent approach for ease of implementation.

Allen & Overy LLP

One Bishops Square, London E1 6AD, United Kingdom

Tel +44 20 3088 0000

Fax +44 20 3088 0088

allenovery.com