INTRODUCTORY NOTE TO RECOMMENDATION OF THE COUNCIL ON ARTIFICIAL INTELLIGENCE (OECD)

Author(s): KAREN YEUNG

Source: *International Legal Materials*, February 2020, Vol. 59, No. 1 (February 2020), pp. 27-34

Published by: Cambridge University Press

Stable URL: https://www.jstor.org/stable/10.2307/26908008

DOI:10.1017/ilm.2020.5

INTRODUCTORY NOTE TO RECOMMENDATION OF THE COUNCIL ON ARTIFICIAL INTELLIGENCE (OECD)

BY KAREN YEUNG*

[May 22, 2019]

Introduction

On May 22, 2019, the Organisation for Economic Co-operation and Development (OECD) Ministerial Council Meeting adopted the Recommendation on Artificial Intelligence, signed by all 36 OECD member countries and non-member countries Argentina, Brazil, Columbia, Costa Rica, Peru, and Romania. Its aim is to foster innovation and trust in artificial intelligence (AI) by promoting the "responsible stewardship of trustworthy AI."

Background

Recent and rapid advances in AI research and technologies are widely expected to bring about pervasive and far-reaching social transformation on a global scale. The celebratory rhetoric that accompanied the emergence of these technologies several years ago, championed by Silicon Valley's high-tech "evangelists," has recently become considerably more muted in the face of rising public anxiety about the possible adverse effects associated with the "rise of the machines."[1] Policy-making communities at the national, regional, and international levels now recognize that the disruptive innovations wrought by the emergence of powerful AI technologies and the resulting "New" Industrial Revolution may not be wholly positive, but in fact might have potentially serious and detrimental impacts for individuals, communities, and society.

Yet policy-makers are also keen to avoid killing the golden goose of technological innovation and entrepreneurship that is widely regarded as a key driver of economic growth and prosperity. Accordingly, despite the prevalence and high profile of contemporary policy discussions about the need to regulate AI and ensure its responsible design, deployment, and operation, policy proposals and measures have hitherto largely remained at the level of voluntary initiatives in the form of "ethics codes," which have proliferated in recent years.[2] These codes consist largely of a set of aspirational principles, which "AI actors"[3] publicly indicate that they will voluntarily uphold, or with which they are exhorted and encouraged by their authors to comply. In this respect, the OECD Recommendation sits comfortably within this larger family of "ethical AI" initiatives, prepared by a multidisciplinary group of AI experts appointed in September 2018 by the OECD's Committee on Digital Economy Policy.

The Recommendation

The Recommendation includes two substantive sections.

Principles of Responsible Stewardship for Trustworthy AI

The first section sets out five principles relevant to all stakeholders, which give flesh to the concept of "responsible stewardship for Trustworthy AI" that the Recommendation aims to secure, calling on AI actors to promote and implement these principles according to their respective roles. These principles are concerned with promoting and implementing: (1) inclusive growth, sustainable development and wellbeing through the responsible stewardship of trustworthy AI; (2) human-centred values and fairness (which is defined to include respect for the rule of law, human rights and democratic values) throughout the AI system lifecycle, including the implementation of safeguards; (3) transparency, explainability, and responsible disclosure regarding AI systems (including an opportunity to contest AI-generated outcomes); (4) robustness, security, and safety, which includes a requirement of traceability and the application of a systematic risk management approach to AI systems; and (5) accountability for the proper functioning of AI systems.

* Karen Yeung is Interdisciplinary Professorial Fellow in Law, Ethics & Informatics, Birmingham Law School and School of Computer Science, the University of Birmingham, United Kingdom and Distinguished Visiting Fellow, Melbourne Law School, Australia.

National Policies and International Cooperation

The second section, addressed to Members and non-Members who have indicated their adherence to the Recommendation, sets out five recommendations through which they are called upon to advance progress on responsible stewardship of trustworthy AI via national policies and intergovernmental cooperation. These recommendations, consistent with the five principles outlined in the first section, call upon governments to: (1) invest, and encourage private investment, in AI research and development (including interdisciplinary efforts) to spur innovation in trustworthy AI; (2) foster the development of, and access to, a digital ecosystem for trustworthy AI, including digital technologies, infrastructure, and mechanisms to support the safe, fair, legal, and ethical sharing of data; (3) shape an enabling policy environment for AI, including consideration of controlled environments for experimentation in which AI systems can be tested and scaled up; (4) work closely with stakeholders to prepare for the transformation of the world of work and of society, providing support to ensure a fair transition for workers as AI is deployed with the aim of facilitating the broad and fair sharing of benefits from AI; and (5) actively cooperate with other regional and global organizations to progress responsible stewardship of trustworthy AI, including the promotion and development of multistakeholder, consensus-driven global technical standards for interoperable and trustworthy AI and the development and use of internationally comparable metrics to monitor the progress of implementation.

Discussion

It is difficult to argue with this laudable collection of principles that, taken together, provide the conceptual underpinnings for the concept of "responsible stewardship of trustworthy AI"[4] or the measures that national governments are encouraged to undertake in fostering the responsible stewardship of trustworthy AI. The principles are largely consistent with other national and regional AI ethics codes, resonating strongly with the core tenets of the European Union's Ethics Guidelines for Trustworthy AI.[5] The significance of the Recommendation can only be understood in light of the larger geopolitical context in which the five national entities at the forefront of AI research and development— China, the United States, the European Union, Canada, and Japan—can be understood to be engaging in an "AI arms race."[6] Not only is the winner of this race expected to reap very substantial economic rewards, but it is also expected to accrue political, military, and economic power from its capacity to produce more advanced, powerful, and reliable machines than its competitors, which can then inform decision making and public policy across an almost limitless array of social domains. Thus, despite its lack of legally binding force, the Recommendation represents an important political and moral commitment at the intergovernmental level in at least three respects. First, it reflects welcome recognition that, although AI has the potential to generate great benefits, these technologies have the potential for great harm—to individuals and to society more generally—including the potential to threaten foundational human and democratic values, thus making appropriate policies and governance mechanisms necessary. Second, the commitment to global cooperation is significant, both in potentially blunting the ferocity of an unbridled "AI arms race," and in paving the way for global cooperation to counter the cross-border risks of AI, which are inevitable given the capacity of AI systems to operate on a planetary scale. Third, the 20 largest industrial nations have expressed their support for the principles for responsible stewardship of trustworthy AI via the G20 Osaka Leaders' Declaration.[7] Given that the G20 countries include China and Russia, neither of which is an OECD member, their willingness to approve the same text along with the United States is both unusual and significant.

Agreement on the rather bland principles contained in the Recommendation conceals considerable differences in national approach between major AI players, with both China and the United States emphasizing the economic importance of AI. In contrast, both the European Union and Canada have underscored the importance of ethical concerns as part of their approach to AI policy, while Japan's national AI strategy lies between these two poles. Moreover, the Trump administration emphatically opposes the need for any international treaty in favor of national approaches to regulation (without ruling out international standards or cooperation). While the current US position may be a product of its general retreat from multilateral engagement, there are other less parochial reasons for doubting whether securing intergovernmental agreement is the most conducive approach to fostering the laudable principles espoused in the Recommendation. Less than a month after the OECD Recommendation was adopted, the UN Secretary-General's High-level Panel on Digital Co-operation released its Report,[8] which also highlighted the critical importance of "digital cooperation."[9] Yet, in addition to intergovernmental cooperation, the UN Report also

emphasized the need for "strengthened cooperation mechanisms," reporting that its consultations revealed a "great deal of dissatisfaction with existing digital cooperation arrangements" and the need for "more inclusive processes and better follow up."[10] Recognizing the heightened interdependence brought about by ongoing digital transformations, while concurrently magnifying the threats of exclusion of those who historically have been marginalized from fairly sharing in the benefits of AI, the UN report states that "intergovernmental work must be balanced with work involving broader stakeholders,"[11] thereby emphasizing the need for an inclusive approach to underpin its support for a "multi-stakeholder 'systems' approach for cooperation" (Recommendation 5B), which may require the creation of new mechanisms involving a wide range of stakeholders that are "adaptive, agile, inclusive, and fit for purpose for the fast-changing digital age."

Conclusion

The Recommendation represents a significant step forward at the intergovernmental level in achieving agreement on the importance of, and core principles for, the responsible stewardship of trustworthy AI, particularly as advances in AI research and technology continue apace. Given the scale and speed at which contemporary networked digital technologies now operate, this calls into question the capacity of conventional intergovernmental approaches to cooperation to meet the challenges and potential dangers that these powerful technologies may portend. From a substantive perspective, it is easy to espouse "motherhood and apple pie" principles at a high level of generality and abstraction, to which states with starkly contrasting political commitments have willingly subscribed. It is only when attempts are made to operationalize those principles to specific contexts that conflicting interpretations of how they should be understood and applied begin to surface. Only time will tell how those conflicts will play out at the level of international standard-setting for AI governance.

ENDNOTES

1 *See, e.g.,* JAMIE BARTLETT, THE PEOPLE VS TECH (2018). TIM O'REILLY, WTF? WHAT'S THE FUTURE AND WHY IT'S UP TO US (2017); SHOSHANA ZUBOFF, THE AGE OF SURVEILLANCE CAPITALISM (2019).

2 *See AI Ethics Guidelines Global Inventory*, ALGORITHM WATCH, https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/.

3 "AI actors" are defined as individuals and organisations who play an active in the AI system lifecycle, including those that deploy or operate AI. *See* OECD, Recommendation of the Council on Artificial Intelligence [hereinafter Recommendation], ¶ I, OECD/LEGAL/0449 (May 21, 2019).

4 *Id*. ¶ II. For the purposes of the Recommendation, an "AI system" is defined as a "machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy." *See* Recommendation ¶ I.

5 *See* EU High Level Expert Group on AI, ETHICS GUIDELINES FOR TRUSTWORTHY AI (2019), *at* https://ec.europa.eu/futurium/en/ai-alliance-consultation.

6 Editorial, *International Ethics Panel Must be Independent*, 572 NATURE 415 (2019), *at* https://www.nature.com/articles/d41586-019-02491-x.

7 G20 Osaka Leaders' Declaration. G20 Ministerial Statement on Trade and Digital Economy, *at* https://g20.org/en.

8 UN Secretary General's High-Level Panel on Digital Cooperation, THE AGE OF DIGITAL INTERDEPENDENCE (June 2019), *at* https://www.un.org/en/pdfs/DigitalCooperation-report-for%20web.pdf.

9 Digital cooperation is defined as "the ways we work together to address the social, ethical, legal and economic impact of digital technologies in order to maximise their benefits and minimise their harm." *See id*. 6.

10 *Id*. 22.

11 *Id*. 23.

DOI:10.1017/ilm.2020.5

RECOMMENDATION OF THE COUNCIL ON ARTIFICIAL INTELLIGENCE (OECD)*

[May 22, 2019]

Recommendation of the Council on Artificial Intelligence

THE COUNCIL,

HAVING REGARD to Article 5 b) of the Convention on the Organisation for Economic Co-operation and Development of 14 December 1960;

HAVING REGARD to the OECD Guidelines for Multinational Enterprises [OECD/LEGAL/0144]; Recommendation of the Council concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data [OECD/LEGAL/0188]; Recommendation of the Council concerning Guidelines for Cryptography Policy [OECD/LEGAL/0289]; Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information [OECD/LEGAL/0362]; Recommendation of the Council on Digital Security Risk Management for Economic and Social Prosperity [OECD/LEGAL/0415]; Recommendation of the Council on Consumer Protection in E-commerce [OECD/LEGAL/0422]; Declaration on the Digital Economy: Innovation, Growth and Social Prosperity (Cancún Declaration) [OECD/LEGAL/0426]; Declaration on Strengthening SMEs and Entrepreneurship for Productivity and Inclusive Growth [OECD/LEGAL/0439]; as well as the 2016 Ministerial Statement on Building more Resilient and Inclusive Labour Markets, adopted at the OECD Labour and Employment Ministerial Meeting;

HAVING REGARD to the Sustainable Development Goals set out in the 2030 Agenda for Sustainable Development adopted by the United Nations General Assembly (A/RES/70/1) as well as the 1948 Universal Declaration of Human Rights;

HAVING REGARD to the important work being carried out on artificial intelligence (hereafter, "AI") in other international governmental and non-governmental fora;

RECOGNISING that AI has pervasive, far-reaching and global implications that are transforming societies, economic sectors and the world of work, and are likely to increasingly do so in the future;

RECOGNISING that AI has the potential to improve the welfare and well-being of people, to contribute to positive sustainable global economic activity, to increase innovation and productivity, and to help respond to key global challenges;

RECOGNISING that, at the same time, these transformations may have disparate effects within, and between societies and economies, notably regarding economic shifts, competition, transitions in the labour market, inequalities, and implications for democracy and human rights, privacy and data protection, and digital security;

RECOGNISING that trust is a key enabler of digital transformation; that, although the nature of future AI applications and their implications may be hard to foresee, the trustworthiness of AI systems is a key factor for the diffusion and adoption of AI; and that a well-informed whole-of-society public debate is necessary for capturing the beneficial potential of the technology, while limiting the risks associated with it;

* This text was reproduced and reformatted from the text available at the Organisation for Economic Co-operation and Development website (visited October 18, 2019), https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

UNDERLINING that certain existing national and international legal, regulatory and policy frameworks already have relevance to AI, including those related to human rights, consumer and personal data protection, intellectual property rights, responsible business conduct, and competition, while noting that the appropriateness of some frameworks may need to be assessed and new approaches developed;

RECOGNISING that given the rapid development and implementation of AI, there is a need for a stable policy environment that promotes a human-centric approach to trustworthy AI, that fosters research, preserves economic incentives to innovate, and that applies to all stakeholders according to their role and the context;

CONSIDERING that embracing the opportunities offered, and addressing the challenges raised, by AI applications, and empowering stakeholders to engage is essential to fostering adoption of trustworthy AI in society, and to turning AI trustworthiness into a competitive parameter in the global marketplace;

On the proposal of the Committee on Digital Economy Policy:

I. AGREES that for the purpose of this Recommendation the following terms should be understood as follows:

- *AI system*: An AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy.
- *AI system lifecycle*: AI system lifecycle phases involve: *i)* 'design, data and models'; which is a context-dependent sequence encompassing planning and design, data collection and processing, as well as model building; *ii)* 'verification and validation'; *iii)* 'deployment'; and *iv)* 'operation and monitoring'. These phases often take place in an iterative manner and are not necessarily sequential. The decision to retire an AI system from operation may occur at any point during the operation and monitoring phase.
- *AI knowledge*: AI knowledge refers to the skills and resources, such as data, code, algorithms, models, research, know-how, training programmes, governance, processes and best practices, required to understand and participate in the AI system lifecycle.
- *AI actors*: AI actors are those who play an active role in the AI system lifecycle, including organisations and individuals that deploy or operate AI.
- *Stakeholders*: Stakeholders encompass all organisations and individuals involved in, or affected by, AI systems, directly or indirectly. AI actors are a subset of stakeholders.

Section 1: Principles for responsible stewardship of trustworthy AI

II. RECOMMENDS that Members and non-Members adhering to this Recommendation (hereafter the "Adherents") promote and implement the following principles for responsible stewardship of trustworthy AI, which are relevant to all stakeholders.

III. CALLS ON all AI actors to promote and implement, according to their respective roles, the following Principles for responsible stewardship of trustworthy AI.

IV. UNDERLINES that the following principles are complementary and should be considered as a whole.

1.1. Inclusive growth, sustainable development and well-being

Stakeholders should proactively engage in responsible stewardship of trustworthy AI in pursuit of beneficial outcomes for people and the planet, such as augmenting human capabilities and enhancing creativity, advancing inclusion of underrepresented populations, reducing economic, social, gender and other inequalities, and protecting natural environments, thus invigorating inclusive growth, sustainable development and well-being.

1.2. Human-centred values and fairness

(a) AI actors should respect the rule of law, human rights and democratic values, throughout the AI system lifecycle. These include freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, fairness, social justice, and internationally recognised labour rights.

(b) To this end, AI actors should implement mechanisms and safeguards, such as capacity for human determination, that are appropriate to the context and consistent with the state of art.

1.3. Transparency and explainability

AI Actors should commit to transparency and responsible disclosure regarding AI systems. To this end, they should provide meaningful information, appropriate to the context, and consistent with the state of art:

i. to foster a general understanding of AI systems,

ii. to make stakeholders aware of their interactions with AI systems, including in the workplace,

iii. to enable those affected by an AI system to understand the outcome, and,

iv. to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision.

1.4. Robustness, security and safety

(a) AI systems should be robust, secure and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety risk.

(b) To this end, AI actors should ensure traceability, including in relation to datasets, processes and decisions made during the AI system lifecycle, to enable analysis of the AI system's outcomes and responses to inquiry, appropriate to the context and consistent with the state of art.

(c) AI actors should, based on their roles, the context, and their ability to act, apply a systematic risk management approach to each phase of the AI system lifecycle on a continuous basis to address risks related to AI systems, including privacy, digital security, safety and bias.

1.5. Accountability

AI actors should be accountable for the proper functioning of AI systems and for the respect of the above principles, based on their roles, the context, and consistent with the state of art.

Section 2: National policies and international co-operation for trustworthy AI

V. RECOMMENDS that Adherents implement the following recommendations, consistent with the principles in section 1, in their national policies and international co-operation, with special attention to small and medium-sized enterprises (SMEs).

2.1. Investing in AI research and development

(a) Governments should consider long-term public investment, and encourage private investment, in research and development, including interdisciplinary efforts, to spur innovation in trustworthy AI that focus on challenging technical issues and on AI-related social, legal and ethical implications and policy issues.

(b) Governments should also consider public investment and encourage private investment in open datasets that are representative and respect privacy and data protection to support an environment for AI research and development that is free of inappropriate bias and to improve interoperability and use of standards.

2.2. Fostering a digital ecosystem for AI

Governments should foster the development of, and access to, a digital ecosystem for trustworthy AI. Such an ecosystem includes in particular digital technologies and infrastructure, and mechanisms for sharing AI knowledge, as appropriate. In this regard, governments should consider promoting mechanisms, such as data trusts, to support the safe, fair, legal and ethical sharing of data.

2.3. Shaping an enabling policy environment for AI

(a) Governments should promote a policy environment that supports an agile transition from the research and development stage to the deployment and operation stage for trustworthy AI systems. To this effect, they should consider using experimentation to provide a controlled environment in which AI systems can be tested, and scaled-up, as appropriate.

(b) Governments should review and adapt, as appropriate, their policy and regulatory frameworks and assessment mechanisms as they apply to AI systems to encourage innovation and competition for trustworthy AI.

2.4. Building human capacity and preparing for labour market transformation

(a) Governments should work closely with stakeholders to prepare for the transformation of the world of work and of society. They should empower people to effectively use and interact with AI systems across the breadth of applications, including by equipping them with the necessary skills.

(b) Governments should take steps, including through social dialogue, to ensure a fair transition for workers as AI is deployed, such as through training programmes along the working life, support for those affected by displacement, and access to new opportunities in the labour market.

(c) Governments should also work closely with stakeholders to promote the responsible use of AI at work, to enhance the safety of workers and the quality of jobs, to foster entrepreneurship and productivity, and aim to ensure that the benefits from AI are broadly and fairly shared.

2.5. International co-operation for trustworthy AI

(a) Governments, including developing countries and with stakeholders, should actively co-operate to advance these principles and to progress on responsible stewardship of trustworthy AI.

(b) Governments should work together in the OECD and other global and regional fora to foster the sharing of AI knowledge, as appropriate. They should encourage international, cross-sectoral and open multi-stakeholder initiatives to garner long-term expertise on AI.

(c) Governments should promote the development of multi-stakeholder, consensus-driven global technical standards for interoperable and trustworthy AI.

(d) Governments should also encourage the development, and their own use, of internationally comparable metrics to measure AI research, development and deployment, and gather the evidence base to assess progress in the implementation of these principles.

VI. INVITES the Secretary-General and Adherents to disseminate this Recommendation.

VII. INVITES non-Adherents to take due account of, and adhere to, this Recommendation.

VIII. INSTRUCTS the Committee on Digital Economy Policy:

(a) to continue its important work on artificial intelligence building on this Recommendation and taking into account work in other international fora, and to further develop the measurement framework for evidence-based AI policies;

(b) to develop and iterate further practical guidance on the implementation of this Recommendation, and to report to the Council on progress made no later than end December 2019;

(c) to provide a forum for exchanging information on AI policy and activities including experience with the implementation of this Recommendation, and to foster multi-stakeholder and interdisciplinary dialogue to promote trust in and adoption of AI; and

(d) to monitor, in consultation with other relevant Committees, the implementation of this Recommendation and report thereon to the Council no later than five years following its adoption and regularly thereafter.