PROCEEDINGS OF SCIENCE

Policies for Artificial Intelligence in Science and Innovation

Alan Paic*†
Senior Policy Analyst, Science and Technology policies, OECD
E-mail: alan.paic@oecd.org

This contribution synthesizes the discussions of the special session on policies for Artificial Intelligence in Science and Innovation, organized by the OECD's Directorate for Science, Technology and Innovation. The session was opened by Dr Judith Arrieta, Minister of the Foreign Service at the Chief of Staff's Office of the Secretary of Foreign Affairs of Mexico, and the two panels included speakers from governments, industry and civil society from European countries, USA, Canada China and Australia. Participants discussed the disruptive nature of AI and the formidable challenges it poses. Most of the discussion focused under the umbrella title of ethics, but they span very different issues of human-centered values, fairness, transparency, explainability, and many more. Other challenges include employment, education, SME policy, enabling environment, access to data and computing technology. Responses by governments were also discussed with a particular focus on national strategies, whose main pillars are oriented toward knowledge creation through AI research, knowledge diffusion through linkages to the private sector, development of human capital which will underpin the development of the sector, and a strong values, ethical and regulatory framework to create the conditions for the development of trustworthy AI. In a world of finite resources, discussants concluded that one cannot apply very stringent requirements to all AI decisions, and there is clearly a need to require more transparency, explainability and robustness from systems which have the greatest impact on human lives. Therefore an approach based on algorithmic impact assessment seems reasonable. Such an approach needs to be further developed and standardized.

Artificial Intelligence for Science, Industry and Society, AISIS2019
October 21-25, 2019
Universidad Nacional Autónoma de México, Mexico City, México

*Speaker.
†Organisation for Economic Co-operation and Development, OECD.

PoS(AISIS2019)045

Introduction

The Symposium "Artificial Intelligence for Science, Industry and Society" took place on 21-25 October 2019 at the Universidad Nacional Autónoma de México (UNAM), Mexico City, Mexico. This contribution presents the highlights of a half-day workshop on policy aspects related to Artificial intelligence (AI).

In recent years the field of artificial intelligence has grown from a niche technology into one that can provide disruptive innovation, including autonomous vehicles, as well as systems which can assist humans in decision making, including in pandemic or other disaster relief situations, judiciary (assessment of risk of repeat offending), traffic control and many more.

As is shown in the various contributions throughout these Proceedings, AI already plays a major role in all stages of the scientific process, for example, AI assists radiologists to help recognise cancers in medical imagery and helps particle physicists identify patterns in immensely complex events. Deep learning, high-performance computing, "big data" and open-source software are transforming the scientific landscape at an ever-increasing pace.

In the context of the current COVID-19 pandemic, AI demonstrates the potential for making significant contributions, as shown by the BlueDot algorithm which sent early warning of the new virus, through analysis of news reports and airline ticketing data. (McCall, 2020[15]). AI analysis of CT scans was also used to assist with diagnosis of COVID-19 as a quick alternative to PCR tests, and may be used to forecast further spread and seasonal influences on the virus (McCall, 2020[15]).[1] Deep neural networks and symbolic AI also have a potential to help in the drug discovery process, by extracting connections from scientific literature, as shown by Benevolent AI which has been able to identify a candidate molecule (Richardson et al., 2020[25]). More broadly, AI helps: (i) understand the virus and accelerate medical research; (ii) detect and diagnose the virus, and predict its evolution; (iii) assist in preventing or slowing the virus' spread through surveillance and contact tracing; (iv) respond to the health crisis through personalized information and learning; and (v) monitor the recovery and improve early health warning signals (OECD, 2020[18]).

Amid the opportunities for AI are important issues of governance, trust ranging from questions of transparency (being able to explain and verify the operation of algorithms), to security (for example ensuring the robustness and dependability of AI systems), privacy (not infringing inappropriately on the privacy of users or third parties), as well as broad questions of governance (such as ensuring that AI systems ultimately operate in ways that align with a broad range of societal preferences).

Rapid technological advancement can sometimes outpace policy highlighting the need for policymakers and institutions to anticipate and adjust to changing circumstances.

In addition to introductory remarks, the policy discussions were organised in two panels, the first one focusing on AI strategies geared at preparing for the future, and the second on new policy ideas and harnessing AI for societal progress.

[1] The acutal contribution of AI in this outbreak has thus far been limited, but could be decisive in the future, as discussed in (Heaven, 2020[10])

Introductory remarks

Dr Judith Arrieta, Minister of the Foreign Service at the Chief of Staff's Office of the Secretary of Foreign Affairs of Mexico introduced Mexico's AI strategy which is under development. With the support of the Office of the Presidency, the Digital National Strategy Office, the ministries of Communications and Transportation, and of Foreign Affairs it aims a quintuple helix approach, including government, academia, business, and civil society. The Ministry of Foreign Affairs is also reaching out to the Mexican highly qualified diaspora, which contains a number of individuals highly skilled in machine learning techniques.

The new AI strategy under development will serve the priorities of the National Development Plan 2019-2024, which aims to leverage societal progress under the 2030 Agenda motto 'to leave no one behind', adding: 'and no one outside'. These priorities are the rule of law, welfare, economic development, gender equality, inclusion and anti-corruption. It has also set the year 2024 to expand internet access to all, that is, to more than 130 million inhabitants.

Mexico's Ministry of Foreign Affairs is implementing the use of new technologies including AI, to reach out to Mexicans abroad, focusing in particular on those in fragile conditions. AI can be used to detect hate speech and fake news and thus defend their human rights. For this purpose, a chatbot is being designed with the support of UNAM, to facilitate services and protection. It will also be translated into indigenous languages. This bot will help streamline information to expedite passports, education, health and financial inclusion services to Mexican communities, particularly in the US, through official websites and popular social media platforms. Therefore, the use of cryptocurrencies is also being evaluated to facilitate remittance payments, under the disposition of a Fintech law enacted by the Mexican Parliament last year.

Partnering with UN Women, Mexico is also promoting the digital inclusion initiative (including AI) of the United Nations Digital Cooperation Report 2019. Best practices will be brought to the 75th UN General Assembly in September 2020. The report and high level discussion were initiated by Mexico, supported by 40 nations, with two resolutions (72/242 and 73/17) which requested the 45 UN agencies that deal with technology to assess yearly how to advance the 2030 sustainable development agenda with the help of new technologies, as well as to address their challenges.

Alan Paic, Senior Policy Analyst, Science and Technology policies, OECD

OECD work in the area of Artificial Intelligence began in the context of the G7 ICT Ministerial Meeting in Takamatsu in 2016. This was followed by a high-level conference organised in 2017 at the OECD Conference Centre under the title "AI: Intelligent Machines, Smart Policies".

This early work has brought to light the realisation that the rapid growth of AI technologies requires swift development and deployment of good policies. Given the development, breadth and universal reach of AI and related technologies, a multi-disciplinary and multi-stakeholder approach is needed, and requires global dialogue and collaboration across borders.

In May 2018, the OECD established an AI group of experts at the OECD (AIGO) to scope principles to foster trust in, and adoption of, AI. AIGO was a multi-stakeholder and multi-disciplinary body comprising more than 50 experts from governments, business academia, international organisations, the technical community, trade unions and civil society.

Building on the work of AIGO, on May 22, 2019, at the annual Ministerial Council Meeting, the OECD's 36 member countries, along with Argentina, Brazil, Colombia, Costa Rica, Peru and Romania (in total 42 countries), formally adopted the OECD AI Principles – the first set of intergovernmental policy guidelines on AI (OECD, 2019[21]).

Figure 1: OECD Principles on AI. Source: (OECD, 2019[21]).

The Principles consist of five values-based principles for the responsible stewardship of trustworthy AI, as well as five recommendations to policy-makers pertaining to national policies and international co-operation (Figure 1). They aim to foster innovation and trust in AI by guiding governments, organisations and individuals to develop and operate AI systems in a way that puts people's best interests first while ensuring that those who play an active role in the AI system lifecycle – including AI system designers, developers and operators – are held accountable for their proper functioning.

The objective is to maximise the benefits from AI while minimizing the risks and concerns, such as the impact on human rights and values, work and jobs. Special attention is dedicated to international cooperation and inclusion of developing countries and underrepresented populations.

While not legally binding, OECD Principles do carry a political commitment, and in other policy areas have proved highly influential in setting international standards and helping governments to design national legislation.

In June 2019, Leaders of the G20 convened in Osaka, Japan, and agreed to commit to a human-centered approach to AI, guided by the G20 AI Principles drawn from the OECD Recommendation

on AI. This is a significant accomplishment given that the G20 accounts for most of the AI development underway globally.

In parallel, the OECD published an analytical report on AI, entitled "Artificial Intelligence in Society" (OECD, 2019[19]), in June 2019. It examines the AI technical and economic landscape, providing a historical overview of the evolution of AI and AI's new role as a general-purpose technology that can lower the cost of prediction and enable better decisions. The report describes areas which are experiencing rapid uptake of AI technologies and highlights key related policy questions. Its goal is to help build a shared understanding of AI in the present and near term, and to encourage a broad dialogue on important policy issues, such as labour market developments and upskilling for the digital age; privacy; accountability in AI-powered decisions; and the responsibility, security and safety questions that AI generates. The report also presents an analytical base supporting the AI principles.

The OECD AI Policy Observatory (OECD.AI), launched on 27 February 2020, aims to help countries encourage, nurture and monitor the responsible development of trustworthy artificial intelligence (AI) systems for the benefit of society. As an inclusive online platform for public policy on AI, the Observatory provides a comprehensive database of AI policies from across the world.

OECD.AI combines resources from across the OECD with those of partners from all stakeholder groups to facilitate dialogue and provide multidisciplinary, evidence-based policy analysis on AI. This includes resources on AI public policy topics; AI policies and initiatives; AI trends and data; and practical guidance on implementing the OECD AI Principles. OECD.AI serves as a centre for policy-oriented evidence, debate and guidance for governments, supported by strong partnerships with a wide spectrum of external actors. It leverages live data from partners to help inform policy by showing dynamic visualisations with timely trends about where, how and at what rate AI is being developed and used, and in which sectors.

Definitions of AI

Artificial intelligence is a broadly used term, but not always precisely defined. In addition to the definition developed by the OECD, we will present some definitions used in the countries represented in the panel.

- The OECD definition of AI is based on a proposal by the AI Group of experts at the OECD (AIGO) that scoped the OECD Principles on AI, in view of delineating the scope of applicability of the OECD Principles: "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. It does so by utilising machine and/or human-based inputs/data to: i) perceive real and/or virtual environments; ii) abstract such perceptions into models manually or automatically; and iii) use model interpretations to formulate options for outcomes. AI systems are designed to operate with varying levels of autonomy." (OECD, 2019[23])

- The French government relies on the definition provided by the Villani Mission on AI: **Imitate human capabilities**. Marvin Lee Minsky, who is considered as one of the founding fathers of AI, defines it as follows: "the science of making machines do things that would

require intelligence if done by men. It requires high-level mental processes such as: perceptual learning, memory and critical thinking. In other words, artificial intelligence is the science of building computer programs that aim to perform tasks that would require some intelligence if they were done by human beings. Therefore, no human activity seems to be out of reach: moving from one place to another, learning, reasoning, socializing, creativity etc. Nevertheless, we are still far from creating a machine that would be able to match or outperform human capabilities in all fields." (Villani Mission on AI, 2018[27])

- The definition used for the term "artificial intelligence" by the German Data Ethics commission is the following: "In this context, we understand 'artificial intelligence' as a collective term for technologies and their applications which process potentially very large and heterogeneous data sets using complex methods modelled on human intelligence to arrive at a result which may be used in automated applications. The most important building blocks of AI as part of computer science are sub-symbolic pattern recognition, machine learning, computerized knowledge representation and knowledge processing, which encompasses heuristic search, inference and planning." (German Ministry of Justice and Consumer Protection, 2018[8])

- The Mexican white paper on AI, prepared by civil society organisations, defines it as follows: "Artificial intelligence (AI) refers to machines, and generally computer systems, that can simulate the processes of natural intelligence displayed by humans. These processes include learning, reasoning, and self-correction[2]. The phrase 'artificial intelligence' is now an umbrella term that refers to a broad range of research approaches and technologies". Additional explanations are provided, referring to 'weak' and 'strong' AI, as well as machine learning as "a subfield of AI that involves building algorithms which learn from experience and make predictions about data, without being explicitly told what to do." (Oxford Insights, CMINDS, British Embassy in Mexico, 2018[24]).

Panel 1: Artificial Intelligence strategies – preparing for the future

This first panel of the session was dedicated to AI strategies and other policies preparing the future for trustworthy, transparent, explainable and ethical development of AI, and explored the following questions:

- Based on the conceptualisation and development of your national AI strategy, what do you consider is the most important and/or the most difficult "ethical" issue that governments must address?

- As AI and its applications evolve, new governance challenges will arise. What foresight mechanisms might be recommended to anticipate such challenges and what role must the private sector play?

[2]For the purposes of this report, "learning" refers to progressively improving performance on a specific task, without being explicitly programmed; "reasoning" refers to the ability to make inferences.

- Most AI-related R&D is undertaken by a small number of firms. In what ways does this reality inform and shape the public R&D element of national AI strategies?

The panel was comprised of speakers from Germany, France, Australia, Canada and the AI for Good Lab at CMINDS (NGO).

Nations across the world are realising that a strategic approach is needed, in order to be able to make society benefit from this new technology, and encourage innovation, while protecting privacy, human dignity and the democratic rights of citizens.

While Mexico's strategy is still under development, France, Germany and Australia presented their AI strategies, and Canada presented its Directive on Automated Decision-making as an initial cornerstone of AI policy making.

Mexico

Building on the introductory remarks by Dr Judith Arrieta, Cristina Martínez Pinto, AI for Good Lab Director at CMINDS presented a diagnostic white paper titled 'Towards an AI Strategy in Mexico - Harnessing the AI revolution' (Oxford Insights, CMINDS, British Embassy in Mexico, 2018[24]) which was co-developed with inputs from 70 experts, and was based on a mapping of the AI ecosystem, and an analysis of international AI strategies. Common themes developed in the report include open data policies, data infrastructure and cybersecurity, governance, government and public services. A draft national vision has been proposed, which calls for the creation of an AI working group in Congress, a national centre for AI research, and strengthened links between government, public service, academia and industry. The report also recommends the development of skills and education in order to develop more computational thinking, as well as soft skills such as empathy and complex thinking. Finally, ethics is an important pillar, and the report calls for the creation of a Mexican AI Ethics Council. An additional pillar has been added to reinforce linkages with Mexicans abroad, leveraging the highly skilled diaspora in building up the AI ecosystem.

Currently, the process is being pushed forward by a citizen-led coalition of 140 people, and a number of pilots are being conducted which will help understand the impact of future legislation, making sure regulation would protect without stifling innovation. Regulatory sandboxes are being developed to help understand how start-ups and companies would abide by certain rules while encouraging innovation. Experiments are being conducted with data trusts, mechanisms for sharing data from private companies for use by the public sector in order to improve public services. This is based on international best practice, adapted to the Mexican national context, and applied to Mexico's national interest. The Commission for Science and Technology in the Chamber of Deputies is closely monitoring the outcomes, and will likely propose legislation based on those outcomes.

One of the pilots, being developed in partnership with the Inter-American Development Bank, is called 'fAIr Jalisco' and aims at setting up an AI centre in the state of Jalisco with the objective of accelerating the responsible use of AI for social policies and social impact entrepreneurship. In this pilot, AI serves social purposes, for those who need it the most, providing better and personalised social services. The focus is on communicating a clear narrative to the beneficiaries, focusing on the impact, rather than the technology, with the objective to achieve a better outcome (CMINDS, n.d.[2]).

Germany

Andreas Hartl, Head of Division on Strategy Artificial Intelligence, Data Economy and Blockchain, German Ministry of Economic Affairs and Energy presented Germany's strategy "AI Made in Germany", adopted in 2018, with a global budget allocation of 3 billion euros until 2025. The strategy builds on the basis of decade-long world class research in the field, but also realisation that it lags in applications of this research and especially take-up by companies. The three main objectives are: (i) make Germany and Europe a leading centre for AI and thus help safeguard Germany's competitiveness in the future, (ii) responsible development of AI, and (iii) integrate AI in society in ethical, legal, cultural and institutional terms in the context of a broad societal dialogue and active political measures. The field is also characterised by worries about lack of explainability ("black boxes") and potential biases.

The additional funding under the Strategy will mainly go into reinforcing research, with a goal further building up the national network of Centres of Excellence for AI, and supporting the transfer of AI expertise into use of especially SME businesses. This will also finance 30 new research professorships by 2024, eventually to be expanded to 100 professorships, and also reinforce 6 national research centres in order to double the research capacity in AI. Knowledge transfer to SMEs is to be ensured particularly through 26 dedicated Centres of Excellence. An innovation competition is organised with 135 consortia bidding for 10 M$\in$ of prizes. 15 consortia will be granted funding. Further calls will follow-up.

Additional fields of action, include: (i) International research and innovation cooperation (bilateral France-Germany, as well as European); (ii) systemic approach to software & processors; (ii) boosting the stock of data while at the same time enabling a high-performance, competitive, secure and trustworthy data infrastructure for Europe with project GAIA-X (German Federal Ministry for Economic Affairs and Energy, 2020[6]).

France

Ana Valcarcel Orti, Inria and Project Manager of the French AI Research Programme presented France's "AI for Humanity" strategy (March 2018) (French Government, 2018[5]). With a budget of 1.5 billion euros for five years, the French Strategy for AI is focusing on three axes: (i) achieving best-in-class level of research for AI, through training and attracting global talent in the AI field, (ii) disseminating AI to the economy and society through spin offs and public-private partnerships and data sharing, as well as (iii) establishing an ethical framework for AI. Those priorities are based on the Villani mission report, written based on hearings of 300 AI experts from around the world (Villani, 2018[28]). The strategy also has a strong international component: bilateral cooperation, actions with international organisations (OECD, UNESCO and G7/G20) and convergence of strategies with the European Union.

The research and talent pillar represents 45% of the total budget. The objective of this axis is to establish France as one of the top 5 countries in AI. To this aim several actions will be carried out in a first phase extending from the end of 2018 to 2022:

- Set up national research network in AI coordinated by INRIA;

- Initiate 4 Interdisciplinary Institutes for Artificial Intelligence: 3IA@Côte d'Azur, ANITI in Toulouse, MIAI in Grenoble and PRAIRIE in Paris;
- Promote programs of talent support throughout the country;
- Contribute to the development of a specific program on AI training;
- Increase the computing resources dedicated to AI and facilitate access to these infrastructures;
- Boost public-private partnerships;
- Boost research in AI through the French National Research Agency (ANR) calls;
- Strengthen bilateral, European and international cooperation.

Australia

Estelle Parker, from the Australian Embassy in Mexico underlined that the focus of Australian policy making concerning AI has been focused mostly on the ethical issues but also on the capacity of Australia to harness the opportunities. The government's strategy called "Australia's Tech Future" covers the domain of AI and focuses on four areas: (i) people's skills as an enabler to take advantage of opportunities which can arise from technology; (ii) improvement of government services; (iii) digital assets, i.e. building up a high quality infrastructure for the digital economy (iv) creating an enabling environment, including regulation to fulfil stakeholder expectations about what government can do to regulate AI and other digital technologies.

Australia is very positive about the opportunities AI could create for productivity increase, for example in agriculture through identification and eradication of weeds and targeted herbicide application. Productivity gains need to be identified, and Australia should not be complacent about economic growth (in spite of having achieved growth in 28 consecutive years).

The government has allocated resources to a cooperative research centre at PhD level, as well as for online resources to teach AI at school level. Australia has a Centre of Excellence for automated decision making and society, which is multidisciplinary, bringing together researchers from humanities, social and technological sciences.

An ethics framework ("AI ethics principles") has been co-developed with community groups, civil society, business and government. Mistrust within the public could exist especially in relation to privacy, since AI enables the government to hold and process a large amount of private information about individuals. It was based on pre-existing ethical frameworks, which were applied to AI. The basic premises of the framework are that AI should do no harm, and should contribute net benefits for the community, with AI applications required to ensure legal compliance, transparency, fairness, privacy and accountability (Australian Department of Industry, Science, Energy and Resurces, 2019[1]).

The Australian AI Ethics Principles include: (i) human, social and environmental wellbeing; (ii) human centred values – respecting human rights, diversity and autonomy of individuals; (iii) fairness – the avoidance of discrimination and bias; (iv) privacy protection and security; (v) reliability and safety; (vi) transparency and explainability; (vii) contestability; and (viii) accountability.

The principle of contestability is to ensure that when the algorithm impacts a person, there is an effective and efficient process for the individual to challenge the output of the algorithm. Another important principle is accountability, which requires people responsible for the creation and implementation of algorithms to be identifiable and accountable for the impacts of the algorithms, even if those impacts are unintended.

Canada

Canada was the first national government to adopt a regulation for automated decision making, as a first policy basis for future AI policies, according to Ashley Casovan, former Director of Data Architecture and Innovation for the Treasury Board of Canada Secretariat. The objective of the Directive on Automated Decision-Making is to ensure that automated decision-making systems "are deployed in a manner that reduces risks to Canadians and federal institutions, and leads to more efficient, accurate, consistent, and interpretable decisions made pursuant to Canadian law" (Government of Canada, 2019[9]). Automated decision-making systems are technologies which assist or replace human judgement (e.g. fire prevention, derailment prevention systems). They rely on technologies such as rules-based systems, regression, predictive analytics, machine learning, deep learning, and neural nets.

Algorithmic impact assessment is the cornerstone of this policy. It is used to assess the impact of algorithms on the public, to achieve a balance between innovation and protection of the public, including protection of the individual as well as the community, and serve to extend protection beyond privacy issues.

This balance between innovation and protection is different, according to the use case: clearly the attribution of a certain campsite by a National Park has less of an impact than the decision about the clearance for crossing the national border. Therefore impact levels are scored on a 1-4 scale, and requirements are adapted to the impact levels, with respect to the necessary peer review of the system, notice given to citizens about the functioning (explainability), necessity of including a human in the loop for the final decision, explanation of the final decision, training, contingency planning and approval of the system.

The objective of the Directive is to provide ethical and responsible services: avoid bias, ensure fairness, robustness, interpretability, explainability. In the context of individual projects standards should assist this. The Directive now needs to be implemented, and standards need to be developed.

Panel 2: New policy ideas – how can we harness AI for societal progress?

AI is only as useful as the quality and volume of data it is trained on. Many AI-intensive start-ups master the technology but are constrained by data access. This panel was structured around the following questions:

- What are the best ways to enhance data availability such that AI-intensive firms can work on a broad set of problems relevant to the public interest?
- Fully explainable AI is not yet a reality, even though some progress is being made. What are the best ways to address public concerns over transparency given the current state of AI's development?

- AI today is an extremely fast-moving technology. How can governments best ensure that adjustments in policies and institutions match the rapidity of change?
- We need to ensure that AI technologies and tools will be implemented in a responsible and ethical manner. How do we balance innovation and protection of the public and environment?

Speakers on this panel included representatives from France, Tambourine ventures, AI Global, ATOS, the University of Pisa, IBM and Huawei.

Artificial Intelligence ethical dilemmas

Artificial intelligence has great potential to assist humans, and reduce some of the weaknesses in decision making. In science, AI can lead to faster discovery, cheaper experimentation, and improved knowledge and data sharing and scientific reproducibility (OECD, 2020[17]).

However, numerous issues including acceptable use of AI, explainability, algorithmic bias, enforcement of ethical principles and democratic control over key decisions remain and were discussed in the second panel.

Towards an understanding of societal impacts of AI uses Artificial intelligence can be leveraged across many sectors of human activity, and can enhance business processes in enterprises, the capacity of problem solving in various fields of science, as well as coping with global challenges such as the 2020 pandemic situation, as noted above.

However, AI can also be used for military purposes. A large number of highly respected personalities including Elon Musk, Stephen Hawking and Steve Wozniak, called for a ban on offensive autonomous weapons such as drones in 2015, and the European Parliament passed legislation in 2018 which "Urges the VP/HR, the Member States and the Council to work towards the start of international negotiations on a legally binding instrument prohibiting lethal autonomous weapon systems" (European Parliament, 2018[4]). At the same time, there are plausible arguments towards the fact that such systems might actually decrease human suffering and collateral victims (Müller, 2016[16]).

In 2018, the public was made aware of Project Maven, a contract given to Google by the Pentagon to build AI for drones, in order to differentiate people from objects on the ground. The project was stopped following a petition by 4,000 Google employees who put forward the view that Google was putting users' trust at risk, as well as ignoring its "moral and ethical responsibility".

Additional issues are linked to the manipulation of public opinion, of which the Cambridge Analytica case is just the tip of the iceberg, and the role of AI in the production and use of fake news. Job displacement is another issue raising the question of the continued validity of the Schumpeterian creative destruction – will the technological advances ensure the creation of number of jobs equivalent to the number being lost? These are some of the issues which were developed by Ulises Cortes in his plenary speech on Trustworthy AI at this Conference (Cortes, 2019[3]).

Are Ethics local or global?

A point of debate concerned the degree of global vs. local values influencing ethical principles. On the one hand, it was argued that a common basis is needed for ethical values applicable

worldwide, due to the global nature of AI technology and its applications worldwide. On the other hand it was argued that ethics depend on values, and since relevant societal values vary, ethical values also vary, and even more so their enactment through regulation.

As an example, Professor Andrea Bertolini, Dirpolis Institute and University of Pisa underlined that western societies are centred on the individual, while the Chinese culture values the community interest above the individual. Within Western societies, European countries on the one hand and the United States on the other hand, who both affirm the relevance of human dignity and privacy – the latter intended as freedom of self-determination – more often than not balance said values differently, as it emerges from the bioethical debate and its principles. Similarly, Europeans will favour the precautionary principle as regards regulation, while in the United States a risk-based approach is more common. The diversity of values will inevitably lead to different ethical principles and different regulations across societies, and this is not a problem – on the contrary the dialogue which exists between those different systems should lead to the advancement of all legal systems alike, favouring the development of competing systems.

Prof Bertolini mentioned that one could imagine some products being banned from some countries as non-compliant to local ethical principles – indeed some existing products and services (e.g. alcoholic beverages) are also banned on similar grounds. He thus proposes that solutions be sought at a regional (e.g. European) level, rather than through a global harmonisation.

Nevertheless, Mirjana Stankovic warned that regulating at national or regional level may lead to "protectionism in disguise". For example, data localization policies could affect the free flow of information across borders, and the maintenance of global supply chains, and ultimately free trade in goods and services. Thus, law and ethics should not be analysed in isolation, but rather changes in law should be seen through the perspective of regulatory impact assessment, i.e. how regulations impact global trade flows.

Avoiding bias (‘Artificial stupidity’)

Discussants warned that with AI, there is a danger of introducing new biases, or perpetuating and amplifying previous human biases, in particular when training of the algorithm occurs on the basis of previous human-arbitraged decision data. More than 180 human biases affecting AI decision-making have been identified by IBM researchers (IBM, n.d.[11]). This phenomenon has been referred to as ’artificial stupidity’.

For example, it was mentioned that studies of court decisions in Israel have concluded that judges are likely to rule unfavourably before a meal break when they are hungry (0% parole grants), while after the meal they are much more lenient (65% parole grants). In the United States, an algorithm called COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is being used to assist judges with sentencing and probation decisions. However, this system has been shown to have a racial bias, with double the probability of incorrectly labelling black people as being at high risk of committing repeat violent crime, compared to white people (Larson, Mattu and Lauren Kirchner, 2016[14]).

Further, discussants underlined that such bias needs to be closely monitored. Researchers are developing automatic bias-detection algorithms, mimicking human anti-bias processes, such as control of consistency of decisions and correlation to potential sources of bias such as race, gender, age and other variables.

Bias also needs to be corrected when identified. For example, the Canadian algorithmic impact assessment cited in the previous section provides for an assessment of the severity of potential impact on the individual, and a policy response adapted to that severity.

The explainability of decisions of AI algorithms is a key to controlling algorithmic bias, and will be treated in the next subsection.

Explainability

A common concern with AI is the complexity of the algorithms which contain multiple connections, impossible for the human mind to comprehend, especially in the case of deep learning, which uses a non-linear approach to learning. Such algorithms are often referred to as "black boxes", because they are not available for review by third parties. This is an issue in particular when the decisions of such algorithms have an impact on individuals, and there can be (perceived or real) bias in those decisions. The natural (human) reaction is to ask for an explanation, which most often is not easy to provide.

Professor Andrea Bertolini pointed out that explainability may or may not be needed, depending on the context, and in particular the impact of the AI decision. He mentioned that in the trivial applications this is not needed, but is very important when AI affects the areas of justice or healthcare for example. Such a differential approach is adopted by the algorithmic impact assessment approach of the Canadian Directive on Automated Decision-making. The Directive foresees a simple "Frequently asked questions" on its website for the lowest impact decisions (such as allocating a campsite), while for the high impact decisions (such as authorising a border crossing) it will "ensure that a meaningful explanation is provided with any decision that resulted in the denial of a benefit, a service, or other regulatory action." (Government of Canada, 2019[9])

The French government supports transparency through publishing of the code of all algorithms developed through public funding as open source, enabling individuals to be able to check the functioning of the algorithms, and enhance explainability. However, it was pointed out that access to the algorithms is not sufficient, one would also need access to the training data sets, as well as sufficient computing power to reproduce the results – which is not always feasible.

Mirjana Stankovic, Vice-President for Emerging Technologies at Tambourine ventures cautioned against the use of AI in justice systems or any public policy decision making, precisely because of the limited explainability and potential biases which can be introduced. Instead, she called for policy experimentation in policy sandboxes and policy labs, which may provide useful insights into future applications.

How can ethical principles be enforced?

Mirjana Stankovic called for ethical approaches to AI which should be human rights-centric, incorporating substantive, procedural, and remedial rights. The proposed approach would address 3 layers: governance (national and international), operational – organizational level, and industry self-regulation or setting standards.

Stankovic continued to explain that initially, an 'AI ethics by design approach' was encouraged by professional associations such as the Institute of Electrical and Electronics Engineers (IEEE). In such approaches, ethical concerns are to be considered before a system is deployed, and accountability and transparency are critical principles that must be built in any AI innovation project.

To maintain transparency, the IEEE recommends developing new standards that describe measurable, testable levels of transparency, allowing for objective assessment and determination of the level of compliance (IEEE, 2016[13]). IEEE also promoted principles for ethically aligned design, including safety and beneficence of artificial general intelligence and artificial superintelligence, personal data and individual access control, reframing autonomous weapons systems, economics/humanitarian and law (IEEE, 2017[12]).

Discussants continued by underlining that in practice, keeping algorithms transparent is a challenge because of heavily interlinked and layered processes of algorithmic programming, in particular where deep learning is involved. Alternatives to 'ethics by design' may include ethical reviews at critical junctures. Such mechanisms include codified data ethics principles or codes of conduct, ethical impact assessments and privacy impact assessments, ethical training for researchers, and ethical review boards according to Stankovic. Privacy impact assessments, in general, allow developers and organizations to effectively assess the risks posed (ensuring compliance with privacy requirements, identifying mitigation measures, and effectively classifying the impacts of data and algorithm use). A stakeholder-inclusive approach that features "the proactive inclusion of users" is also desirable. The context of data use should also always be considered, thus requiring human intervention, and at times, context-specific expertise.

An approach based on human-centred values and fairness can be constructed based on human rights impact assessments, which evaluate technology against a wide range of possible human rights impacts. Such assessments can help determine the risks focusing on incidental human impacts (OECD, 2019[20]).

Self-regulation by the private sector

Private sector representatives at the panel underlined the importance of ethical principles for their everyday operations. Cédric Bourrasset, Artificial Intelligence product manager at ATOS mentioned an ATOS product which enables searching and tracking people without using face recognition technology, while complying with the General Data Protection (GDPR). This means that the AI system cannot be based on any pre-existing dataset of personal data, but only allow tracking of a given person within the same recording, with a specific purpose and under specific rules. Andrea Escobedo Lastiri, Government & Regulatory Affairs Leader at IBM Mexico presented IBM's Ethics for AI principles: accountability, value alignment, explainability, fairness and user data rights. She underlined that IBM relies on good practices and open source tools for trust and transparency. In France, a manifesto has been signed by Air Liquide, Dassault Aviation, EDF, Renault, Safran, Thales, Total and Valeo, whereby these companies commit to a coordinated action plan on French AI ecosystem participation including start-ups, SMEs and the public sector.

However, according to Mirjana Stankovic, the effectiveness of self-regulation is questionable: for example, when Google acquired DeepMind in 2014, perhaps the world's most important AI lab, the company agreed to set up an external review board that would ensure the lab's research would not be used in military applications or otherwise unethical projects. But five years later, it is still unclear whether this board even exists. Google, Microsoft, Facebook and other companies have created organizations like the Partnership on AI that aim to guide the practices of the entire industry, but the effectiveness of these operations remains to be proven. According to Stankovic, the most significant changes have been driven by employee protests, e.g. when Amazon employees

protested against the sale of facial recognition services to police departments and various academic studies highlighted the biases within these services (Singer, 2019[26]). Amazon and Microsoft have since called for government regulation in this area.

Safeguards for ethical issues: accountability, contestability, and human in the loop

Estelle Parker from the Australian Embassy in Mexico presented contestability, a key principle of the Australian AI Ethics Framework: when an AI system significantly impacts a person, community, group or environment, there should be a timely process to allow people to challenge the use or output of the AI system. Another important principle according to Ms. Parker is accountability, which requires people responsible for the creation and implementation of algorithms to be identifiable and accountable for the impacts of the algorithms, even if those impacts are unintended.

The discussion further focused on safeguards proportional to the impact on fundamental human rights. Ashley Casovan presented the Canadian Directive on automated decision-making, where in cases labelled as 'high impact' or 'very high impact', "decisions cannot be made without having specific human intervention points during the decision-making process; and the final decision must be made by a human." (Government of Canada, 2019[9]). This is equally so in Germany, where "in the areas with high implications for fundamental rights that require the central decisions to be made by the democratically elected legislator." (German Ministry of Justice and Consumer Protection, 2019[7])

Regulation and democratic control shall prevail

The German data ethics commission presented the 'Opinion of the Data Ethics Commission' to the federal ministers in October 2019. The report presents 75 recommendations, including general ethical and legal principles, such as: (i) human dignity, self-determination, privacy, security, democracy, justice and solidarity, sustainability, and further goes on with more detailed recommendations concerning data and algorithmic systems, including a reference to the Canadian algorithmic impact assessment. The report nevertheless concludes that "regulation is necessary, and cannot be replaced by ethical principles, [...] in particular in the areas with high implications for fundamental rights that require the central decisions to be made by the democratically elected legislator." (German Ministry of Justice and Consumer Protection, 2019[7])

Regulation and democratic control shall prevail

The German data ethics commission presented the 'Opinion of the Data Ethics Commission' to the federal ministers in October 2019. The report presents 75 recommendations, including general ethical and legal principles, such as: (i) human dignity, self-determination, privacy, security, democracy, justice and solidarity, sustainability, and further goes on with more detailed recommendations concerning data and algorithmic systems, including a reference to the Canadian algorithmic impact assessment. The report nevertheless concludes that "regulation is necessary, and cannot be replaced by ethical principles, [...] in particular in the areas with high implications for fundamental rights that require the central decisions to be made by the democratically elected legislator." (German Ministry of Justice and Consumer Protection, 2019[7])

Andreas Hartl underlines that fears are driven by the potential power of AI on human lives, in particular fuelled by depictions in science-fiction on the evolution to a sort of 'superintelligence'.

In its present form the term 'artificial intelligence' may be misleading, since AI is capable in single tasks, not intelligent in the multiple ways humans are. The concerns are about the tipping point when the impact of artificial intelligence starts impacting human rights, the values of the enlightenment, and ethical codes more broadly. We need a discussion: can we accept that a machine does something which is not totally explainable? In some fields this may be socially acceptable, in others not. The bottom line, as underlined by Hartl, is that a machine cannot substitute the democratic will of the people, and therefore in the political sphere, decisions need to be democratically legitimated. Such decision making can be complemented and assisted by machines (e.g. forecasts of migration movements, natural disasters), but the final legislative decision must remain under democratic control, and cannot be replaced by machine learning.

Conclusion

This conference session discussed the promises of artificial intelligence to assist humans in numerous applications and situations, including the current pandemic situation. The power of the technology has a potential to drive innovation, contribute to tackling global challenges and significantly improve societal well-being in general.

Participants discussed its disruptive nature and the formidable challenge to policy makers. Most of the discussion focused under the umbrella title of ethics, but they span very different issues of human-centered values, fairness, transparency, explainability, and many more. Other challenges include employment, education, SME policy, enabling environment, access to data and computing technology.

Responses by governments were also discussed with a particular focus on national strategies, whose main pillars are oriented toward knowledge creation through AI research, knowledge diffusion through linkages to the private sector, development of human capital which will underpin the development of the sector, and a strong values, ethical and regulatory framework to create the conditions for the development of trustworthy AI.

In a world of finite resources, discussants concluded that one cannot apply very stringent requirements to all AI decisions, and there is clearly a need to require more transparency, explainability and robustness from systems which have the greatest impact on human lives. Therefore an approach based on algorithmic impact assessment seems reasonable. Such an approach needs to be further developed and standardized.

References

[1] Australian Department of Industry, Science, Energy and Resurces (2019), *AI Ethics Principles*, https://www.industry.gov.au/data-and-publications/building-australias-artificial-intelligence-capability/ai-ethics-framework/ai-ethics-principles.

[2] CMINDS (n.d.), *Preparing the Global South for the AI Revolution*, https://www.cminds.co/aiforgoodlab.

[3] Cortes, U., *Thrustwothy AI, The AI4EU approach*, PoS(AISIS2019)014, https://pos.sissa.it/372/014.

[4] European Parliament (2018), *European Parliament resolution of 12 September 2018 on autonomous weapon systems (2018/2752(RSP))*, https://www.europarl.europa.eu/doceo/document/TA-8-2018-0341_EN.html.

[5] French Government (2018), *AI for Humanity*, https://www.aiforhumanity.fr/en/.

[6] German Federal Ministry for Economic Affairs and Energy (2020), *A Federated Data Infrastructure as the Cradle of a Vibrant European Ecosystem*, https://www.bmwi.de/Redaktion/EN/Artikel/Digital-World/data-infrastructure.html.

[7] German Ministry of Justice and Consumer Protection (2019), *Opinion of the Data Ethics Commission*, https://www.bmjv.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN.pdf?__blob=publicationFile&v=2.

[8] German Ministry of Justice and Consumer Protection (2018), *Recommendations of the Data Ethics Commission for the Federal Government's Strategy on Artificial Intelligence*, https://www.bmjv.de/SharedDocs/Downloads/DE/Ministerium/ForschungUndWissenschaft/DEK_Empfehlungen_englisch.pdf?__blob=publicationFile&v=3.

[9] Government of Canada (2019), *Directive on Automated Decision-Making*, https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592.

[10] Heaven. (2020), *AI could help with the next pandemic but not with this one*, https://www.technologyreview.com/2020/03/12/905352/ai-could-help-with-the-next-pandemicbut-not-with-this-one.

[11] IBM (n.d.), *AI and Bias*,https://www.research.ibm.com/5-in-5/ai-and-bias/.

[12] IEEE (2017), *Ethically Aligned Design - A vision for Prioritising Human Well-Being with Autonomous and Intelligent Systems*, https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf.

[13] IEEE (2016), *P7001 - Transparency of Autonomous Systems*, https://standards.ieee.org/project/7001.html.

[14] Larson, J., S. Mattu and J. Lauren Kirchner (2016), *How We Analyzed the COMPAS Recidivism Algorithm*, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[15] B. MacCall, (2020), *COVID-19 and artificial intelligence: protecting healtcare workers and curbing the spread*, The Lancet, Vol. 2/4, pp. PE166/E167, https://doi.org/10.1016/S2589-7500(20)30054-6.

[16] Müller, V. (2016), *Autonomous Killer Robots Are Probably Good News* *, http://www.sophia.de (accessed on 18 February 2020).

[17] OECD (2020), *The Digitalisation of Science, Technology and Innovation: Key Developments and Policies*, OECD Publishing, Paris, https://dx.doi.org/10.1787/b9e4a2c0-en.

[18] OECD (2020), *Using artificial intelligence to help combat COVID-19*, https://read.oecd-ilibrary.org/view/?ref=130_130771-3jtyra9uoh&title=Using-artificial-intelligence-to-help-combat-COVID-19.

[19] OECD (2019), *Artificial Intelligence in Society*, OECD Publishing, Paris, https://dx.doi.org/10.1787/eedfee77-en.

[20] OECD (2019), *Artificial Intelligence in Society*, OECD Publishing, https://doi.org/10.1787/eedfee77-en.

[21] OECD (2019), *Recommendation of the Council on Artificial Intelligence*, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 (accessed on 17 February 2020).

[22] OECD (2019), *Recommendation of the Council on Artificial Intelligence*, https://oecd.ai/assets/files/OECD-LEGAL-0449-en.pdf.

[23] OECD (2019), *Scoping the OECD AI principles: Deliberations of the Expert Group on Artificial Intelligence at the OECD (AIGO)*, OECD Publishing, https://doi.org/10.1787/d62f618a-en.

[24] Oxford Insights, CMINDS, British Embassy in Mexico (2018), *Towards an AI Strategy in Mexico - Harnessing the AI Revolution*, https://docs.wixstatic.com/ugd/7be025_e726c582191c49d2b8b6517a590151f6.pdf.

[25] P. Richardson, et al. (2020), *Baricitinib as potential treatment for 2019-nCoV acute respiratory disease*, The Lancet, Vol. 395/10223, pp. PE30-E31, https://doi.org/10.1016/S0140-6736(20)30304-4.

[26] Singer, N. (2019), *Amazon Faces Investor Pressure Over Facial Recognition*, https://www.nytimes.com/2019/05/20/technology/amazon-facial-recognition.html.

[27] Villani Mission on AI (2018), *What is Artificial Intelligence*, https://www.aiforhumanity.fr/pdfs/MissionVillani_WhatisAI_ENG(1)VF.pdf.

[28] Villani, C. (2018), *For a meaningful artificial intelligence - Towards a French and European Strategy*, https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf .