BRIEFING

European Parliament

EU guidelines on ethics in artificial intelligence: Context and implementation

SUMMARY

The discussion around artificial intelligence (AI) technologies and their impact on society is increasingly focused on the question of whether AI should be regulated. Following the call from the European Parliament to update and complement the existing Union legal framework with guiding ethical principles, the EU has carved out a 'human-centric' approach to AI that is respectful of European values and principles. As part of this approach, the EU published its guidelines on ethics in AI in April 2019, and European Commission President-elect, Ursula von der Leyen, has announced that the Commission will soon put forward further legislative proposals for a coordinated European approach to the human and ethical implications of AI.

Against this background, this paper aims to shed some light on the ethical rules that are now recommended when designing, developing, deploying, implementing or using AI products and services in the EU. Moreover, it identifies some implementation challenges and presents possible further EU action ranging from soft law guidance to standardisation to legislation in the field of ethics and AI. There are calls for clarifying the EU guidelines, fostering the adoption of ethical standards and adopting legally binding instruments to, inter alia, set common rules on transparency and common requirements for fundamental rights impact assessments, and to provide an adequate legal framework for face recognition technology. Finally, the paper gives an overview of the main ethical frameworks for AI under development outside the EU (e.g. in the United States and China).

In this Briefing

EPRS | European Parliamentary Research Service

Author: Tambiama Madiega
Members' Research Service

PE 640.163 – September 2019

EN

EU human-centric approach to artificial intelligence

Background

Artificial intelligence (AI) commonly refers to a combination of: machine learning techniques used for searching and analysing large volumes of data; robotics dealing with the conception, design, manufacture and operation of programmable machines; and algorithms and automated decision-making systems (ADMS) able to predict human and machine behaviour and to make autonomous decisions.[1] AI technologies can be **extremely beneficial from an economic and social point of view** and are already being used in areas such as healthcare (for instance, to find effective treatments for cancer) and transport (for instance, to predict traffic conditions and guide autonomous vehicles), or to efficiently manage energy and water consumption. AI increasingly affects our daily lives, and its potential range of application is so broad that it is sometimes referred to as the fourth industrial revolution.[2]

However, while most studies concur that AI brings many benefits, they also highlight a number of **ethical, legal** and **economic concerns**, relating primarily to the risks facing human rights and fundamental freedoms. For instance, AI poses risks to the right to personal data protection and privacy, and equally so a risk of discrimination when algorithms are used for purposes such as to profile people or to resolve situations in criminal justice.[3] There are also some concerns about the impact of AI technologies and robotics on the labour market (e.g. jobs being destroyed by automation). Furthermore, there are calls to assess the impact of algorithms and automated decision-making systems (ADMS) in the context of defective products (safety and liability), digital currency (blockchain), disinformation-spreading (fake news) and the potential military application of algorithms (autonomous weapons systems and cybersecurity). Finally, the question of how to develop **ethical principles** in algorithms and AI design has also been raised.[4]

A recent report by Algorithmwatch – a not-for-profit organisation promoting more transparency in the use of algorithms – lists examples of ADMS already in use in the EU. AI applications are wide-ranging. For instance, the Slovenian Ministry of Finance uses a machine-learning system to **detect tax evasion** and tax fraud. In Belgium, the police are using a predictive algorithm to **predict car robberies**. In Poland, this technology is used to **profile unemployed people** and decide upon the type of assistance appropriate for them.

EU approach

Policy-makers across the world are looking at ways to tackle the risks associated with the development of AI. That said, the EU can be considered a front-runner with regard to establishing a **framework on ethical rules for AI**.

Leading the EU-level debate, the **European Parliament** called on the European Commission to assess the impact of AI, and made wide-ranging recommendations on civil law rules on robotics in January 2017. The Parliament drew up **a code of ethics for robotics engineers** and asked the Commission to consider the creation of a European agency for robotics and AI, tasked with providing the technical, ethical and regulatory expertise needed in an AI-driven environment.[5] Against this background, in 2018 the **Commission** adopted a communication to promote the development of AI in Europe, and in 2019 it published a coordinated plan on AI – endorsed by the **Council of the European Union** – to coordinate the EU Member States' national AI strategies.[6]

Building on this groundwork, in April 2019 the Commission published a set of **non-binding** Ethics guidelines for trustworthy AI. Prepared by the Commission's High-Level Expert Group on AI, composed of 52 independent experts, this document aims to offer guidance on how to foster and secure the development of ethical AI systems in the EU.

Notion of human-centric AI

The **core principle** of the EU guidelines is that the EU must develop a **'human-centric' approach** to AI that is respectful of European values and principles.

> *The human-centric approach to AI strives to ensure that human values are central to the way in which AI systems are developed, deployed, used and monitored, by ensuring respect for fundamental rights, including those set out in the Treaties of the European Union and Charter of Fundamental Rights of the European Union, all of which are united by reference to a common foundation rooted in respect for human dignity, in which the human being enjoys a unique and inalienable moral status. This also entails consideration of the natural environment and of other living beings that are part of the human ecosystem, as well as a sustainable approach enabling the flourishing of future generations to come.*[7]

While this approach will unfold in the context of the global race on AI, EU policy-makers have adopted a frame of analysis to differentiate the **EU strategy** on AI from the **US strategy** (developed mostly through private-sector initiatives and self-regulation) and the **Chinese strategy** (essentially government-led and characterised by strong coordination of private and public investment into AI technologies).[8] In its approach, the EU seeks to remain faithful to its cultural preferences and its **higher standard of protection against the social risks** posed by AI – in particular those affecting privacy, data protection and discrimination rules – unlike other more lax jurisdictions.[9]

To that end, the EU ethics guidelines promote a **trustworthy AI system** that is lawful (complying with all applicable laws and regulations), ethical (ensuring adherence to ethical principles and values) and robust (both from a technical and social perspective) in order to avoid causing unintentional harm. Furthermore, the guidelines highlight that **AI software and hardware systems need to be human-centric**, i.e. developed, deployed and used in adherence to the key ethical requirements outlined below.

Key ethical requirements

The guidelines are **addressed to all AI stakeholders** designing, developing, deploying, implementing, using or being affected by AI in the EU, including companies, researchers, public services, government agencies, institutions, civil society organisations, individuals, workers and consumers. Stakeholders can **voluntarily** opt to use these guidelines and follow the **seven key requirements** (see box on the right) when they are developing, deploying or using AI systems in the EU.

Human agency and oversight

Respect for human autonomy and fundamental rights is at the heart of the seven EU ethical rules. The EU guidelines prescribe three measures to ensure this requirement is reflected in practice:

- to make sure that an AI system does not hamper EU fundamental rights, **a fundamental rights impact assessment** should be undertaken prior to its development. Mechanisms should be put in place afterwards to allow for external feedback on any potential infringement of fundamental rights;
- **human agency** should be ensured, i.e. users should be able to understand and interact with AI systems to a satisfactory degree. The **right of end users not to be subject to a decision based solely on automated processing** (when this produces a legal effect on users or significantly affects them) should be enforced in the EU;

The key EU requirements for achieving trustworthy AI

- human agency and oversight
- robustness and safety
- privacy and data governance
- transparency
- diversity, non-discrimination and fairness
- societal and environmental well-being
- accountability

- a machine cannot be in full control. Therefore, there should always be **human oversight**. Humans should always have the possibility ultimately to over-ride a decision made by a system. When designing an AI product or service, AI developers should consider the type of technical measures that should be implemented to ensure human oversight. For instance, they should provide a stop button or a procedure to abort an operation to ensure human control.

Different types of **fundamental rights impact assessments** are already being used in the EU. The European Commission adopted a set of guidelines **on fundamental rights in impact assessments** and uses this checklist to identify which fundamental rights could be affected by a proposal and to assess systematically the impact of each envisaged policy option on these rights. The **General Data Protection Regulation** (GDPR) provides for a regulatory framework that obliges data controllers to apply a Data Protection Impact Assessment (DPIA). The Government of Canada has also developed an **Algorithmic Impact Assessment** that assesses the potential impact of an algorithm on citizens, with a digital questionnaire evaluating the potential risk of a public-facing automated decision system. This tool will be mandatory in Canada as of 2020.

Technical robustness and safety

Another essential requirement is to have **secure and reliable systems and software**. Trustworthy AI requires algorithms to be secure, reliable and robust enough to deal with errors or inconsistencies during all life-cycle phases of an AI system. This requirement is about ensuring **cybersecurity**. In practice, all vulnerabilities should be taken into account when building algorithms. This requires testing AI systems to understand and mitigate the risks of cyber-attacks and hacking. AI developers should put in place processes capable of assessing the safety risks involved, in case someone uses the AI system they are building for harmful purposes. For instance, if the system is compromised, it should be possible for human control to take over and abort the system. To tackle this important question, the EU applies a twofold approach: first, **fostering cooperation** between the AI community and the security community, and second, reflecting on how to modify the legal framework governing liabilities in the EU, and to go from a **human-conduct-based liability** regime to a more **machine-based liability regime**.

Privacy and data protection

In the EU, there is a lot of consideration for data protection and privacy, and all AI stakeholders must comply with the General Data Protection Regulation (GDPR) as a matter of principle. Furthermore, the EU guidelines on AI advise the AI community to ensure privacy and personal data are protected, both when building and when running an AI system. Citizens should have **full control over their own data**, and their data should not be used to harm or discriminate against them. In practice, this means that AI systems should be designed to guarantee privacy and data protection. To this end, AI developers should apply **design techniques** such as data encryption and data anonymisation. Moreover, they should ensure the **quality of the data**, i.e. avoid socially constructed biased, inaccuracies, errors and mistakes. To that end, data collection should not be biased and AI developers should put in place oversight mechanisms to control the quality of data sets.

Transparency

Transparency is paramount to ensuring that AI is not biased. The AI guidelines introduce a number of measures to ensure transparency in the AI industry. For instance, the data sets and processes that are used in building AI systems should be documented and traceable. Also, AI systems should be identifiable as such, and **humans need to be aware** that they are interacting with an AI system. Furthermore, AI systems and related human decisions are subject to the principle of **explainability**, according to which it should be possible for them to be understood and traced by humans.

Diversity, non-discrimination and fairness

The guidelines focus strongly on **avoiding unfair bias** when AI products and services are designed. In practice, AI developers should make sure that **the design of their algorithms is not biased** (e.g. by the use of an inadequate data set). **Stakeholders** that may be directly or indirectly affected by AI systems should be consulted and involved in their development and implementation. AI systems should be conceived with consideration for the whole range of human abilities, skills and requirements, and ensure accessibility to persons with disabilities.

Explainability – part I

The wide-ranging concept of explainability is about **making explanations on an algorithmic decision-making system available.** The requirement for explainable AI addresses the fact that complex machines and algorithms often cannot provide insights into their behaviour and processes. This sometimes results in a **black box** effect, i.e. a situation where AI systems are capable of producing results, but the process by which the results are produced and the reasons why the algorithm makes specific decisions are not fully understandable by humans.

Explainability is therefore particularly important to **ensure fairness** in the use of algorithms and to identify potential **bias** in the training data. This far-reaching requirement means that an explanation should be available on **how AI systems influence and shape the decision-making process**, on how they are **designed**, and on what is the **rationale** for deploying them. Explainability must address both the technical processes of an AI system and the related human decisions taken in accordance with the EU guidelines.

Societal and environmental well-being

AI systems should be used to enhance positive social change and encourage sustainability and environmental responsibility of AI systems. In other words, measures securing the **environmental friendliness** of AI systems should be encouraged (e.g. opting for a less harmful energy consumption method) and the **social impacts** of these systems (i.e. on people's physical and mental wellbeing) must be monitored and considered. Moreover, the **effects of AI systems on society and democracy** (including regarding the electoral context) should be assessed.

Accountability

Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes. Internal and external independent **audits** should be put in place, especially for AI systems whose use affects fundamental rights. **Reporting** of the AI systems' negative impacts should be available (including for whistle-blowers), and **impact assessment** tools should be used to that end. In situations where the implementation of the key ethical requirements creates conflicts between them, decisions on the **trade-off** (i.e. the decision to choose to fulfil one ethical requirement over another) should be evaluated continuously. **Accessible redress** mechanisms should be implemented.

Implementation challenges

While the implementing phase of the guidelines has started, academics and stakeholders have warned about a number of implementation challenges.

Need for clarification

The lack of clarity in the wording of the guidelines has been criticised in many respects. Thomas Metzinger, professor of theoretical philosophy at the University of Mainz and a member of the Commission's expert group on AI, warns that the guidelines are **short-sighted**, deliberately **vague** and do **not take long-term risks into consideration**. Furthermore, he regrets that the 'red-lines' (i.e. non-negotiable ethical principles) in the draft guidelines were deleted or watered down in the final text. Two of these 'red lines' had been that AI should never be used to build **autonomous lethal weapons** or **social scoring systems**. However, after protracted negotiations, the final version of the

text instead referred to these issues as 'critical concerns' and did not include a clearly formulated prohibition.

Another expert group member, Andrea Renda, together with the AI task force of the Centre for European Policy Studies (CEPS), also published a report highlighting some shortcomings of the draft ethics guidelines. The report warns in particular about the lack of a **hierarchy of principles** that would otherwise have allowed EU institutions to tailor their policy approach.

Lack of regulatory oversight

The EU ethics guidelines are **non-binding**. However, concerns have been raised regarding the lack of regulatory oversight to support their implementation. Non-profit research and advocacy organisation AlgorithmWatch stresses that most of the recommendations and guidelines on AI issued so far do not provide any oversight mechanisms to ensure and enforce compliance with voluntary commitments. Without such mechanisms, however, there is **little incentive to adhere to these ethical principles**. Others also warn about the risk of the technology industry financing and shaping the ethical debate about algorithms and automated decision-making systems. The lack of regulatory oversight raises the issue of the empowerment of public bodies or authorities to **monitor the enforcement of the EU ethical guidelines**.

The AI Now Institute, argues for **expanding the powers of regulators** to oversee, audit, and monitor AI technologies by domain. The institute favours a sector-specific approach that focuses on the application of AI technologies within individual domains (e.g. health, education, transport).

Need for coordination of actions at EU and national levels

Several EU Member States have started work on establishing their own national frameworks on ethics and AI in parallel to the EU initiatives. Below is an outline of these moves, by country.

France

Dating from March 2018, the French AI strategy sets out as one of its core principles the requirement that AI technologies must be explainable to be socially acceptable. To that end, the government is required to: put in place several policies in order to **develop algorithm transparency and audits**; include ethics in training for AI engineers and researchers; carry out a discrimination impact assessment (to encourage AI designers to consider the social implications of the algorithms they produce); and ensure that the principle of human responsibility is applied (e.g. by setting boundaries for the use of predictive algorithms in the law enforcement context). Furthermore, it is proposed that a **consultative ethics committee for digital technologies and AI** be set up for the purpose of organising a public debate in this field.

Germany

Initially, the ethics-related debate was essentially driven by sector-specific industry interests, and resulted in the adoption in June 2017 of a set of ethical rules for automated and connected vehicular traffic by the Transport Ministry's Ethics Commission. In November 2018, the national AI strategy was launched, setting out a range of measures on ethics. For instance, the document advocates using an **'ethics by, in and for design' approach** for all development stages and uses of AI. It pledges to promote research into novel ways for pseudonymising and anonymising data and for differential privacy. Furthermore, the federal government will review whether the German AI-related legal framework covers all aspects related to algorithm-based and AI-based decisions, services and products and, if necessary, adapt it in order to make it possible to verify whether there is any undue discrimination or bias. The legislation governing the **use of personal and non-personal data** for AI-based applications will be reviewed, and the possibility to establish and/or expand government agencies or private-sector **auditing** institutions to verify algorithmic decision-making processes will be examined.

Finland

In August 2018, the Ministry of Economic Affairs issued a report recommending to set up a parliamentary **monitoring group to promote the ethical value** base of AI more extensively in society, and to monitor and evaluate pilots and technology developments associated with the ethical aspects of artificial intelligence. This group would be tasked with the creation of rules and the assessment of practices in the context of defining responsibilities in situations where a machine is taking decisions autonomously.

United Kingdom

The UK Committee on Standards in Public Life announced in March 2019 it was launching an inquiry into the use of AI in public services, with the aim of examining whether the rules were sufficient to ensure that high standards of conduct are upheld as technologically assisted decision-making is adopted more widely across the public sector. In the UK, there is particular focus on the **risks of biometrics** in ongoing discussions. A 2019 report from the UK Biometrics and Forensics Ethics Group outlines some of the ethical issues raised by the use of live (real-time) face recognition technology (FRT) based on machine-learning techniques and recommends the development of an adequate legal framework. Against this background, the House of Commons Science and Technology Committee has urged the UK government to issue a moratorium on the current use of FRT and to prohibit further FRT trials until a proper legislative framework has been introduced, and guidance on trial protocols and an oversight and evaluation system have been established.

Risk of fragmentation. EU Member States are likely to enact some diverging national ethical rules on AI that could fragment the EU landscape in this domain. Such fragmentation may hamper the emergence of pan-European AI services. Therefore, coordinated actions at EU and national levels will be key to ensuring coherent harmonisation of the EU ethical guidelines and avoiding any discrepancies within the EU.

Possible further EU action

Following the publication of the EU guidelines on ethics in AI, the Commission launched a **pilot phase** in June 2019 and invited all stakeholders to provide feedback on the practical implementation of the key requirements by the end of 2019. To this end, companies participating in the pilot will report on their experience in implementing the guidelines. Based on the feedback received, the High-Level Expert Group on AI will propose a revised version of the compliance assessment list to the Commission in early 2020.

Something crucial in this context is to reflect on the following question: to what extent will voluntary ethical rules, driven by industry's pace and strategies, be sufficient to address the ethical issues raised by AI development. There are calls for stronger intervention on the part of public authorities to influence the development and enforcement of these rules. The **European Commission President-elect, Ursula von der Leyen,** has announced that she will put forward **legislative proposals** for a coordinated European approach on the human and ethical implications of AI within her first 100 days in office. Policy-makers, academics and stakeholders have called for further action to implement and complement the ethics guidelines, and equally to ensure a harmonised approach and avoid fragmentation. Possible further action focusing on ethical issues ranges from **soft law guidance**, **hard law legislation** and **standardisation**.[10]

Clarification of the guidelines

One of the main recommendations of the CEPS task force report on AI is to adopt some **guidance** allowing to identify which applications or business models are potentially problematic and which should be prohibited because they are incompatible with EU core values and legislation. The report further stresses that extensive explanations should be provided to establish effective **fairness standards** and that it is necessary to focus more extensively on setting up appropriate **redress**

mechanisms for individuals.[11] Another great challenge is to **clarify how to implement the requirement of explainability** in a context where the complexity of AI algorithms can make it difficult to provide a clear explanation and justification for a decision made by a machine (i.e. **black box effect**). Ensuring a harmonised application of the guidelines throughout the EU would require spelling out this concept in more detail.

Explainability – part II

While AI systems can be made explainable, this may result in **a trade-off between cost and interpretability**.[12] In order to apply the guidelines consistently and efficiently, stakeholders would need additional recommendations on key questions such as: i) do they need to ensure **explainability by design**; ii) can they **differentiate the level of transparency** required when they face cases where AI supports decision-making by humans which may raise fewer explainability issues than fully automated decision-making systems;[13] and iii) to what extent should **intellectual property rights** and **trade secret protection** be limited by the implementation of the explainability requirement. In this regard, the AI Now Institute argues that AI companies should waive trade secrecy and other legal claims that inhibit full auditing and understanding of their software, because such trade secrecy contributes to the black box effect and makes it hard to assess bias, contest decisions or remedy errors.

Standardisation

Standardisation is expected to play an essential role in driving AI market adoption. Standards can influence the development and deployment of particular AI systems through product certification, and serve to disseminate best practices in AI as is the case in cybersecurity or environmental sustainability.[14] A number of standardisation organisations are working on AI technical standards; in parallel, **ethical AI standards** are also being developed.[15] For instance, the joint technical committee of the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) is working on developing standards to ensure trustworthiness in AI technology from the outset. Expert working groups are considering how to technically achieve AI systems' robustness, resiliency, reliability, accuracy, safety, security and privacy. Another leading standardisation organisation, the Institute of Electrical and Electronics Engineers (IEEE) published in 2019 an ethical framework setting out more than 100 ethical issues and recommendations to serve as a reference for policy-makers, engineers, developers and companies deploying, selling and using AI systems. In practice, the IEEE seeks to develop specific **industry standards and processes** – related to transparency, accountability, and algorithmic bias – for the **certification** of AI systems.[16]

AI ethical standards

An IEEE standard establishes a process model by which engineers and technologists can address ethical consideration throughout the various stages of system initiation, analysis and design of new IT products and systems.

Other standards address the manner in which personal privacy terms are offered and how they can be read and agreed to by machines, or describe specific methodologies (e.g. selection of data sets) to address and eliminate bias when algorithms are created.

Against this background, the Commission's 2019 Rolling plan for ICT standardisation identifies three main actions in relation to standard-setting in AI, namely: i) fostering coordination of standardisation efforts on AI in Europe; ii) ensuring coordination between standardisation efforts on AI in Europe and other international standardisation efforts; and iii) integrating the outcomes of the High-Level Expert Group on Artificial Intelligence within the standardisation roadmaps.

However, launching a standardisation process raises many questions. Similar to technical standards, ethical standards are **voluntary measures**. Standards can be made **mandatory** or become a **condition for awarding procurement contracts**[17] so as to ensure that industry players implement them. Some researchers stress, however, that there are not sufficient grounds for the adoption of public certification or mandatory standards on AI in Europe, as the self-certification framework is

evolving and it is too early to anticipate with enough certainty how the AI market will develop over time.[18] The fact that **standards are vague** and **certification enforcement and oversight are unclear** (i.e. who performs the ethical certifications?) has been criticised too. One concern is that the standardisation and certification bodies are focused on enabling AI to become 'ethically' marketable and that existing **market logic will control AI development**.[19] The risk of a **race to the bottom in regulatory oversight** because AI development organisations may choose to locate in jurisdictions that impose more lax rules for implementing ethical standards has also been pointed out.[20]

EU regulatory framework on AI

A number of proposals on AI legislation have been discussed,[21] including several described below.

Legislation on transparency of decision-making systems

Transparency is paramount to ensuring that AI is not biased and AI systems are **explainable**. There are calls to legislate and **make the transparency requirement mandatory**. For instance, the Finnish national AI strategy paper recommends assessing how ethical obligations could be imposed on platforms as is done in the GDPR. The paper stresses in particular that certain parts of an algorithm developed and used by the platforms could be prohibited if it distorts or restricts competition without justification.[22] The EU could build on existing legislative initiatives and research on transparency conducted in recent years. In July 2019, the EU adopted the new Regulation (EU) 2019/1150 requiring providers of online intermediation services and online search engines to implement a set of measures to ensure transparency and fairness in the contractual relations they have with online businesses (e.g. online retailers, hotels and restaurants businesses, app stores) that use such online platforms to sell and provide their services to customers in the EU. The Commission is also carrying out an in-depth analysis on algorithmic transparency.

Against this background, a 2019 Parliament study recommends the **creation of a regulatory body for algorithmic decision-making** tasked with defining i) criteria that can be used to differentiate **acceptable algorithmic decision-making systems** (that should be subject to an algorithmic impact assessment) and systems that should be prohibited; and ii) the **obligations** falling on algorithmic decision-making system providers (such as the obligation to make their systems auditable). New EU legislation could also address the responsibility for informing the persons affected by such systems, while also **clarifying the explainability requirements** and setting **specific liability** and **certifications regimes**.[23]

Sector-specific legislation in the health sector

It is arguably more important to ensure rigorous implementation of the ethical rules in specific sectors, such as healthcare, where human control over algorithms and decision-making systems is paramount. Against this background, the Finnish national AI strategy proposes to formulate AI ethics **rules specific to the healthcare ecosystem**. A 2018 study by the University of Oxford stresses the need to analyse the implementation of the GDPR in the field of health research, and where needed, amend laws or create more clarity through interpretation and guidance.

Legislation on face recognition technology

The use of **face recognition technology** (**FRT)** is becoming widespread across Europe and is giving rise to growing concerns. FRT is considered as processing 'biometric data' under the GDPR, and is in principle subject to strict terms and conditions of use. However, technology experts disagree on whether the GDPR framework is robust enough to address all issues created by the growing use of AI-based FRT, or whether additional legislation will be necessary to ensure EU fundamental rights are protected.[24] Already, the adoption of **national FRT legislation** is being discussed in some Member States (see in particular the UK debate mentioned above).[25]

A number of **legally binding instruments** could be adopted to translate ethical rules into **hard law** and make them **mandatory** for the most influential AI industry players in the EU.

International context

While the EU is clearly a front-runner in the debate on the ethical and social implications of AI, other government entities in the world are also looking at these issues.

United States

In the US, while a range of industry players have already developed some codes of conduct on ethics and AI, there are **calls for more government-led regulation**. Collaborative industry groups, such as the Partnership on AI (including Microsoft, Amazon, Facebook and Apple), have pledged to develop and share best practices, including on ethics. The Association for Computing Machinery (ACM) also published in 2018 a Code of Ethics and Professional Conduct to guide the ethical conduct of computing professionals. Furthermore, companies are developing their own ethical guidelines. For instance, Microsoft has its own AI advisory board and Google has disclosed its AI principles, an ethics charter to guide the responsible development and use of AI in research and products.[26] However, there is growing concern that **self-regulation will not be enough** to tackle the ethical challenges posed by the development of AI. In 2018 the AI Now Institute issued a report concluding that internal governance structures in most technology companies are failing to ensure accountability for AI systems. It argues therefore that government agencies need greater power to oversee, audit and monitor AI technologies, especially those involving face recognition.

China

In China, there is **growing interest** in setting up an ethical framework for the development of AI. In 2017, China released its Next Generation Artificial Intelligence Development Plan setting out long-term strategic goals for AI development in the country by 2030. One objective is to **establish regulatory and ethical frameworks** to ensure the healthy development of AI in China. China would promote self-discipline of the AI industry and enterprises, and increase punishments for data abuse, violations of personal privacy and unethical activities in this regard. The Artificial Intelligence Industry Alliance, which brings together Chinese tech firms and universities, released draft guidelines for **self-regulation** in the field of AI in May 2019.[27] These call for implementing principles of 'human-oriented', 'secure/safe and controllable' and 'transparent and explainable' AI similar to the ones enshrined in the EU AI ethical guidelines. Furthermore, the New Generation AI Governance Expert Committee, established by the Ministry of Science and Technology, released in June 2019 a document outlining **eight non-binding principles** to guide AI development in China. These principles largely mirror the EU rules on AI. For instance, AI development should conform to 'human values, ethics, and morality'; 'should be based on the premise of safeguarding societal security and respecting human rights'; should 'eliminate bias and discrimination in the process of data acquisition, algorithm design, technology development, product R&D, and application'; and should 'respect and protect personal privacy'.[28]

Other countries and organisations in the world

Canada has already adopted a number of guiding principles governing the use of AI in the administration and public services. Public institutions are required to incorporate some ethical principles (including privacy and transparency concerns) in their application of AI.[29] The 2018 Directive on Automated Decision-Making for Federal Institutions outlines the responsibilities of federal institutions and provides rules to help them assess and mitigate the risks associated with deploying an automated decision system. **Australia** is also well advanced. The Office of the Australian Information Commissioner published a Guide to data analytics and the Australian privacy principles in 2018 and is working on a national ethics framework to address standards and codes of

conduct in the field of AI. Preparatory work for establishing an AI ethical framework is also ongoing in **India, New Zealand, Singapore, South Korea** and **Japan.**[30] Some international organisations are engaging in setting international rules in the field of ethics and AI. In May 2019, the **OECD** and associated nations adopted a non-binding list of guidelines for the development and use of AI.

The AlgorithmWatch AI Ethics Guidelines Global Inventory lists all of the ethical frameworks and principles being developed across the globe. Recent years have seen a flurry of initiatives from companies, governments, NGOs and research bodies to propose ethical rules on AI. The ethical principles laid down in other jurisdictions seem relatively similar to those of the EU (though less detailed) and are essentially of a self-regulatory nature, even though there is growing demand for more government oversight.

Outlook

Policy-makers all over the globe are looking at how to tackle the risks associated with the development of AI. In April 2019, the EU published its guidelines on ethics in AI, becoming a front-runner in the setting up of a framework for AI. Ethical rules on AI, where such exist, are so far essentially of **a self-regulatory nature**, and there is **growing demand for more government oversight.** In the EU, there are strong calls for clarifying the EU guidelines, fostering the adoption of ethical standards and adopting legally biding instruments in order to, inter alia, set **common rules on transparency**, set common **requirements for fundamental rights impact assessments** and provide an adequate legal framework for **face recognition technology**.

MAIN REFERENCES

Boucher P., How artificial intelligence works, briefing, EPRS, March 2019.

The Age of Artificial Intelligence, European Political Strategy Centre, European Commission, March 2018.

Ethics Guidelines for Trustworthy AI, Independent High-Level Expert Group on Artificial Intelligence, European Commission, April 2019.

Artificial Intelligence: A European Perspective, JRC, European Commission, 2018.

Kritikos M., Artificial Intelligence ante portas: Legal & ethical reflections, briefing, EPRS, March 2019.

A governance framework for algorithmic accountability and transparency, study, Scientific Foresight Unit (STOA), EPRS, April 2019.

Understanding algorithmic decision-making: Opportunities and challenges, study, STOA, EPRS, March 2019.

ENDNOTES

1 See definition provided by the High-Level Expert Group on Artificial Intelligence (glossary section of the Ethics Guidelines for Trustworthy AI). There is no commonly agreed definition for AI. For an overview of the notion of AI and the difficulty in defining it, see Philip Boucher's 2019 EPRS briefings on How artificial intelligence works and on Why artificial intelligence matters.

2 See EPRS briefing on Economic impacts of artificial intelligence by Marcin Szczepański, July 2019.

3 See Data quality and artificial intelligence – mitigating bias and error to protect fundamental rights, FRA Focus, EU Agency for Fundamental Rights, June 2019.

4 See two EPRS publications by Mihalis Kritikos: What if algorithms could abide by ethical principles?, 'at a glance' note, November 2018; and Artificial Intelligence ante portas: Legal & ethical reflections, briefing, March 2019.

5 Furthermore, the Parliament adopted an own-initiative-resolution on a Comprehensive European industrial policy on artificial intelligence and robotics in February 2019. The European Economic and Social Committee issued an own-initiative opinion on AI in May 2017 also calling for a code of ethics for the development, application and use of AI.

6 Furthermore, on 10 April 2018, 25 European countries signed a Declaration of cooperation on Artificial Intelligence.

7 See definition provided by the High-Level Expert Group on Artificial Intelligence (glossary section of the Ethics Guidelines for Trustworthy AI).

8 See The Age of Artificial Intelligence, European Political Strategy Centre, European Commission, March 2018.

9 Ibid. See as well Artificial Intelligence: A European Perspective, JRC, European Commission, 2018.

10 EU action could also be taken to clarify the issue of liability of damages and measures to encourage data-sharing.

[11] See Andrea Renda, Artificial Intelligence: ethics, governance and policy challenges, report of a CEPS Task Force, 2019, p. 117.

[12] See PricewaterhouseCoopers, 2018 AI predictions: 8 insights to shape business strategy, 2018. The report argues that most AI can be made explainable but at a cost, and explains that, if every step must be documented and explained, the process becomes slower and may be more expensive.

[13] See OECD, above at p. 93.

[14] See Peter Cihon, Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development, University of Oxford, April 2019.

[15] See Jens Popper et al., Artificial intelligence across industries, International Electrotechnical Commission (IEC) whitepaper, October 2018, pp. 71-78.

[16] See Marc Böhlen cited in Industry standards won't give artificial intelligence a conscience, Matthew Linares, openDemocracy media platform, February 2019.

[17] See Alan F. T. Winfield, Ethical standards in robotics and AI, University of the West of England, Bristol, February 2019.

[18] See A. Renda, op.cit., p. 118.

[19] See Marc Böhlen cited in M. Linares, op.cit.

[20] See P. Cihon, op.cit., p. 17.

[21] See a draft European Commission priorities working document published by *Politico* in August 2019.

[22] See Work in the age of artificial intelligence – Four perspectives on the economy, employment, skills and ethics, Ministry of Economic Affairs and Employment of Finland, 2018 p. 54.

[23] See the following EPRS studies: A governance framework for algorithmic accountability and transparency, April 2019; Understanding algorithmic decision-making: Opportunities and challenges, March 2019; Cost of non-Europe in robotics and artificial intelligence - Liability, insurance and risk management, June 2019.

[24] On 21 August 2019, the Swedish Data Protection Authority (DPA) imposed its first fine since the EU GDPR came into effect in May 2018. The fine was imposed on a school for creating a facial recognition program in violation of the GDPR. More generally, how the GDPR affects machine learning techniques remains unclear in many respects. See also Mapping regulatory proposals for artificial Intelligence in Europe, Access Now non-profit advocacy group, November 2018.

[25] Work has started being undertaken on the application of FRT in Schengen border control. See Biometrics and the Schengen Information System – Fostering identification capabilities, JRC blog, European Commission, July 2019.

[26] Google also established an Advanced Technology External Advisory Council (ATEAC) featuring prominent academics and designed to monitor its use of artificial intelligence, but decided to dissolved it in April 2019 amid a controversy about the appointment and independence of some of the Council members.

[27] See Joint Pledge on Artificial Intelligence Industry Self-Discipline (Draft for Comment), June 17, 2019.

[28] See Chinese Government Governance Principles for a New Generation of Artificial Intelligence: Develop Responsible Artificial Intelligence, 17 June 2019. The eight principles with regard to AI are: harmony and friendliness; fairness and justice; inclusivity and sharing; respect for privacy; secure/safe and controllable; shared responsibility; open collaboration; and agile governance.

[29] See Responsible Artificial Intelligence in the Government of Canada (Digital Disruption White Paper Series), Treasury Board of Canada Secretariat, 10 April 2019.

[30] For an overview, see Regulation of Artificial Intelligence in Selected Jurisdictions, the Law Library of Congress, United States, January 2019.

DISCLAIMER AND COPYRIGHT

eprs@ep.europa.eu (contact)

www.eprs.ep.parl.union.eu (intranet)

www.europarl.europa.eu/thinktank (internet)

http://epthinktank.eu (blog)