cepInput

07 | 2019

EU and OECD Ethics Guidelines on Artificial Intelligence

A comparison of the two documents

Alessandro Gasparotti

In April 2019, the EU published ethics guidelines for the development of Artificial Intelligence (AI). In May 2019, the OECD published a separate set of such guidelines. The G20 heads of state and government adopted the OECD guidelines. This cep**Input** compares the two sets of guidelines. The main findings are:

- There is no major difference between the EU and OECD guidelines. Both aim to create a framework for trustworthy AI. The EU guidelines, however, are clearer and go more into detail.
- The G20's endorsement of the OECD guidelines, then, does not undermine the relevance of the EU ones.
- The importance of the EU guidelines might even increase, as the EU guidelines include practical instructions for their implementation, which the OECD is currently developing. The OECD might use the EU's practical instructions as a template.
- The intended EU legislative action on AI should refrain from making the guidelines binding: As the EU is lagging behind in the development and deployment of AI, binding AI guidelines would further increase the EU's competitive disadvantage vis-à-vis the USA and China.

Table of contents

1 Introduction ... 3
2 The drafting of the EU and the OECD guidelines ... 4
3 Content of the EU and OECD guidelines ... 6
3.1 The EU ethics guidelines ... 7
3.1.1 Three Components ... 7
3.1.2 Four Ethical Principles ... 7
3.1.3 Seven key requirements ... 8
3.2 The OECD ethics guidelines ... 9
3.3 Comparison of the EU and OECD ethics guidelines ... 10
4 The future of the EU ethics guidelines ... 11
5 Annex ... 12

1 Introduction

Following swift progress in computing power and machine learning techniques through, inter alia, deep neural networks, governments around the world have drafted national strategies regarding artificial intelligence (AI). These strategies outline visions and policies on how to boost research in the field of AI and how to adapt to the social changes that will come along with increased use of AI. Some of these strategies also deal with the question of how to regulate AI.

The rationale for regulating the development and the use of AI is that AI may have serious implications for democracy[1], human rights[2], privacy[3], and digital security[4]. These possible negative consequences might be aggravated by the fact that AI is a general-purpose technology, i.e. can be applied to theoretically every sector of the economy.

The two leading countries in the field of AI development and adoption are the USA and China. In 2016, AI investment totalled around € 15 billion in North America and around € 8 billion in Asia, most of it in China. In the same year, European investment in AI only amounted to around € 2.6 billion[5]. Although they are aware of the possible negative consequences of AI, the USA and China have so far decided not to regulate the development and use of AI. Both countries are afraid that imposing strong regulation would reduce their relative competitiveness in AI development.

However, as many stakeholders see a need for regulation, various organizations have published documents on the minimum ethical standards to be met by AI. These include research centres[6], industries[7], and international organizations[8].

The EU is one of these organizations. On 8 April 2019, the EU published ethics guidelines on AI. These guidelines are part of the European AI Strategy that the EU Commission announced on 25 April 2018[9]. The strategy consists of three pillars aimed at boosting the EU's technological capacity and investment in R&D[10], preparing for the socio-economic changes resulting from an increased use of AI[11] and ensuring an appropriate ethical and legal framework for AI[12]. In implementing the strategy, the EU Commission and Member States have pledged resources to boost research, increase training of EU citizens and assess the fitness of existing legislation vis-à-vis the challenges posed by AI. By the end of 2020, total investment in AI in the EU shall increase to at least 20 billion euros. Furthermore, every Member State will draft its own national AI strategy, with the EU coordinating national efforts, facilitating joint projects and developing economies of scale.

The most recent step towards implementation of the EU's AI strategy was publication of the ethics guidelines. On 25 April 2018, the EU Commission appointed an independent High-Level Expert Group

1 E.g. fake news spread by chat-bots on social media can impact election results.
2 E.g. automated decision-making by companies could also perpetuate human bias and result in discriminatory outcomes.
3 E.g. mass surveillance enabled by face recognition technologies or algorithms capable of extrapolating sexual orientation and political preferences from user's information.
4 E.g. AI could be used to build new, more effective malware that will be able to learn and adapt to launch further attacks.
5 10 imperatives for Europe in the age of AI and automation, McKinsey, 2017 [quoted by COM(2018) 237 p.4].
6 E.g. the "Bejing Academy of Artificial Intelligence" or the "Montreal Declaration for A Responsible Development of AI".
7 E.g. Google, Deutsche Telekom, SAP, IBM, and Sony.
8 E.g. The UNESCO and the Council of Europe.
9 See COM (2018) 237: Artificial Intelligence for Europe.
10 See cepPolicyBrief No.2019-10.
11 See cepPolicyBrief No.2019-12.
12 See cepPolicyBrief No.2019-13.

on Artificial Intelligence (AIHLEG) with the goal of drafting ethics guidelines for the development and use of AI. On 8 April 2019, the AIHLEG presented the final version of its "Ethics Guidelines for Trustworthy AI".[13] While not binding, these guidelines will be the backbone of possible EU legislative action in the future.

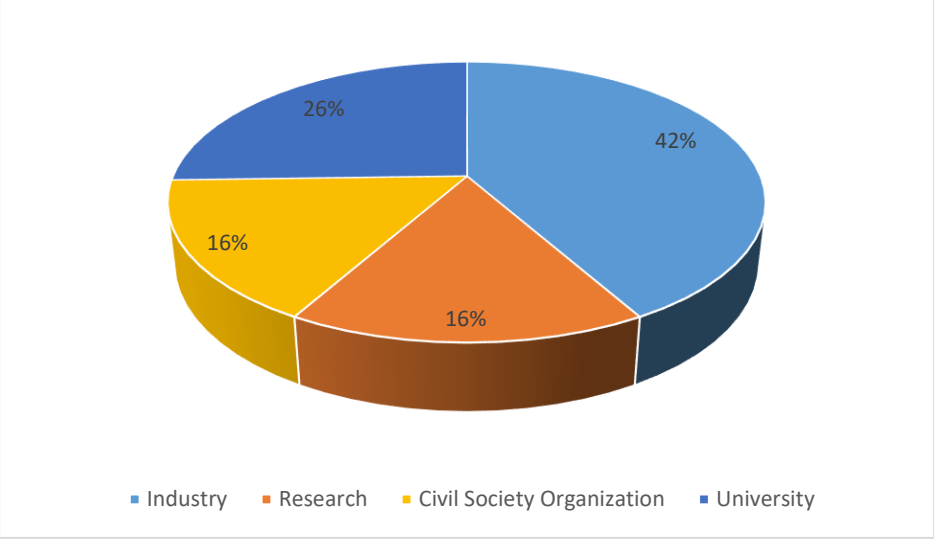
One month after presentation of the EU guidelines, on 22 May 2019, forty-two countries[14] adopted the first intergovernmental standards on AI in the form of an "Organization for Economic Cooperation and Development (OECD) recommendation on Artificial Intelligence"[15]. Ministers from major AI global players, among them Israel, Japan, Korea, and the USA, agreed on a common set of principles for guiding the future development of AI. The majority of EU Member States – also being OECD members – agreed on these standards for AI. At the G20 Summit in June 2019, the heads of state and government also endorsed the OECD ethics guidelines.

This cep**Input** assesses the consequences of this endorsement for the relevance of the EU guidelines. Therefore, section 2 compares the process of drafting the OECD and EU guidelines, while sections 3 compares their content. Section 4 then discusses the relevance of the EU guidelines.

2 The drafting of the EU and the OECD guidelines

The EU Commission appointed an expert group (AIHLEG) and asked it to develop ethics guidelines for AI. The AIHLEG consists of 52 members coming from academia, industry, and civil society[16]. Figure 1 illustrates the composition of the AIHLEG. The AIHLEG is supported by the European AI Alliance[17], a group of more than 2000 stakeholders that provides constant feedback on the work of the AIHLEG.

Fig. 1: AIHLEG members



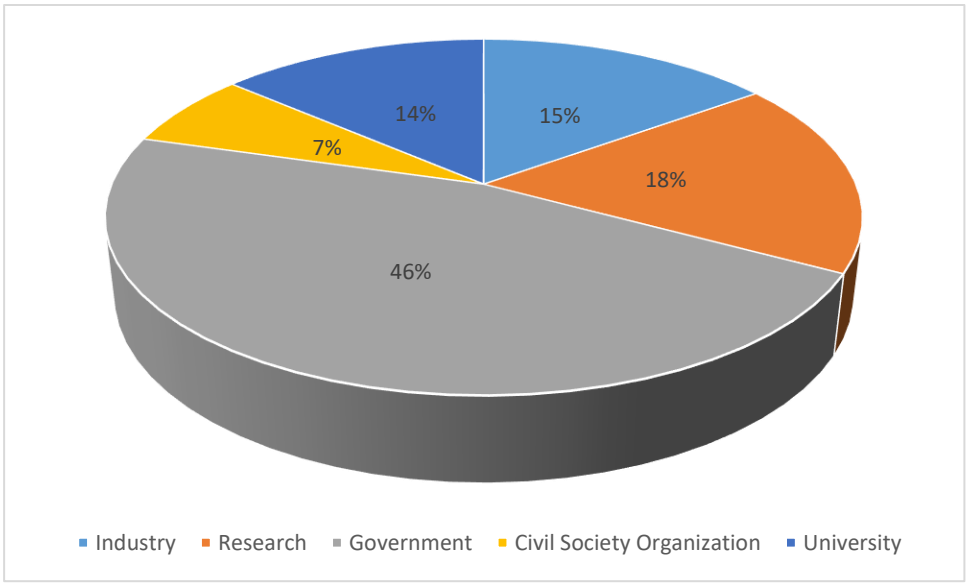
Source: author's calculation based on the official bio of members. Whenever a member could be associated with two groups, it was included in both.

13 See COM (2019) 168 and cepPolicyBrief No. 2019-13.
14 Other than the 36 OECD Member States, the recommendation has already been signed by six non-Member Countries.
15 https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.
16 https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence.
17 https://ec.europa.eu/digital-single-market/en/european-ai-alliance.

The AIHLEG started working on the ethics guidelines on 1 June 2018. Six month later – on 18 December 2018 – it released its draft ethics guidelines[18]. Following a comprehensive revision, the AIHLEG presented the final version of the ethics guidelines[19] on 8 April 2019. On 26 June 2019 the AIHLEG launched a pilot phase for assessment of the guidelines. Up until the end of the year, companies in the EU can voluntarily adopt these guidelines and report to the AIHLEG about the implementation. Before the guidelines are officially transmitted to the EU Commission, the AIHLEG will consider further amendments. These amendments will be based on further consultations with the public and information gathered during the pilot phase, e.g. through in-depth interviews with selected companies. It is planned that in 2020, the EU Commission will have guidelines at its disposal, which have been tested in the real world, and a reasonable stock of information on the opinions and requests of stakeholders. Further action could then follow.

In May 2018, one month after the EU Commission had announced its plan to develop ethics guidelines, the OECD Committee on Digital Economy Policy (CDEP) agreed to form an expert group tasked with drafting ethics principles for the development of AI. The AI Group of Experts at the OECD (AIGO) is composed of 56 members who are supported by 17 external experts. Members and experts are from academia, industry, civil society, and otherwise mainly governments. Interestingly, members of the EU Commission, of the Institute of Electrical and Electronics Engineers (IEEE)[20], and of the United Nations Educational, Scientific and Cultural Organization (UNESCO) are also part of the AIGO. All three organisations produced their own guidelines, suggesting that the OECD worked closely with its partners to align its principles on AI. Figure 2 illustrates the composition of the AIGO.

Fig. 2: AIGO members



Source: Author's calculation based on the official bio of members. Whenever a member could be associated with two groups, it was included in both.

18 https://ec.europa.eu/digital-single-market/en/news/draft-ethics-guidelines-trustworthy-ai.
19 https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.
20 The "IEEE inspires a global community to innovate for a better tomorrow through its more than 423,000 members in over 160 countries. IEEE is the trusted "voice" for engineering, computing, and technology information around the globe".

On 13 and 14 March 2019, the CDEP approved a draft recommendation for AI ethics guidelines and agreed to transmit it to the OECD Council. On the 22 and 23 May 2019, the OECD Council adopted[21] the recommendation at its meeting at ministerial level. Such a recommendation is the first legislative act on the ethical regulation of AI in the world. Although it is a legislative instrument, the OECD guidelines are non-binding for signatory countries. Nonetheless, given the wide global reach of OECD membership, ranging from EU countries, to Japan, Korea, Israel, Canada and the USA, consensus on AI ethics standards seems to be a remarkable achievement.

In order to bolster the implementation of the OECD guidelines, the OECD Council tasked the CDEP to develop practical instructions for their implementation. No timetable currently exists for the release of this practical document.

Table 1 compares the development of the two guidelines. It shows that the drafting of the two guidelines was quite similar. They were both drafted by a group of various actors, and both ended-up being a non-binding document. The most striking difference between the two guidelines is that in the drafting of the OECD guidelines almost 50% of the members are government representatives, while there were no government representatives directly involved in drafting the EU guidelines. Specifically, the EU guidelines were drafted by the users of the guidelines and parties directly affected by them. The next section will assess whether this results in differences in the content of the two guidelines.

Tab. 1: Information on the EU and OECD guidelines

		AIHLEG (EU Commission)	AIGO (OECD)
Composition expert group	Research	16%	18%
	University	26%	14%
	Industry	42%	15%
	Civil Society	16%	7%
	Government	0%	46%
Supranational nature of body requesting		+	+
Binding character		-	-
Further steps ongoing		+	+

3 Content of the EU and OECD guidelines

Both guidelines aim to provide a general framework, which will eventually be complemented by sectoral[22] or contractual[23] regulation. The OECD and the EU argue that this is appropriate given the unforeseeable developments of AI and its diverse applications. According to both organisations, any regulatory framework should be comprehensive, i.e. encompass the whole life cycle of the technology, whilst also being agile, i.e. adaptable to new scenarios without unintentionally hindering technological progress.

21 https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

22 AI can be used in different products, e.g. voice recognition software, surveillance drones, autonomous cars, and medical devices. Given that these applications operate in different fields, it might not be possible to apply the same rules to them.

23 Especially in business to business relations, there should be a flexible framework that guarantee freedom of contract. Business to consumer relations may require more structured rules due to the possible asymmetries of power and information between the parties.

The next section briefly describes the two guidelines, starting with the EU guidelines. At the end of this section, similarities and differences of the guidelines are presented.

3.1 The EU ethics guidelines[24]

As for the EU, trust is a key element in the acceptance and adoption of AI technologies[25], the EU ethics guidelines are proposed as an instrument for establishing trust between producers, deployers and end users of AI. For this reason, the AIHLEG provides an extensive definition of what "trustworthy AI" is, through the declaration of three general components that AI needs to fulfil to be considered trustworthy. One of the components contains four ethical principles. As these principles are very general and therefore hard to implement in practice, the ethical principles are transformed into seven key requirements that give guidance – notably to developers – about how the principles can be implemented in practice. The guidelines furthermore include a questionnaire that helps – notably developers – to assess whether a product complies with the ethics guidelines. Figure 3 outlines the logic of the EU ethics guidelines.

The rest of this section describes the three components, the ethical principles, and the key requirements.

Fig 3: Logic of the EU ethics guidelines

3 components → 4 ethical principles → 7 key requirements → questionnaire

3.1.1 Three Components

Trustworthy AI has three components that should be met throughout the system's entire life cycle. This means that trustworthy AI concerns not only the trustworthiness of the AI system itself. It also encompasses the trustworthiness of all processes and actors that are part of the system's life cycle. The three components of trustworthy AI are:[26]

- **Lawful AI:** AI should comply with all applicable laws and regulations. Given that the law provides both positive and negative obligations, this component of trustworthy AI refers to both what may be done and what should not be done.
- **Ethical AI:** AI should demonstrate respect for, and ensure adherence to, ethical principles and values.
- **Robust AI:** AI should ensure that its use will not cause any unintentional harm.

3.1.2 Four Ethical Principles

The EU guidelines include four ethical principles that are rooted in fundamental rights. The guidelines emphasise that AI practitioners should always strive to adhere to them. These four principles are:[27]

---

24 For a precise assessment see cepPolicyBrief No. 2019-13. A cepPolicyBrief that assesses the content of the EU guidelines will follow.

25 Ethics guidelines for trustworthy AI [p. 4].

26 Ethics guidelines for trustworthy AI [p. 6-7, p. 38].

27 Ethics guidelines for trustworthy AI [p. 11-13].

- **Respect for human autonomy**: Humans should retain full self-determination and oversight over AI; furthermore AI should not unjustifiably deceive, coerce or manipulate humans.
- **Prevention of harm**: AI should neither cause nor exacerbate harm, or otherwise adversely affect human beings, especially considering that AI can cause asymmetries of power or information[28].
- **Fairness**: The costs and benefits of AI should be justly shared, discrimination and unfair bias avoided and effective redress mechanisms against AI-driven decisions made available.
- **Explicability**: The purpose of AI systems should be communicated, processes transparent and decisions explainable, to the extent possible, depending on the context and severity of the consequences of an erroneous output.

These principles are already largely reflected in existing law, e.g. in data protection law. AI is therefore already required to comply with most of the principles. Where a trade-off between these four principles arises, it should be solved through evidence-based reflection[29]. The existence of trade-offs and the way they are addressed should also be presented to, and be understandable for, all stakeholders in order to ensure trust.

3.1.3 Seven key requirements

Due to the general nature of the ethical principles, they are hard to implement, and have therefore been transformed into seven key requirements that provide guidance – notably to developers – on how to implement them. The seven requirements must be met in order to achieve trustworthy AI[30]. Such compliance is necessary but not sufficient to guarantee trustworthy AI.

The seven requirements are:

- **Human agency and oversight:**
  - Negative effects on fundamental rights should be assessed before AI development and reduced or justified;
  - AI systems should not decrease human autonomy but support individuals in making better, informed choices;
  - appropriate human oversight over AI should be achieved through governance mechanisms, e.g. allowing decisions on the use of AI in specific situations, monitoring AI activity, and ensuring levels of human discretion and the ability to override AI decisions; the less oversight that is possible, the more extensive is the need for AI testing;
  - public enforcers must have the mandate and ability to exercise oversight over AI users and outputs.
- **Technical robustness and safety**:
  - AI should be reliable, secure and resilient to attacks, e.g. hacking and manipulation.
  - AI should have safeguards that enable a fall-back plan in case of problems, e.g. involve a human operator.
  - AI should have a level of safety proportionate to the magnitude of the risk posed by its output.
  - AI should be accurate and inform users about the likelihood of possible mistakes.
  - AI should create the same, i.e. "reproducible" results under the same conditions, so that its behaviour can be described.

---

28 E.g. a company could impose higher insurance premiums on people with a higher likelihood of falling ill.

29 E.g. facial recognition technologies can reduce crime (prevent harm), while limiting privacy and individual liberty (i.e. human autonomy).

30 Ethics guidelines for trustworthy AI p. 4-6, p. 15-20.

- **Privacy and data governance**:
  - Individuals must have full control over their data collected for or by AI; data must not be used unlawfully.
  - The quality, integrity and relevance of data fed into an AI system must be ensured to avoid bias and mistakes.
- **Transparency** of data, AI systems and business models must be ensured. This includes:
  - Traceability of AI systems, inter alia by documenting their decisions and the underlying process (including the data used in the analysis and training of the AI machine).
  - Explainability of algorithmic decision-making processes to the extent possible, weighed against a possible reduction in accuracy, and subject to the condition that the AI has "a significant impact on people's lives".
  - Appropriate notification of individuals about the fact that they are interacting with an AI system, about the system's capabilities and limitations (e.g. limited accuracy) and how to opt out and reach a human, where this is necessary to ensure compliance with fundamental rights.
- **Diversity, non-discrimination and fairness**: AI should be "user-centric" so that everybody can use the product or service. Unfair bias e.g. in data sets must be avoided, as it could lead to discrimination
- **Societal and environmental well-being**: AI should be sustainable as well as ecologically and societally friendly.
- **Accountability**: AI should be designed so that it can be audited – without the need to publish intellectual property-related or other proprietary information –, especially when AI applications affect fundamental rights. Negative impacts should be reported and minimised, and adequate redress mechanisms provided.

Together with the ethics guidelines, the EU released a questionnaire. It includes a series of questions on each of the seven requirements which shall offer all AI stakeholders a chance to evaluate whether an AI product complies with the seven requirements and thus also the four ethical principles.

3.2 The OECD ethics guidelines

The trustworthiness of AI is also the key requirement for the OECD. The OECD ethics guidelines therefore aim to develop trustworthy AI. Instead of providing an exhaustive definition – as the EU did – it simply defines trustworthy AI as AI that is compliant with the guidelines. Compared to the EU guidelines, the OECD has developed a much more concise document. While the former is 38 pages long, the latter fits onto one page on which the OECD lists five ethical principles with which AI must comply in order to be considered trustworthy. These five principles are:[31]

- **AI should support inclusive growth**, **sustainable development and well-being**, i.e. pursue beneficial outcomes for people and the planet. These outcomes include augmenting human capabilities, advancing inclusion of underrepresented populations, reducing economic, social, gender and other inequalities, and protecting the environment.
- **AI should support human-centred values and fairness**, i.e.
  - AI actors should respect the rule of law, human rights and democratic values, throughout the AI system lifecycle. These include freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, fairness, social justice, and internationally recognised labour rights.
  - AI actors should then implement mechanisms and safeguards, such as capacity for human determination, that are appropriate to the context and consistent with the state of art.

[31] OECD/LEGAL/0449 section 1.

- **AI should be transparent and explainable**, i.e. AI actors should provide meaningful information, appropriate to the context, and consistent with the state of art:
  - to foster a general understanding of AI systems,
  - to make stakeholders aware of their interactions with AI systems,
  - to enable those affected by an AI system to understand the outcome,
  - to enable those adversely affected by an AI system to challenge its outcome based on information on the factors, and the logic that served as the basis for the decision.
- **AI should be robust, secure and safe**:
  - In conditions of normal use, foreseeable use or misuse, they function appropriately and do not pose unreasonable safety risk.
  - AI actors should ensure traceability, e.g. of datasets, processes and decisions made during the AI system lifecycle, in a way that is appropriate to the context.
  - AI actors should apply a systematic risk management approach to each phase of the AI system lifecycle to address risks related to AI systems, e.g. privacy and digital security.
- **AI should be accountable**, i.e. AI actors should be accountable for the proper functioning of AI systems and for the respect of the above principles.

Although there is an explanation of each principle, the OECD guidelines are very short and generally worded. By contrast with the EU's seven requirements, no further details on the implementation of these five principles is provided and no questionnaire is included in the guidelines. This makes it difficult for AI practitioners, e.g. developers, to assess compliance with the principles.

3.3 Comparison of the EU and OECD ethics guidelines

This section compares the EU and OECD ethics guidelines. As mentioned in Section 3.2, both guidelines aim to create a framework for the development of trustworthy AI. To achieve this goal, both guidelines contain a list of ethical principles that, in the case of the EU, are converted into seven key requirements.

The two sets of guidelines are very similar. The OECD's five principles overlap with the EU's four principles. As shown in Table 2 in the Annex, the seven requirements described in the EU guidelines can be subsumed under the five principles defined by the OECD. No major provision differs in the two sets of guidelines. However, the EU guidelines are clearer as they provide two tools – the requirements and the questionnaire – that can facilitate their implementation. But, even if the EU's guidelines are more precise than those of the OECD, both guidelines remain general and lack practical examples. A considerable degree of discretion is therefore to be expected in the implementation of both sets of guidelines. Nevertheless, the presence of a questionnaire makes compliance and, possibly, future enforcement[32] of the EU guidelines more feasible. The tentative transformation of vague ethical principles into more precise – yet still vague – requirements and steps to be followed is the main difference between the two sets of guidelines.

[32] While the EU and OECD guidelines are by nature non-binding, the European Commission intends to put forward legislation for AI. It is unclear whether there is the intention to take steps toward regulating the use of AI or promoting research and innovation.

4 The future of the EU ethics guidelines

During the G20 Ministerial Meeting on Trade and Digital Economy, held in Tsukuba City (Japan) on 8 and 9 June 2019, ministers from the biggest global economies – among which, most notably, USA and China – agreed on ethical principles for a human-centered approach to AI. The principles are an Annex to the G20 Ministerial Statement, endorsed in the final G20 Declaration published on 29 June 2019[33], and are non-binding in nature[34]. These principles are "drawn" from the OECD guidelines[35], meaning they are a complete endorsement of the OECD's recommendation. As a result, it appears that the USA, China, and the EU[36] are accepting the OECD guidelines.

Consensus, however, is a long way off. The same paragraph of the Declaration stresses the importance of regulatory approaches that are agile and flexible, "including through the use of regulatory sandboxes"[37]. Sandboxes are by nature national temporary solutions to circumvent existing standards. In future, therefore, it is possible that, instead of finding common, robust ethics standards for AI, even national legislation, e.g. on data protection, could be halted in order to foster AI development.

The endorsement of non-binding ethics standards that are broadly defined, difficult to implement, and can be circumvented via sandboxes, is not a leap forward in establishing the global regulation of AI.

In the global context, therefore, the EU has produced the more precise and practical guidelines. Commission president-elect Ursula von der Leyen announced legislative action on AI. At this stage, it is not clear what legislative action for AI the European Commission intends to push forward. If the guidelines were to be made binding, a lot of work would be needed to reduce vagueness and clear ambiguities. Such a task would be lengthy and entail a lot of uncertainty, given the fast pace at which AI is developing. Furthermore, it is hard to imagine that China and the USA will accept EU legislation setting standards on AI.

Therefore, as the EU is currently lagging behind in the development and deployment of AI, binding AI guidelines would further increase the EU's competitive disadvantage.

Given the similarity of the EU and OECD guidelines, as well as the OECD's plan to release a practical guide on the implementation of their ethical principles, the EU can play an important role in shaping the future of global standards for AI. This would be the case if the OECD's practical guide were to be based on the EU's key requirements and its questionnaire which is not to be ruled out considering that some major corporations[38] have already pledged to implement the EU ethics guidelines.

---

33 https://g20.org/pdf/documents/en/FINAL_G20_Osaka_Leaders_Declaration.pdf .
34 paragraph 12 final declaration.
35 Ibid.
36 Jean-Claude Juncker represents the European Commission at the G20 Summit.
37 Ibid.
38 https://www.ibm.com/blogs/policy/ai-ethics-eu/ IBM for example will apply them from now on.

5 Annex

Tab. 2: Comparison of OECD's ethical principles and EU's key requirements

OECD principles	Corresponding EU key requirements
Inclusive and sustainable growth and well being Trustworthy AI in pursuit of beneficial outcomes for people and the planet, reducing inequality, protecting natural resources, invigorating inclusive growth and sustainable development.	**Societal and environmental well being** Sustainability and ecological responsibility of AI systems should be encouraged, and research should be fostered into AI solutions addressing areas of global concern, such as for instance the Sustainable Development Goals.
Human centred values and fairness AI actors should respect rule of law, human rights and democratic values. These include freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, fairness, social justice, and labour rights.	**Human agency and oversight** Given the reach and capacity of AI systems, they can also negatively affect fundamental rights. In situations where such risks exist, a fundamental rights impact assessment should be undertaken. This should be done prior to the system's development and include an evaluation of whether those risks can be reduced or justified as necessary in a democratic society in order to respect the rights and freedoms of others.
	Privacy and data governance AI systems must guarantee privacy and data protection throughout a system's entire lifecycle. This includes the information initially provided by the user, as well as the information generated about the user over the course of their interaction with the system (e.g. outputs that the AI system generated for specific users or how users responded to particular recommendations). Digital records of human behaviour may allow AI systems to infer not only individuals' preferences, but also their sexual orientation, age, gender, religious or political views. To allow individuals to trust the data gathering process, it must be ensured that data collected about them will not be used to unlawfully or unfairly discriminate against them.
	Diversity, non-discrimination, and fairness Identifiable and discriminatory bias should be removed in the collection phase where possible. The way in which AI systems are developed (e.g. algorithms' programming) may also suffer from unfair bias. This could be counteracted by putting in place oversight processes to analyse and address the system's purpose, constraints, requirements and decisions in a clear and transparent manner.
	Societal and environmental well-being Beyond assessing the impact of an AI system's development, deployment and use on individuals, this impact should also be assessed from a societal perspective, taking into account its effect on institutions, democracy and society at large. The use of AI systems should be given careful consideration particularly in situations relating to the democratic process, including not only political decision-making but also electoral contexts.
Human centred values and fairness AI actors should implements mechanisms and safeguards, such as capacity for human determination, appropriate to the context and consistent with the state of art.	**Accountability** Both the ability to report on actions or decisions that contribute to a certain system outcome, and to respond to the consequences of such an outcome, must be ensured. Identifying, assessing, documenting and minimising the potential negative impacts of AI systems is especially crucial for those (in)directly affected.

	Technical robustness and safety AI systems should have safeguards that enable a fallback plan in case of problems. This can mean that AI systems switch from a statistical to rule-based procedure, or that they ask for a human operator before continuing their action.
	Human agency and oversight Users should be able to make informed autonomous decisions regarding AI systems. They should be given the knowledge and tools to comprehend and interact with AI systems to a satisfactory degree and, where possible, be enabled to reasonably self-assess or challenge the system. (..)Human oversight helps ensuring that an AI system does not undermine human autonomy or causes other adverse effects. (..)This can include the decision not to use an AI system in a particular situation, to establish levels of human discretion during the use of the system, or to ensure the ability to override a decision made by a system. Moreover, it must be ensured that public enforcers have the ability to exercise oversight in line with their mandate.
Transparency and explainability AI actors should commit to responsible disclosure regarding AI systems. They should provide meaningful information to foster general understanding of AI system.	**Transparency** Explainability concerns the ability to explain both the technical processes of an AI system and the related human decisions (e.g. application areas of a system). Technical explainability requires that the decisions made by an AI system can be understood and traced by human beings. Moreover, trade-offs might have to be made between enhancing a system's explainability (which may reduce its accuracy) or increasing its accuracy (at the cost of explainability). Whenever an AI system has a significant impact on people's lives, it should be possible to demand a suitable explanation of the AI system's decision-making process. Such explanation should be timely and adapted to the expertise of the stakeholder concerned (e.g. layperson, regulator or researcher). In addition, explanations of the degree to which an AI system influences and shapes the organisational decision-making process, design choices of the system, and the rationale for deploying it, should be available (hence ensuring business model transparency).
Transparency and explainability Provide meaningful information to make stakeholders aware of their interactions with AI systems.	**Transparency** AI systems should not represent themselves as humans to users; humans have the right to be informed that they are interacting with an AI system. This entails that AI systems must be identifiable as such. In addition, the option to decide against this interaction in favour of human interaction should be provided where needed to ensure compliance with fundamental rights.
Transparency and explainability Provide meaningful information to those affected by AI to understand its outcome.	**Transparency** The AI system's capabilities and limitations should be communicated to AI practitioners or end-users in a manner appropriate to the use case at hand. This could encompass communication of the AI system's level of accuracy, as well as its limitations.
Transparency and explainability Provide meaningful information to enable those adversely affected by AI to challenge its outcome.	**Human agency and oversight** Users should be able to make informed autonomous decisions regarding AI systems. They should be given the knowledge and tools to comprehend and interact with AI systems to a satisfactory degree and, where possible, be enabled to reasonably self-assess or challenge the system. AI systems should support individuals in making better,

	more informed choices in accordance with their goals. AI systems can sometimes be deployed to shape and influence human behaviour through mechanisms that may be difficult to detect, since they may harness sub-conscious processes, including various forms of unfair manipulation, deception, herding and conditioning, all of which may threaten individual autonomy. The overall principle of user autonomy must be central to the system's functionality.
	Accountability When unjust adverse impact occurs, accessible mechanisms should be foreseen that ensure adequate redress. Knowing that redress is possible when things go wrong is key to ensure trust. Particular attention should be paid to vulnerable persons or groups.
Robustness and safety AI systems should be robust, safe and secure throughout the life cycle so that they do not pose unreasonable safety risk in condition of normal use and foreseeable use or misuse.	**Technical robustness and safety** It must be ensured that the system will do what it is supposed to do without harming living beings or the environment. This includes the minimisation of unintended consequences and errors. In addition, processes to clarify and assess potential risks associated with the use of AI systems, across various application areas, should be established. The level of safety measures required depends on the magnitude of the risk posed by an AI system, which in turn depends on the system's capabilities. Where it can be foreseen that the development process or the system itself will pose particularly high risks, it is crucial for safety measures to be developed and tested proactively.
Robustness and safety AI actors should ensure traceability, including in relation to datasets, processes and decisions made, to enable analysis of outcomes and responses to inquiry, appropriate to the context and consistent with the state of art.	**Technical robustness and safety** It is critical that the results of AI systems are reproducible, as well as reliable. A reliable AI system is one that works properly with a range of inputs and in a range of situations. This is needed to scrutinise an AI system and to prevent unintended harms. Reproducibility describes whether an AI experiment exhibits the same behaviour when repeated under the same conditions. This enables scientists and policy makers to accurately describe what AI systems do. Replication files can facilitate the process of testing and reproducing behaviours.
	Transparency The data sets and the processes that yield the AI system's decision, including those of data gathering and data labelling as well as the algorithms used, should be documented to the best possible standard to allow for traceability and an increase in transparency. This also applies to the decisions made by the AI system. This enables identification of the reasons why an AI-decision was erroneous which, in turn, could help prevent future mistakes. Traceability facilitates auditability as well as explainability.
	Accountability Auditability entails the enablement of the assessment of algorithms, data and design processes. This does not necessarily imply that information about business models and intellectual property related to the AI system must always be openly available. Evaluation by internal and external auditors, and the availability of such evaluation reports, can contribute to the trustworthiness of the technology. In applications affecting fundamental rights, including safety-critical applications, AI systems should be able to be independently audited.

<table>
<tr><td rowspan="2">Robustness and safety
AI actors should, based on context, role and capabilities, apply systemic risk management approach to each phase of AI system lifecycle, to address risks related to AI.</td><td>Human agency and oversight
Given the reach and capacity of AI systems, they can also negatively affect fundamental rights. In situations where such risks exist, a fundamental rights impact assessment should be undertaken. This should be done prior to the system’s development and include an evaluation of whether those risks can be reduced or justified as necessary in a democratic society in order to respect the rights and freedoms of others. Moreover, mechanisms should be put into place to receive external feedback</td></tr>
<tr><td>Accountability
Auditability entails the enablement of the assessment of algorithms, data and design processes. This does not necessarily imply that information about business models and intellectual property related to the AI system must always be openly available. Evaluation by internal and external auditors, and the availability of such evaluation reports, can contribute to the trustworthiness of the technology. In applications affecting fundamental rights, including safety-critical applications, AI systems should be able to be independently audited.</td></tr>
<tr><td rowspan="2">Accountability
AI actors should be accountable for the proper functioning of the AI system and for respect of the above principles.</td><td>Accountability
The requirement of accountability complements the above requirements, and is closely linked to the principle of fairness. It necessitates that mechanisms be put in place to ensure responsibility and accountability for AI systems and their outcomes, both before and after their development, deployment and use.</td></tr>
<tr><td>Human agency and oversight
It must be ensured that public enforcers have the ability to exercise oversight in line with their mandate. Oversight mechanisms can be required in varying degrees to support other safety and control measures, depending on the AI system’s application area and potential risk. All other things being equal, the less oversight a human can exercise over an AI system, the more extensive testing and stricter governance is required.</td></tr>
</table>

Recently published in this series:

06/2019: Gas Supply in the EU (July 2019)

05/2019: Road Safety Management (July 2019)

04/2019: Internal Electricity Market (June 2019)

03/2019: The EU Green Bond Standard (June 2019)

02/2019: Energy Union Governance (May 2019)

01/2019: Renewable Energy in the EU (April 2019)

05/2018: Energy Efficiency Policy (December 2018)

04/2018: Climate Protection outside the EU ETS (August 2018)

03/2018: Climate Protection by way of the EU ETS (July 2018)

02/2018: French Vocational Training (February 2018)

The Author:

Alessandro Gasparotti is policy analyst in the Internal Market department at the Centre for European policy.

cep | Centrum für Europäische Politik
Kaiser-Joseph-Straße 266 | D-79098 Freiburg
Telefon +49 761 38693-0 | www.cep.eu

cep is the European-policy think tank of the non-profit-making foundation Stiftung Ordnungspolitik. It is an independent centre of excellence for the examination, analysis and evaluation of EU policy.