EUROPEAN
COMMISSION

Brussels, 8.4.2019
COM(2019) 168 final

COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS

Building Trust in Human-Centric Artificial Intelligence

1. INTRODUCTION — THE EUROPEAN AI STRATEGY

Artificial intelligence (AI) has the potential to transform our world for the better: it can improve healthcare, reduce energy consumption, make cars safer, and enable farmers to use water and natural resources more efficiently. AI can be used to predict environmental and climate change, improve financial risk management and provides the tools to manufacture, with less waste, products tailored to our needs. AI can also help to detect fraud and cybersecurity threats, and enables law enforcement agencies to fight crime more efficiently.

AI can benefit the whole of society and the economy. It is a strategic technology that is now being developed and used at a rapid pace across the world. Nevertheless, AI also brings with it new challenges for the future of work, and raises legal and ethical questions.

To address these challenges and make the most of the opportunities which AI offers, the Commission published a European strategy[1] **in April 2018.** The strategy places people at the centre of the development of AI — **human-centric AI**. It is a three-pronged approach to boost the EU's technological and industrial capacity and AI uptake across the economy, prepare for socio-economic changes, and ensure an appropriate ethical and legal framework.

To deliver on the AI strategy, **the Commission developed together with Member States a coordinated plan on AI**[2], which it presented in December 2018, to create synergies, pool data — the raw material for many AI applications — and increase joint investments. The aim is to foster cross-border cooperation and mobilise all players to increase public and private investments to **at least EUR 20 billion** annually over the next decade[3]. The Commission doubled its investments in AI in Horizon 2020 and plans to invest EUR 1 billion annually from Horizon Europe and the Digital Europe Programme, in support notably of common data spaces in health, transport and manufacturing, and large experimentation facilities such as smart hospitals and infrastructures for automated vehicles and a strategic research agenda.

To implement such a common strategic research, innovation and deployment agenda the Commission has intensified its **dialogue with all relevant stakeholders** from industry, research institutes and public authorities. The new Digital Europe programme will also be crucial in helping to make AI available to small and medium-size enterprises across all Member States through digital innovation hubs, strengthened testing and experimentation facilities, data spaces and training programmes.

Building on its reputation for safe and high-quality products, Europe's ethical approach to AI strengthens citizens' trust in the digital development and aims at building a competitive advantage for European AI companies. The purpose of this Communication is to launch a comprehensive piloting phase involving stakeholders on the widest scale in order to test the practical implementation of ethical guidance for AI development and use.

2. BUILDING TRUST IN HUMAN-CENTRIC AI

The European AI strategy and the coordinated plan make clear that **trust is a prerequisite to ensure a human-centric approach to AI**: AI is not an end in itself, but a tool that has to serve people with the ultimate aim of increasing human well-being. To achieve this, the

[1] COM(2018) 237.

[2] COM(2018) 795.

[3] To help reach this goal, the Commission proposed, under the next programming period 2021-2027, that the Union allocates at least EUR 1 billion per year in funding from the Horizon Europe and Digital Europe programmes to invest in AI.

trustworthiness of AI should be ensured. The values on which our societies are based need to be fully integrated in the way AI develops.

The Union is founded on **the values of respect for human dignity, freedom, democracy, equality, the rule of law and respect for human rights**, including the rights of persons belonging to minorities[4]. These values are common to the societies of all Member States in which pluralism, non-discrimination, tolerance, justice, solidarity and equality prevail. In addition, the **EU Charter of Fundamental Rights** brings together – in a single text – the personal, civic, political, economic and social rights enjoyed by people within the EU.

The EU has a **strong regulatory framework** that will set the global standard for human-centric AI. The General Data Protection Regulation ensures a high standard of protection of personal data, and requires the implementation of measures to ensure data protection by design and by default[5]. The Free Flow of Non-Personal Data Regulation removes barriers to the free movement of non-personal data and ensures the processing of all categories of data anywhere in Europe. The recently adopted Cybersecurity Act will help to strengthen trust in the online world, and the proposed ePrivacy Regulation[6] also aims at this goal.

Nevertheless, AI brings new challenges because it enables machines to "learn" and to take and implement decisions without human intervention. Before long, this kind of functionality will become standard in many types of goods and services, from smart phones to automated cars, robots and online applications. Yet, decisions taken by algorithms could result from data that is incomplete and therefore not reliable, they may be tampered with by cyber-attackers, or they may be biased or simply mistaken. Unreflectively applying the technology as it develops would therefore lead to problematic outcomes as well as reluctance by citizens to accept or use it.

Instead, AI technology should be developed in a way that puts people at its centre and is thus worthy of the public's trust. This implies that AI applications should not only be consistent with the law, but also adhere to ethical principles and ensure that their implementations avoid unintended harm. Diversity in terms of gender, racial or ethnic origin, religion or belief, disability and age should be ensured at every stage of AI development. AI applications should empower citizens and respect their fundamental rights. They should aim to enhance people's abilities, not replace them, and also enable access by people with disabilities.

Therefore, there is a need for **ethics guidelines** that build on the existing regulatory framework and that should be applied by developers, suppliers and users of AI in the internal market, establishing an ethical level playing field across all Member States. This is why the Commission has set up a **high-level expert group on AI**[7] representing a wide range of stakeholders and has tasked it with drafting AI ethics guidelines as well as preparing a set of recommendations for broader AI policy. At the same time, the **European AI Alliance**[8], an open multi-stakeholder platform with over 2700 members, was set up to provide broader input for the work of the AI high-level expert group.

[4] In addition, the EU is a party to the UN Convention on the Rights of persons with disabilities.

[5] Regulation (EU) 2016/679. The General Data Protection Regulation (GDPR) guarantees the free flow of personal data within the Union. It contains provisions on decision-making based solely on automated processing, including profiling. The individuals concerned have the right to be informed about the existence of automated-decision making and to receive meaningful information about the logic involved in the automated decision-making and about the significance and envisaged consequences of the processing for them. They also have the right in such cases to obtain human intervention, to express their point of view and to contest the decision.

[6] COM(2017) 10.

[7] https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence

[8] https://ec.europa.eu/digital-single-market/en/european-ai-alliance

The AI high-level expert group published a first draft of the ethics guidelines in December 2018. Following a **stakeholder consultation**[9] and **meetings with representatives from Member States**[10], the AI expert group has delivered a revised document to the Commission in March 2019. In their feedback so far, stakeholders overall have welcomed the practical nature of the guidelines and the concrete guidance they offer to developers, suppliers and users of AI on how to ensure trustworthiness.

2.1. Guidelines for trustworthy AI drafted by the AI high-level expert group

The guidelines drafted by the AI high-level expert group, to which this Communication refers[11], build in particular on the work done by the European Group on Ethics in Science and New Technologies and the Fundamental Rights Agency.

The guidelines postulate that in order to achieve 'trustworthy AI', three components are necessary: (1) it should comply with the law, (2) it should fulfil ethical principles and (3) it should be robust.

Based on these three components and the European values set out in section 2, the guidelines identify seven key requirements that AI applications should respect to be considered trustworthy. The guidelines also include an assessment list to help check whether these requirements are fulfilled.

The seven key requirements are:

- Human agency and oversight
- Technical robustness and safety
- Privacy and data governance
- Transparency
- Diversity, non-discrimination and fairness
- Societal and environmental well-being
- Accountability

While these requirements are intended to apply to all AI systems in different settings and industries, the specific context in which they are applied should be taken into account for their concrete and proportionate implementation, taking an impact-based approach. For illustration, AI application suggesting an unsuitable book to read is much less perilous than misdiagnosing a cancer and could therefore be subject to less stringent supervision.

The guidelines drafted by the AI high-level expert group are non-binding and as such do not create any new legal obligations. However, many existing (and often use- or domain-specific)

[9] This consultation resulted in comments from 511 organisations, associations, companies, research institutes, individuals and others. A summary of the feedback received is available at: https://ec.europa.eu/futurium/en/system/files/ged/consultation_feedback_on_draft_ai_ethics_guidelines_4.pdf

[10] The work of the expert group was positively received by Member States, with the Council conclusions adopted on 18 February 2019 *inter alia* taking note of the forthcoming publication of the ethics guidelines and supporting the Commission's effort to bring an EU ethical approach to the global stage: https://data.consilium.europa.eu/doc/document/ST-6177-2019-INIT/en/pdf

[11] https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top

provisions of Union law of course already reflect one or several of these key requirements, for example safety, personal data protection, privacy or environmental protection rules.

The Commission welcomes the work of the AI high-level expert group and considers it valuable input for its policy-making.

2.2. Key requirements for trustworthy AI

The Commission supports the following key requirements for trustworthy AI, which are based on European values. It encourages stakeholders to apply the requirements and to test the assessment list that operationalises them in order to create the right environment of trust for the successful development and use of AI. The Commission welcomes feedback from stakeholders to evaluate whether this assessment list provided in the guidelines requires further adjustment.

I. Human agency and oversight

AI systems should support individuals in making better, more informed choices in accordance with their goals. They should act as enablers to a flourishing and equitable society by supporting human agency and **fundamental rights**, and not decrease, limit or misguide human autonomy. The overall **wellbeing of the user** should be central to the system's functionality.

Human oversight helps ensuring that an AI system does not undermine human autonomy or causes other adverse effects. Depending on the specific AI-based system and its application area, the appropriate degrees of **control measures**, including the adaptability, accuracy and explainability of AI-based systems, should be ensured[12]. **Oversight** may be achieved through governance mechanisms such as ensuring a human-in-the-loop, human-on-the-loop, or human-in-command approach.[13] It must be ensured that public authorities have the ability to exercise their oversight powers in line with their mandates. 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.

II. Technical robustness and safety

Trustworthy AI requires algorithms to be secure, reliable and robust enough to deal with errors or inconsistencies during all life cycle phases of the AI system, and to adequately cope with erroneous outcomes. AI systems need to be **reliable**, secure enough to be

[12] The General Data Protection Regulation gives individuals the right not to be subject to a decision based solely on automated processing when this produces legal effects on users or similarly significantly affects them (Article 22 GDPR).

[13] Human-in-the-loop (HITL) refers to the human intervention in every decision cycle of the system, which in many cases is neither possible nor desirable. Human-on-the-loop (HOTL) refers to the capability for human intervention during the design cycle of the system and monitoring the system’s operation. Human-in-command (HIC) refers to the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impact) and the ability to decide when and how to use the system in any particular situation. 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 the system.

resilient against both overt attacks and more subtle attempts to manipulate data or algorithms themselves, and they must ensure a **fall-back plan** in case of problems. Their decisions must be **accurate**, or at least correctly reflect their level of accuracy, and their outcomes should be **reproducible**.

In addition, AI systems should integrate safety and security-by-design mechanisms to ensure that they are **verifiably safe** at every step, taking at heart the physical and mental safety of all concerned. This includes the minimisation and where possible the reversibility of unintended consequences or errors in the system's operation. Processes to clarify and assess potential risks associated with the use of AI systems, across various application areas, should be put in place.

III. Privacy and Data Governance

Privacy and **data protection** must be guaranteed at **all stages** of the AI system's life cycle. Digital records of human behaviour may allow AI systems to infer not only individuals' preferences, age and gender but also their sexual orientation, religious or political views. To allow individuals to trust the data processing, it must be ensured that they have full control over their own data, and that data concerning them will not be used to harm or discriminate against them.

In addition to safeguarding privacy and personal data, requirements must be fulfilled to ensure high quality AI systems. The quality of the data sets used is paramount to the performance of AI systems. When data is gathered, it may reflect socially constructed biases, or contain inaccuracies, errors and mistakes. This needs to be addressed prior to training an AI system with any given data set. In addition, the **integrity** of the data must be ensured. Processes and data sets used must be tested and documented at each step such as planning, training, testing and deployment. This should also apply to AI systems that were not developed in-house but acquired elsewhere. Finally, the **access** to data must be adequately governed and controlled.

IV. Transparency

The **traceability** of AI systems should be ensured; it is important to log and document both the decisions made by the systems, as well as the entire process (including a description of data gathering and labelling, and a description of the algorithm used) that yielded the decisions. Linked to this, **explainability** of the algorithmic decision-making process, adapted to the persons involved, should be provided to the extent possible. Ongoing research to develop explainability mechanisms should be pursued. In addition, explanations of the degree to which an AI system influences and shapes the organisational decision-making process, design choices of the system, as well as the rationale for deploying it, should be available (hence ensuring not just data and system transparency, but also business model transparency).

Finally, it is important to adequately **communicate** the AI system's capabilities and limitations to the different stakeholders involved in a manner appropriate to the use case at hand. Moreover, AI systems should be identifiable as such, ensuring that users know they are interacting with an AI system and which persons are responsible for it.

V. Diversity, non-discrimination and fairness

Data sets used by AI systems (both for training and operation) may suffer from the inclusion of inadvertent historic bias, incompleteness and bad governance models. The continuation of such biases could lead to (in)direct discrimination. Harm can also result from the intentional exploitation of (consumer) biases or by engaging in unfair competition. Moreover, the way in which AI systems are developed (e.g. the way in which the programming code of an algorithm is written) may also suffer from bias. Such concerns should be tackled from the beginning of the system' development.

Establishing **diverse design teams** and setting up mechanisms ensuring **participation,** in particular of citizens, in AI development can also help to address these concerns. It is advisable to consult stakeholders who may directly or indirectly be affected by the system throughout its life cycle. AI systems should consider the whole range of human abilities, skills and requirements, and ensure accessibility through a universal design approach to strive to achieve equal access for persons with disabilities.

VI. Societal and environmental well-being

For AI to be trustworthy, its impact on the **environment and other sentient beings** should be taken into account. Ideally, all humans, including future generations, should benefit from biodiversity and a habitable environment. Sustainability and **ecological responsibility** of AI systems should hence be encouraged. The same applies to AI solutions addressing areas of global concern, such as for instance the UN Sustainable Development Goals.

Furthermore, the impact of AI systems should be considered not only from an individual perspective, but also from the perspective of **society as a whole**. The use of AI systems should be given careful consideration particularly in situations relating to the democratic process, including opinion-formation, political decision-making or electoral contexts. Moreover, AI's **social impact** should be considered. While AI systems can be used to enhance social skills, they can equally contribute to their deterioration.

VII. Accountability

Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes, both before and after their implementation. **Auditability** of AI systems is key in this regard, as the assessment of AI systems by internal and external auditors, and the availability of such evaluation reports, strongly contributes to the trustworthiness of the technology. External auditability should especially be ensured in applications affecting fundamental rights, including safety-critical applications.

Potential negative impacts of AI systems should be identified, assessed, documented and minimised. The use of impact assessments facilitates this process. These assessments should be proportionate to the extent of the risks that the AI systems pose. **Trade-offs** between the requirements – which are often unavoidable – should be addressed in a rational and methodological manner, and should be accounted for. Finally, when unjust adverse impact occurs, accessible mechanisms should be foreseen that ensure **adequate redress**.

2.3. Next steps: a piloting phase involving stakeholders on the widest scale

Reaching consensus on these key requirements for AI systems is a first important milestone towards guidelines for ethical AI. As a next step, the Commission will ensure that this guidance can be tested and implemented in practice.

To this end, the Commission will now launch a targeted piloting phase designed to obtain structured feedback from stakeholders. This exercise will focus in particular on the assessment list which the high-level expert group has drawn up for each of the key requirements.

This work will have two strands: (i) a piloting phase for the guidelines involving stakeholders who develop or use AI, including public administrations, and (ii) a continued stakeholder consultation and awareness-raising process across Member States and different groups of stakeholders, including industry and service sectors:

(i) Starting in June 2019, all stakeholders and individuals will be invited to test the assessment list and provide feedback on how to improve it. In addition, the AI high-level expert group will set up an in-depth review with stakeholders from the private and the public sector to gather more detailed feedback on how the guidelines can be implemented in a wide range of application domains. All feedback on the guidelines' workability and feasibility will be evaluated by the end of 2019.

(ii) In parallel, the Commission will organise further outreach activities, giving representatives of the AI high-level expert group the opportunity to present the guidelines to relevant stakeholders in the Member States, including industry and service sectors, and providing these stakeholders with an additional opportunity to comment on and contribute to the AI guidelines.

The Commission will take into account the work of the group of experts on ethics for connected and automated driving[14] and work with EU-funded research projects on AI and with relevant public-private partnerships on implementing the key requirements[15]. For example, the Commission will support, in coordination with Member States, the development of a common database of health images initially dedicated to the most common forms of cancer, so that algorithms can be trained to diagnose symptoms with very high accuracy. Similarly, the cooperation of the Commission and Member States enables an increasing number of cross-border corridors for testing connected and automated vehicles. The guidelines should be applied in these projects and tested, and the results will feed into the evaluation process.

The piloting phase and the stakeholder consultation will benefit from the contribution of the European AI Alliance and AI4EU, the AI on-demand platform. The AI4EU project[16], launched in January 2019, brings together algorithms, tools, datasets and services to help organisations, in particular small and medium-size enterprises, to implement AI solutions. The European AI Alliance, together with AI4EU, will continue to mobilise the AI ecosystem across Europe, also in view of piloting the AI ethics guidelines and promoting the respect for human-centric AI.

At the beginning of 2020, building on the evaluation of feedback received during the piloting phase, **the AI high-level expert group will review and update the guidelines**. Based on the

[14] See the Commission's Communication on connected and automated mobility, COM(2018) 283.

[15] In the Framework of the European Defence Fund, the Commission will also develop specific ethical guidance for the evaluation of project proposals in the area of AI for defence.

[16] https://ec.europa.eu/digital-single-market/en/news/artificial-intelligence-ai4eu-project-launches-1-january-2019

review and on the experience acquired, **the Commission will evaluate the outcome and propose any next steps.**

Ethical AI is a win-win proposition. Guaranteeing the respect for fundamental values and rights is not only essential in itself, it also facilitates acceptance by the public and increases the competitive advantage of European AI companies by establishing a brand of human-centric, trustworthy AI known for ethical and secure products. This builds more generally on the strong reputation of European companies for providing safe and secure products of high quality. The pilot phase will help to ensure that AI products fulfil this promise.

2.4. Towards international AI ethics guidelines

International discussions on AI ethics have intensified after Japan's G7 Presidency put the topic high on the agenda in 2016. Given the international interlinkages of AI development in terms of data circulation, algorithmic development and research investments, **the Commission will continue its efforts to bring the Union's approach to the global stage and build a consensus on a human-centric AI**[17].

The work done by the AI high-level expert group, and more specifically the list of requirements and the engagement process with stakeholders, provides the Commission with additional valuable input for contributing to the international discussions. The European Union can have a leadership role in developing international AI guidelines and, if possible, a related assessment mechanism.

Therefore, the Commission will:

Strengthen cooperation with like-minded partners:

- exploring the extent to which convergence can be achieved with third countries' draft ethics guidelines (e.g. Japan, Canada, Singapore) and, building on this group of like-minded countries, to prepare for a broader discussion, supported by actions implementing the Partnership Instrument for cooperation with Third Countries[18]; and
- exploring how companies from non-EU countries and international organisations can contribute to the 'pilot phase' of the guidelines through testing and validation.

Continue to play an active role in international discussions and initiatives:

- contributing to multilateral fora such as the G7 and G20;
- engaging in dialogues with non-EU countries and organising bilateral and multilateral meetings to build a consensus on human-centric AI;

[17] The High Representative of the Union for Foreign Affairs and Security Policy will, with the support of the Commission, build on consultations in the United Nations, the Global Tech Panel, and other multilateral fora, and in particular coordinate proposals for addressing the complex security challenges involved.

[18] Regulation (EU) No 234/2014 of the European Parliament and of the Council of 11 March 2014 establishing a Partnership Instrument for cooperation with third countries (OJ L 77, 15.3.2014, p. 77). For instance the planned project on 'An international alliance for a human-centric approach to artificial intelligence' will facilitate joint initiatives with like-minded partners, in order to promote ethical guidelines and to adopt common principles and operational conclusions. It will enable the EU and like-minded countries to discuss operational conclusions resulting from the ethical guidelines on AI proposed by the expert group in order to reach a common approach. Moreover, it will provide for monitoring the uptake of AI technology globally. Finally, the project plans to organise public diplomacy activities accompanying international events e.g. by the G7, G20 and the Organisation for Economic Cooperation and Development.

- contributing to relevant standardisation activities in international standards development organisations to promote this vision; and
- strengthening the collection and diffusion of insights on public policies, working jointly with relevant international organisations.

3. CONCLUSIONS

The EU is founded on a set of fundamental values and has constructed a strong and balanced regulatory framework on these foundations. Building on this existing regulatory framework, there is a need for ethics guidelines for the development and use of AI due to its novelty and the specific challenges this technology brings. Only if AI is developed and used in a way that respects widely-shared ethical values, it can be considered trustworthy.

With a view to this objective, the Commission welcomes the input prepared by the AI high-level expert group. Based on the key requirements for AI to be considered trustworthy, the Commission will now launch a targeted piloting phase to ensure that the resulting ethical guidelines for AI development and use can be implemented in practice. The Commission will also work to forge a broad societal consensus on human-centric AI, including with all involved stakeholders and our international partners.

The ethical dimension of AI is not a luxury feature or an add-on: it needs to be an integral part of AI development. By striving towards human-centric AI based on trust, we safeguard the respect for our core societal values and carve out a distinctive trademark for Europe and its industry as a leader in cutting-edge AI that can be trusted throughout the world.

To ensure the ethical development of AI in Europe in its wider context, the Commission is pursuing a comprehensive approach including in particular the following lines of action to be implemented by the third quarter of 2019:

- It will start launching a set of **networks of AI research excellence** centres through Horizon 2020. It will select up to four networks, focusing on scientific or technological major challenges such as explainability and advanced human-machine interaction, which are key ingredients for trustworthy AI.
- It will begin setting up **networks of digital innovation hubs**[19] focussing on AI in manufacturing and on big data.
- Together with Member States and stakeholders, the Commission will start preparatory discussions to develop and implement **a model for data sharing and making best use of common data spaces,** with a focus notably on transport, healthcare and industrial manufacturing.[20]

In addition, the Commission is working on a report on the challenges posed by AI to the safety and liability frameworks and a guidance document on the implementation of the Product Liability Directive[21]. At the same time, the European High-Performance Computing Joint Undertaking (EuroHPC)[22] will develop the next generation of supercomputers because

[19] http://s3platform.jrc.ec.europa.eu/digital-innovation-hubs

[20] The necessary resources will be mobilised from Horizon 2020 (under which close to 1.5 billion EUR are dedicated to AI for the period 2018-2020) and its planned successor Horizon Europe, the Digital part of the Connecting Europe Facility and especially the future Digital Europe Programme. Projects will also draw on resources from the private sector and Member State programmes.

[21] See the Commission's Communication Artificial Intelligence for Europe, COM (2018) 237.

[22] https://eurohpc-ju.europa.eu

computing capacity is essential for processing data and training AI and Europe needs to master the full digital value chain. The ongoing partnership with Member States and industry on microelectronic components and systems (ECSEL)[23] as well as the European Processor Initiative[24] will contribute to the development of low-power processor technology for trustworthy and secure high-performance and edge computing.

Just like the work on ethical guidelines for AI, all these initiatives build on **close cooperation of all concerned stakeholders**, Member States, industry, societal actors and citizens. Overall, Europe's approach to Artificial Intelligence shows how economic competitiveness and societal trust must start from the same fundamental values and mutually reinforce each other.

[23] www.ecsel.eu

[24] www.european-processor-initiative.eu