European Commission

INDEPENDENT

HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE

SET UP BY THE EUROPEAN COMMISSION

ETHICS GUIDELINES FOR TRUSTWORTHY AI

ETHICS GUIDELINES FOR TRUSTWORTHY AI

High-Level Expert Group on Artificial Intelligence

This document was written by the High-Level Expert Group on AI (AI HLEG). The members of the AI HLEG named in this document support the overall framework for Trustworthy AI put forward in these Guidelines, although they do not necessarily agree with every single statement in the document.

The Trustworthy AI assessment list presented in Chapter III of this document will undergo a piloting phase by stakeholders to gather practical feedback. A revised version of the assessment list, taking into account the feedback gathered through the piloting phase, will be presented to the European Commission in early 2020.

The AI HLEG is an independent expert group that was set up by the European Commission in June 2018.

Contact Nathalie Smuha - AI HLEG Coordinator
E-mail CNECT-HLG-AI@ec.europa.eu

European Commission
B-1049 Brussels

Document made public on 8 April 2019.

A first draft of this document was released on 18 December 2018 and was subject to an open consultation which generated feedback from more than 500 contributors. We wish to explicitly and warmly thank all those who contributed their feedback on the document's first draft, which was considered in the preparation of this revised version.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The aim of the Guidelines is to promote Trustworthy AI. Trustworthy AI has **three components**, which should be met throughout the system's entire life cycle: (1) it should be **lawful**, complying with all applicable laws and regulations (2) it should be **ethical**, ensuring adherence to ethical principles and values and (3) it should be **robust**, both from a technical and social perspective since, even with good intentions, AI systems can cause unintentional harm. Each component in itself is necessary but not sufficient for the achievement of Trustworthy AI. Ideally, all three components work in harmony and overlap in their operation. If, in practice, tensions arise between these components, society should endeavour to align them.

These Guidelines set out a **framework for achieving Trustworthy AI.** The framework does not explicitly deal with Trustworthy AI's first component (lawful AI).[1] Instead, it aims to offer guidance on the second and third components: fostering and securing ethical and robust AI. Addressed to all stakeholders, these Guidelines seek to go beyond a list of ethical principles, by providing guidance on how such principles can be operationalised in socio-technical systems. Guidance is provided in three layers of abstraction, from the most abstract in Chapter I to the most concrete in Chapter III, closing with examples of opportunities and critical concerns raised by AI systems.

I. Based on an approach founded on fundamental rights, **Chapter I** identifies the **ethical principles** and their correlated values that must be respected in the development, deployment and use of AI systems.

Key guidance derived from Chapter I:

- ✓ Develop, deploy and use AI systems in a way that adheres to the ethical principles of: *respect for human autonomy, prevention of harm, fairness and explicability.* Acknowledge and address the potential tensions between these principles.
- ✓ Pay particular attention to situations involving more vulnerable groups such as children, persons with disabilities and others that have historically been disadvantaged or are at risk of exclusion, and to situations which are characterised by asymmetries of power or information, such as between employers and workers, or between businesses and consumers.[2]
- ✓ Acknowledge that, while bringing substantial benefits to individuals and society, AI systems also pose certain risks and may have a negative impact, including impacts which may be difficult to anticipate, identify or measure (e.g. on democracy, the rule of law and distributive justice, or on the human mind itself.) Adopt adequate measures to mitigate these risks when appropriate, and proportionately to the magnitude of the risk.

II. Drawing upon Chapter I, **Chapter II** provides guidance on how Trustworthy AI can be realised, by listing **seven requirements** that AI systems should meet. Both technical and non-technical methods can be used for their implementation.

Key guidance derived from Chapter II:

- ✓ Ensure that the development, deployment and use of AI systems meets the seven key requirements for Trustworthy AI: (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination and fairness, (6) environmental and societal well-being and (7) accountability.
- ✓ Consider technical and non-technical methods to ensure the implementation of those requirements.

[1] All normative statements in this document aim to reflect guidance towards achieving the second and third component of trustworthy AI (ethical and robust AI). These statements are hence not meant to provide legal advice or to offer guidance on compliance with applicable laws, though it is acknowledged that many of these statements are to some extent already reflected in existing laws. In this regard, see §21 and following.

[2] See articles 24 to 27 of the Charter of Fundamental Rights of the EU (EU Charter), dealing with the rights of the child and the elderly, the integration of persons with disabilities and workers' rights. See also article 38 dealing with consumer protection.

- ✓ Foster research and innovation to help assess AI systems and to further the achievement of the requirements; disseminate results and open questions to the wider public, and systematically train a new generation of experts in AI ethics.
- ✓ Communicate, in a clear and proactive manner, information to stakeholders about the AI system's capabilities and limitations, enabling realistic expectation setting, and about the manner in which the requirements are implemented. Be transparent about the fact that they are dealing with an AI system.
- ✓ Facilitate the traceability and auditability of AI systems, particularly in critical contexts or situations.
- ✓ Involve stakeholders throughout the AI system's life cycle. Foster training and education so that all stakeholders are aware of and trained in Trustworthy AI.
- ✓ Be mindful that there might be fundamental tensions between different principles and requirements. Continuously identify, evaluate, document and communicate these trade-offs and their solutions.

III. **Chapter III** provides a concrete and non-exhaustive Trustworthy AI assessment list aimed at operationalising the key requirements set out in Chapter II. This **assessment list** will need to be tailored to the specific use case of the AI system.[3]

Key guidance derived from Chapter III:

- ✓ Adopt a Trustworthy AI assessment list when developing, deploying or using AI systems, and adapt it to the specific use case in which the system is being applied.
- ✓ Keep in mind that such an assessment list will never be exhaustive. Ensuring Trustworthy AI is not about ticking boxes, but about continuously identifying and implementing requirements, evaluating solutions, ensuring improved outcomes throughout the AI system's lifecycle, and involving stakeholders in this.

A final section of the document aims to concretise some of the issues touched upon throughout the framework, by offering examples of beneficial opportunities that should be pursued, and critical concerns raised by AI systems that should be carefully considered.

While these Guidelines aim to offer guidance for AI applications in general by building a horizontal foundation to achieve Trustworthy AI, different situations raise different challenges. It should therefore be explored whether, in addition to this horizontal framework, a sectorial approach is needed, given the context-specificity of AI systems.

These Guidelines do not intend to substitute any form of current or future policymaking or regulation, nor do they aim to deter the introduction thereof. They should be seen as a living document to be reviewed and updated over time to ensure their continuous relevance as the technology, our social environments, and our knowledge evolve. This document is a starting point for the discussion about "Trustworthy AI for Europe".[4]

Beyond Europe, the Guidelines also aim to foster research, reflection and discussion on an ethical framework for AI systems at a global level.

[3] In line with the scope of the framework, this assessment list does not provide any advice on ensuring legal compliance (lawful AI), but limits itself to offering guidance on meeting the second and third components of trustworthy AI (ethical and robust AI).

[4] This ideal is intended to apply to AI systems developed, deployed and used in the Member States of the European Union (EU), as well as to systems developed or produced elsewhere but deployed and used in the EU. When referring to "Europe" in this document, we mean this to encompass the EU Member States. However, these Guidelines also aspire to be relevant outside the EU. In this regard, it can also be noted that both Norway and Switzerland are part of the Coordinated Plan on AI agreed and published in December 2018 by the Commission and Member States.

A. INTRODUCTION

In its Communication of 25 April 2018 and 7 December 2018, the European Commission set out its vision for artificial intelligence (AI), which supports "ethical, secure and cutting-edge AI made in Europe".[5] Three pillars underpin the Commission's vision: (i) increasing public and private investments in AI to boost its uptake, (ii) preparing for socio-economic changes, and (iii) ensuring an appropriate ethical and legal framework to strengthen European values.

To support the implementation of this vision, the Commission established the High-Level Expert Group on Artificial Intelligence (AI HLEG), an independent group mandated with the drafting of two deliverables: (1) AI Ethics Guidelines and (2) Policy and Investment Recommendations.

This document contains the AI Ethics Guidelines, which have been revised following further deliberation by our Group in light of feedback received from the public consultation on the draft published on 18 December 2018. It builds on the work of the European Group on Ethics in Science and New Technologies[6] and takes inspiration from other similar efforts.[7]

Over the past months, the 52 of us met, discussed and interacted, committed to the European motto: united in diversity. We believe that AI has the potential to significantly transform society. AI is not an end in itself, but rather a promising means to increase human flourishing, thereby enhancing individual and societal well-being and the common good, as well as bringing progress and innovation. In particular, AI systems can help to facilitate the achievement of the UN's Sustainable Development Goals, such as promoting gender balance and tackling climate change, rationalising our use of natural resources, enhancing our health, mobility and production processes, and supporting how we monitor progress against sustainability and social cohesion indicators.

To do this, AI systems[8] need to be **human-centric**, resting on a commitment to their use in the service of humanity and the common good, with the goal of improving human welfare and freedom. While offering great opportunities, AI systems also give rise to certain risks that must be handled appropriately and proportionately. We now have an important window of opportunity to shape their development. We want to ensure that we can trust the socio-technical environments in which they are embedded. We also want producers of AI systems to get a competitive advantage by embedding Trustworthy AI in their products and services. This entails seeking to **maximise the benefits of AI systems** while at the same time **preventing and minimising their risks**.

In a context of rapid technological change, we believe it is essential that trust remains the bedrock of societies, communities, economies and sustainable development. We therefore identify **Trustworthy AI as our foundational ambition,** since human beings and communities will only be able to have confidence in the technology's development and its applications when a clear and comprehensive framework for achieving its trustworthiness is in place.

This is the path that we believe Europe should follow to become the home and leader of cutting-edge and ethical technology. It is through Trustworthy AI that we, as European citizens, will seek to reap its benefits in a way that is aligned with our foundational values of respect for human rights, democracy and the rule of law.

Trustworthy AI

Trustworthiness is a prerequisite for people and societies to develop, deploy and use AI systems. Without AI systems – and the human beings behind them – being demonstrably worthy of trust, unwanted consequences may ensue and their uptake might be hindered, preventing the realisation of the potentially vast social and economic

---

5 COM(2018)237 and COM(2018)795. Note that the term "made in Europe" is used throughout the Commission's communication. The scope of these Guidelines however aims to encompass not only those AI systems made in Europe, but also those developed elsewhere and deployed or used in Europe. Throughout this document, we hence aim to promote trustworthy AI "for" Europe.

6 The European Group on Ethics in Science and New Technologies (EGE) is an advisory group of the Commission.

7 See Section 3.3 of COM(2018)237.

8 The Glossary at the end of this document provides a definition of AI systems for the purpose of this document. This definition is further elaborated on in a dedicated document prepared by the AI HLEG that accompanies these Guidelines, titled "*A definition of AI: Main capabilities and scientific disciplines*".

benefits that they can bring. To help Europe realise those benefits, our vision is to ensure and scale Trustworthy AI.

Trust in the development, deployment and use of AI systems concerns not only the technology's inherent properties, but also the qualities of the socio-technical systems involving AI applications.[9] Analogous to questions of (loss of) trust in aviation, nuclear power or food safety, it is not simply components of the AI system but the system in its overall context that may or may not engender trust. Striving towards Trustworthy AI hence concerns not only the trustworthiness of the AI system itself, but requires a holistic and systemic approach, encompassing the trustworthiness of all actors and processes that are part of the system's socio-technical context throughout its entire life cycle.

Trustworthy AI has **three components,** which should be met throughout the system's entire life cycle:

1. it should be **lawful**, complying with all applicable laws and regulations;
2. it should be **ethical**, ensuring adherence to ethical principles and values; and
3. it should be **robust**, both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm.

Each of these three components is necessary but not sufficient in itself to achieve Trustworthy AI.[10] Ideally, all three work in harmony and overlap in their operation. In practice, however, there may be tensions between these elements (e.g. at times the scope and content of existing law might be out of step with ethical norms). It is our individual and collective responsibility as a society to work towards ensuring that all three components help to secure Trustworthy AI.[11]

A trustworthy approach is key to enabling "responsible competitiveness", by providing the foundation upon which all those affected by AI systems can trust that their design, development and use are lawful, ethical and robust. These Guidelines are intended to foster responsible and sustainable AI innovation in Europe. They seek to make ethics a core pillar for developing a unique approach to AI, one that aims to benefit, empower and protect both individual human flourishing and the common good of society. We believe that this will enable Europe to position itself as a global leader in cutting-edge AI worthy of our individual and collective trust. Only by ensuring trustworthiness will European individuals fully reap AI systems' benefits, secure in the knowledge that measures are in place to safeguard against their potential risks.

Just as the use of AI systems does not stop at national borders, neither does their impact. Global solutions are therefore required for the global opportunities and challenges that AI systems bring forth. We therefore encourage all stakeholders to work towards a global framework for Trustworthy AI, building international consensus while promoting and upholding our fundamental rights-based approach.

Audience and Scope

These guidelines are addressed to all AI stakeholders designing, developing, deploying, implementing, using or being affected by AI, including but not limited to companies, organisations, researchers, public services, government agencies, institutions, civil society organisations, individuals, workers and consumers. Stakeholders committed towards achieving Trustworthy AI can **voluntarily** opt to use these Guidelines as a method to operationalise their commitment, in particular by using the practical assessment list of Chapter III when developing, deploying or using AI systems. This assessment list can also complement – and hence be incorporated in – existing assessment processes.

The Guidelines aim to provide guidance for AI applications in general, building a horizontal foundation to achieve Trustworthy AI. However, **different situations raise different challenges**. AI music recommendation systems do not

[9] These systems comprise humans, state actors, corporations, infrastructure, software, protocols, standards, governance, existing laws, oversight mechanisms, incentive structures, auditing procedures, best practices reporting and others.

[10] This does not exclude the fact that additional conditions may be(come) necessary.

[11] This also means that the legislature or policy-makers may need to review the adequacy of existing law where these might be out of step with ethical principles.

raise the same ethical concerns as AI systems proposing critical medical treatments. Likewise, different opportunities and challenges arise from AI systems used in the context of business-to-consumer, business-to-business, employer-to-employee and public-to-citizen relationships, or more generally, in different sectors or use cases. Given the context-specificity of AI systems, the implementation of these Guidelines needs to be adapted to the particular AI-application. Moreover, the necessity of an additional sectorial approach, to complement the more general horizontal framework proposed in this document, should be explored.

To gain a better understanding of how this guidance can be implemented at a horizontal level, and of those matters that require a sectorial approach, we invite all stakeholders to pilot the Trustworthy AI assessment list (Chapter III) that operationalises this framework and to provide us feedback. Based on the feedback gathered through this piloting phase, we will revise the assessment list of these Guidelines by early 2020. The piloting phase will be launched by the summer of 2019 and last until the end of the year. All interested stakeholders will be able to participate by indicating their interest through the European AI Alliance.

B. A FRAMEWORK FOR TRUSTWORTHY AI

These Guidelines articulate a framework for achieving Trustworthy AI based on fundamental rights as enshrined in the Charter of Fundamental Rights of the European Union (EU Charter), and in relevant international human rights law. Below, we briefly touch upon Trustworthy AI's three components.

Lawful AI

AI systems do not operate in a lawless world. A number of legally binding rules at European, national and international level already apply or are relevant to the development, deployment and use of AI systems today. Legal sources include, but are not limited to: EU primary law (the Treaties of the European Union and its Charter of Fundamental Rights), EU secondary law (such as the General Data Protection Regulation, the Product Liability Directive, the Regulation on the Free Flow of Non-Personal Data, anti-discrimination Directives, consumer law and Safety and Health at Work Directives), the UN Human Rights treaties and the Council of Europe conventions (such as the European Convention on Human Rights), and numerous EU Member State laws. Besides horizontally applicable rules, various domain-specific rules exist that apply to particular AI applications (such as for instance the Medical Device Regulation in the healthcare sector).

The law provides both positive and negative obligations, which means that it should not only be interpreted with reference to what *cannot* be done, but also with reference to what *should* be done and what *may* be done. The law not only prohibits certain actions but also enables others. In this regard, it can be noted that the EU Charter contains articles on the 'freedom to conduct a business' and the 'freedom of the arts and sciences', alongside articles addressing areas that we are more familiar with when looking to ensure AI's trustworthiness, such as for instance data protection and non-discrimination.

The Guidelines do not explicitly deal with the first component of Trustworthy AI (lawful AI), but instead aim to offer guidance on fostering and securing the second and third components (ethical and robust AI). While the two latter are to a certain extent often already reflected in existing laws, their full realisation may go beyond existing legal obligations.

Nothing in this document shall be construed or interpreted as providing legal advice or guidance concerning how compliance with any applicable existing legal norms and requirements can be achieved. Nothing in this document shall create legal rights nor impose legal obligations towards third parties. We however recall that it is the duty of any natural or legal person to comply with laws – whether applicable today or adopted in the future according to the development of AI. **These Guidelines proceed on the assumption that all legal rights and obligations that apply to the processes and activities involved in developing, deploying and using AI systems remain mandatory and must be duly observed.**

Ethical AI

Achieving Trustworthy AI requires not only compliance with the law, which is but one of its three components. Laws

are not always up to speed with technological developments, can at times be out of step with ethical norms or may simply not be well suited to addressing certain issues. For AI systems to be trustworthy, they should hence also be ethical, ensuring alignment with ethical norms.

Robust AI

Even if an ethical purpose is ensured, individuals and society must also be confident that AI systems will not cause any unintentional harm. Such systems should perform in a safe, secure and reliable manner, and safeguards should be foreseen to prevent any unintended adverse impacts. It is therefore important to ensure that AI systems are robust. This is needed both from a technical perspective (ensuring the system's technical robustness as appropriate in a given context, such as the application domain or life cycle phase), and from a social perspective (in due consideration of the context and environment in which the system operates).

Ethical and robust AI are hence closely intertwined and complement each other. The principles put forward in Chapter I, and the requirements derived from these principles in Chapter II, address both components.

The framework

The Guidance in this document is provided in three chapters, from most abstract in Chapter I to most concrete in Chapter III:

- **Chapter I – Foundations of Trustworthy AI:** sets out the foundations of Trustworthy AI by laying out its fundamental-rights[12] based approach. It identifies and describes the ethical principles that must be adhered to in order to ensure ethical and robust AI.
- **Chapter II – Realising Trustworthy AI**: translates these ethical principles into seven key requirements that AI systems should implement and meet throughout their entire life cycle. In addition, it offers both technical and non-technical methods that can be used for their implementation.
- **Chapter III – Assessing Trustworthy AI:** sets out a concrete and non-exhaustive Trustworthy AI assessment list to operationalise the requirements of Chapter II, offering AI practitioners practical guidance. This assessment should be tailored to the particular system's application.

The document's final section lists examples of beneficial opportunities and critical concerns raised by AI systems, which should serve to stimulate further debate.

The Guidelines' structure is illustrated in *Figure 1* below

[12] Fundamental rights lie at the foundation of both international and EU human rights law and underpin the legally enforceable rights guaranteed by the EU Treaties and the EU Charter. Being legally binding, compliance with fundamental rights hence falls under trustworthy AI's first component (lawful AI). Fundamental rights can however also be understood as reflecting special moral entitlements of all individuals arising by virtue of their humanity, regardless of their legally binding status. In that sense, they hence also form part of the second component of trustworthy AI (ethical AI).

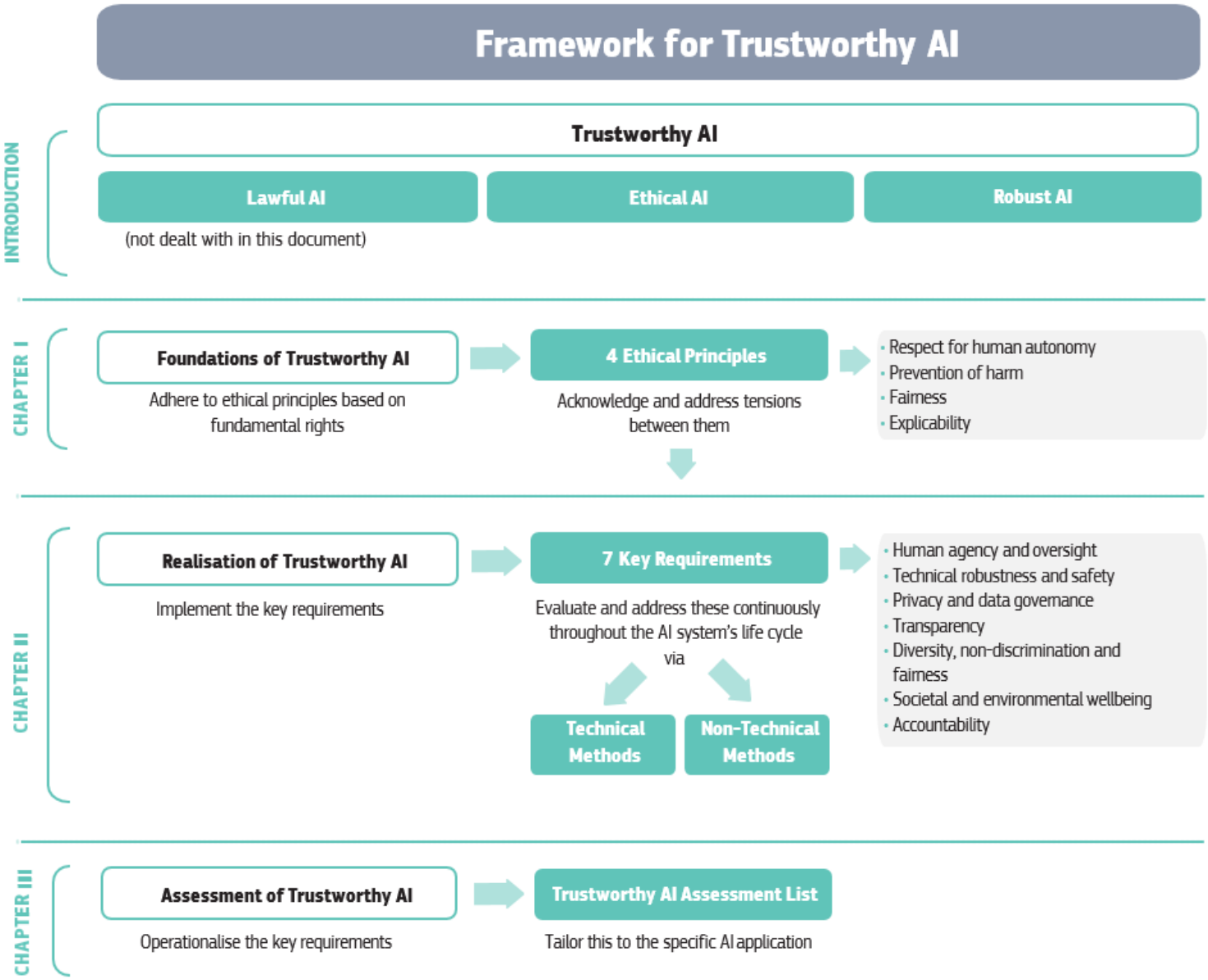

Figure 1: The Guidelines as a framework for Trustworthy AI

I. Chapter I: Foundations of Trustworthy AI

This Chapter sets out the foundations of Trustworthy AI, grounded in fundamental rights and reflected by four ethical principles that should be adhered to in order to ensure ethical and robust AI. It draws heavily on the field of ethics.

AI ethics is a sub-field of applied ethics, focusing on the ethical issues raised by the development, deployment and use of AI. Its central concern is to identify how AI can advance or raise concerns to the good life of individuals, whether in terms of quality of life, or human autonomy and freedom necessary for a democratic society.

Ethical reflection on AI technology can serve multiple purposes. First, it can stimulate reflection on the need to protect individuals and groups at the most basic level. Second, it can stimulate new kinds of innovations that seek to foster ethical values, such as those helping to achieve the UN Sustainable Development Goals[13], which are firmly embedded in the forthcoming EU Agenda 2030.[14] While this document mostly concerns itself with the first purpose mentioned, the importance that ethics could have in the second should not be underestimated. Trustworthy AI can improve individual flourishing and collective wellbeing by generating prosperity, value creation and wealth maximization. It can contribute to achieving a fair society, by helping to increase citizens' health and well-being in ways that foster equality in the distribution of economic, social and political opportunity.

It is therefore imperative that we understand how to best support AI development, deployment and use to ensure that everyone can thrive in an AI-based world, and to build a better future while at the same time being globally competitive. As with any powerful technology, the use of AI systems in our society raises several ethical challenges, for instance relating to their impact on people and society, decision-making capabilities and safety. If we are increasingly going to use the assistance of or delegate decisions to AI systems, we need to make sure these systems are fair in their impact on people's lives, that they are in line with values that should not be compromised and able to act accordingly, and that suitable accountability processes can ensure this.

Europe needs to define what normative vision of an AI-immersed future it wants to realise, and understand which notion of AI should be studied, developed, deployed and used in Europe to achieve this vision. With this document, we intend to contribute to this effort by introducing the notion of Trustworthy AI, which we believe is the right way to build a future with AI. A future where democracy, the rule of law and fundamental rights underpin AI systems and where such systems continuously improve and defend democratic culture will also enable an environment where innovation and responsible competitiveness can thrive.

A domain-specific ethics code – however consistent, developed and fine-grained future versions of it may be – can never function as a substitute for ethical reasoning itself, which must always remain sensitive to contextual details that cannot be captured in general Guidelines. Beyond developing a set of rules, ensuring Trustworthy AI requires us to build and maintain an ethical culture and mind-set through public debate, education and practical learning.

1. Fundamental rights as moral and legal entitlements

We believe in an approach to AI ethics based on the fundamental rights enshrined in the EU Treaties,[15] the EU Charter and international human rights law.[16] Respect for fundamental rights, within a framework of democracy and the rule of law, provides the most promising foundations for identifying abstract ethical principles and values, which can be operationalised in the context of AI.

The EU Treaties and the EU Charter prescribe a series of fundamental rights that EU member states and EU institutions are legally obliged to respect when implementing EU law. These rights are described in the EU Charter

[13] https://ec.europa.eu/commission/publications/reflection-paper-towards-sustainable-europe-2030_en

[14] https://sustainabledevelopment.un.org/?menu=1300

[15] The EU is based on a constitutional commitment to protect the fundamental and indivisible rights of human beings, to ensure respect for the rule of law, to foster democratic freedom and promote the common good. These rights are reflected in Articles 2 and 3 of the Treaty on European Union, and in the Charter of Fundamental Rights of the EU.

[16] Other legal instruments reflect and provide further specification of these commitments, such as for instance the Council of Europe's European Social Charter or specific legislation such as the EU's General Data Protection Regulation.

by reference to dignity, freedoms, equality and solidarity, citizens' rights and justice. The common foundation that unites these rights can be understood as rooted in respect for human dignity – thereby reflecting what we describe as a "human-centric approach" in which the human being enjoys a unique and inalienable moral status of primacy in the civil, political, economic and social fields.[17]

While the rights set out in the EU Charter are legally binding,[18] it is important to recognise that fundamental rights do not provide comprehensive legal protection in every case. For the EU Charter, for instance, it is important to underline that its field of application is limited to areas of EU law. International human rights law and in particular the European Convention on Human Rights are legally binding on EU Member States, including in areas that fall outside the scope of EU law. At the same time, fundamental rights are also bestowed on individuals and (to a certain degree) groups by virtue of their moral status as human beings, independently of their legal force. Understood as legally enforceable rights, fundamental rights therefore fall under the first component of Trustworthy AI (lawful AI), which safeguards compliance with the law. Understood as the rights of everyone, rooted in the inherent moral status of human beings, they also underpin the second component of Trustworthy AI (ethical AI), dealing with ethical norms that are not necessarily legally binding yet crucial to ensure trustworthiness. Since this document does not aim to offer guidance on the former component, for the purpose of these non-binding guidelines, references to fundamental rights reflect the latter component.

2. From fundamental rights to ethical principles

2.1 Fundamental rights as a basis for Trustworthy AI

Among the comprehensive set of indivisible rights set out in international human rights law, the EU Treaties and the EU Charter, the below families of fundamental rights are particularly apt to cover AI systems. Many of these rights are, in specified circumstances, legally enforceable in the EU so that compliance with their terms is legally obligatory. But even after compliance with legally enforceable fundamental rights has been achieved, ethical reflection can help us understand how the development, deployment and use of AI systems may implicate fundamental rights and their underlying values, and can help provide more fine-grained guidance when seeking to identify what we *should* do rather than what we (currently) *can* do with technology.

Respect for human dignity. Human dignity encompasses the idea that every human being possesses an "intrinsic worth", which should never be diminished, compromised or repressed by others – nor by new technologies like AI systems.[19] In this context, respect for human dignity entails that all people are treated with respect due to them as moral *subjects*, rather than merely as *objects* to be sifted, sorted, scored, herded, conditioned or manipulated. AI systems should hence be developed in a manner that respects, serves and protects humans' physical and mental integrity, personal and cultural sense of identity, and satisfaction of their essential needs.[20]

Freedom of the individual. Human beings should remain free to make life decisions for themselves. This entails freedom from sovereign intrusion, but also requires intervention from government and non-governmental organisations to ensure that individuals or people at risk of exclusion have equal access to AI's benefits and opportunities. In an AI context, freedom of the individual for instance requires mitigation of (in)direct illegitimate coercion, threats to mental autonomy and mental health, unjustified surveillance, deception and unfair manipulation. In fact, freedom of the individual means a commitment to enabling individuals to wield even higher control over their lives, including (among other rights) protection of the freedom to conduct a business, the freedom of the arts and science, freedom of expression, the right to private life and privacy, and freedom of

---

17 It should be noted that a commitment to human-centric AI and its anchoring in fundamental rights requires collective societal and constitutional foundations in which individual freedom and respect for human dignity is both practically possible and meaningful, rather than implying an unduly individualistic account of the human.

18 Pursuant to Article 51 of the Charter, it applies to EU Institutions and to EU member states when implementing EU law.

19 C. McCrudden, Human Dignity and Judicial Interpretation of Human Rights, *EJIL*, 19(4), 2008.

20 For an understanding of "human dignity" along these lines see E. Hilgendorf, Problem Areas in the Dignity Debate and the Ensemble Theory of Human Dignity, in: D. Grimm, A. Kemmerer, C. Möllers (eds.), *Human Dignity in Context. Explorations of a Contested Concept*, 2018, pp. 325 ff.

assembly and association.

Respect for democracy, justice and the rule of law. All governmental power in constitutional democracies must be legally authorised and limited by law. AI systems should serve to maintain and foster democratic processes and respect the plurality of values and life choices of individuals. AI systems must not undermine democratic processes, human deliberation or democratic voting systems. AI systems must also embed a commitment to ensure that they do not operate in ways that undermine the foundational commitments upon which the rule of law is founded, mandatory laws and regulation, and to ensure due process and equality before the law.

Equality, non-discrimination and solidarity *- including the rights of persons at risk of exclusion.* Equal respect for the moral worth and dignity of all human beings must be ensured. This goes beyond non-discrimination, which tolerates the drawing of distinctions between dissimilar situations based on objective justifications. In an AI context, equality entails that the system's operations cannot generate unfairly biased outputs (e.g. the data used to train AI systems should be as inclusive as possible, representing different population groups). This also requires adequate respect for potentially vulnerable persons and groups,[21] such as workers, women, persons with disabilities, ethnic minorities, children, consumers or others at risk of exclusion.

Citizens' rights. Citizens benefit from a wide array of rights, including the right to vote, the right to good administration or access to public documents, and the right to petition the administration. AI systems offer substantial potential to improve the scale and efficiency of government in the provision of public goods and services to society. At the same time, citizens' rights could also be negatively impacted by AI systems and should be safeguarded. When the term "citizens' rights" is used here, this is not to deny or neglect the rights of third-country nationals and irregular (or illegal) persons in the EU who also have rights under international law, and – therefore – in the area of AI systems.

2.2 Ethical Principles in the Context of AI Systems[22]

Many public, private, and civil organizations have drawn inspiration from fundamental rights to produce ethical frameworks for AI systems.[23] In the EU, the European Group on Ethics in Science and New Technologies ("EGE") proposed a set of 9 basic principles, based on the fundamental values laid down in the EU Treaties and Charter.[24] We build further on this work, recognising most of the principles hitherto propounded by various groups, while clarifying the ends that all principles seek to nurture and support. These ethical principles can inspire new and specific regulatory instruments, can help interpreting fundamental rights as our socio-technical environment evolves over time, and can guide the rationale for AI systems' development, deployment and use – adapting dynamically as society itself evolves.

AI systems should improve individual and collective wellbeing. This section lists **four ethical principles**, rooted in fundamental rights, which must be respected in order to ensure that AI systems are developed, deployed and used in a trustworthy manner. They are specified as **ethical imperatives**, such that AI practitioners should always strive to adhere to them. Without imposing a hierarchy, we list the principles here below in manner that mirrors the order of appearance of the fundamental rights upon which they are based in the EU Charter.[25]

---

21 For a description of the term as used throughout this document, see the Glossary.

22 These principles also apply to the development, deployment and use of other technologies, and hence are not specific to AI systems. In what follows, we have aimed to set out their relevance specifically in an AI-related context.

23 Reliance on fundamental rights also helps to limit regulatory uncertainty as it can build on the basis of decades of practice of fundamental rights protection in the EU, thereby offering clarity, readability and foreseeability.

24 More recently, the AI4People's taskforce has surveyed the aforementioned EGE principles as well as 36 other ethical principles put forward to date and subsumed them under four overarching principles: L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, E. J. M. Vayena (2018), "AI4People —An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations", Minds and Machines 28(4): 689-707.

25 Respect for human autonomy is strongly associated with the right to human dignity and liberty (reflected in Articles 1 and 6 of the Charter). The prevention of harm is strongly linked to the protection of physical or mental integrity (reflected in Article 3).

These are the principles of:

(i) Respect for human autonomy
(ii) Prevention of harm
(iii) Fairness
(iv) Explicability

Many of these are to a large extent already reflected in existing legal requirements for which mandatory compliance is required and hence also fall within the scope of lawful AI, which is Trustworthy AI's first component.[26] Yet, as set out above, while many legal obligations reflect ethical principles, adherence to ethical principles goes beyond formal compliance with existing laws.[27]

- *The principle of respect for human autonomy*

The fundamental rights upon which the EU is founded are directed towards ensuring respect for the freedom and autonomy of human beings. Humans interacting with AI systems must be able to keep full and effective self-determination over themselves, and be able to partake in the democratic process. AI systems should not unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans. Instead, they should be designed to augment, complement and empower human cognitive, social and cultural skills. The allocation of functions between humans and AI systems should follow human-centric design principles and leave meaningful opportunity for human choice. This means securing human oversight[28] over work processes in AI systems. AI systems may also fundamentally change the work sphere. It should support humans in the working environment, and aim for the creation of meaningful work.

- *The principle of prevention of harm*

AI systems should neither cause nor exacerbate harm[29] or otherwise adversely affect human beings.[30] This entails the protection of human dignity as well as mental and physical integrity. AI systems and the environments in which they operate must be safe and secure. They must be technically robust and it should be ensured that they are not open to malicious use. Vulnerable persons should receive greater attention and be included in the development, deployment and use of AI systems. Particular attention must also be paid to situations where AI systems can cause or exacerbate adverse impacts due to asymmetries of power or information, such as between employers and employees, businesses and consumers or governments and citizens. Preventing harm also entails consideration of the natural environment and all living beings.

- *The principle of fairness*

The development, deployment and use of AI systems must be fair. While we acknowledge that there are many different interpretations of fairness, we believe that fairness has both a substantive and a procedural dimension. The substantive dimension implies a commitment to: ensuring equal and just distribution of both benefits and costs, and ensuring that individuals and groups are free from unfair bias, discrimination and stigmatisation. If unfair biases can be avoided, AI systems could even increase societal fairness. Equal opportunity in terms of access to education, goods, services and technology should also be fostered. Moreover, the use of AI systems should never lead to people being deceived or unjustifiably impaired in their freedom of choice. Additionally, fairness implies that AI practitioners should respect the principle of proportionality between means and ends, and consider carefully how to

---

Fairness is closely linked to the rights to Non-discrimination, Solidarity and Justice (reflected in Articles 21 and following). Explicability and Responsibility are closely linked to the rights relating to Justice (as reflected in Article 47).

26 Think for instance of the GDPR or EU consumer protection regulations.

27 For further reading on this subject, see for instance L. Floridi, Soft Ethics and the Governance of the Digital, *Philosophy & Technology,* March 2018, Volume 31, Issue 1, pp 1–8.

28 The concept of human oversight is further developed as one of the key requirements set out in Chapter II here below.

29 Harms can be individual or collective, and can include intangible harm to social, cultural and political environments.

30 This also encompasses the way of living of individuals and social groups, avoiding for instance cultural harm.

balance competing interests and objectives.[31] The procedural dimension of fairness entails the ability to contest and seek effective redress against decisions made by AI systems and by the humans operating them.[32] In order to do so, the entity accountable for the decision must be identifiable, and the decision-making processes should be explicable.

- *The principle of explicability*

Explicability is crucial for building and maintaining users' trust in AI systems. This means that processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions – to the extent possible – explainable to those directly and indirectly affected. Without such information, a decision cannot be duly contested. An explanation as to why a model has generated a particular output or decision (and what combination of input factors contributed to that) is not always possible. These cases are referred to as 'black box' algorithms and require special attention. In those circumstances, other explicability measures (e.g. traceability, auditability and transparent communication on system capabilities) may be required, provided that the system as a whole respects fundamental rights. The degree to which explicability is needed is highly dependent on the context and the severity of the consequences if that output is erroneous or otherwise inaccurate.[33]

2.3 Tensions between the principles

Tensions may arise between the above principles, for which there is no fixed solution. In line with the EU fundamental commitment to democratic engagement, due process and open political participation, methods of accountable deliberation to deal with such tensions should be established. For instance, in various application domains, *the principle of prevention of harm* and *the principle of human autonomy* may be in conflict. Consider as an example the use of AI systems for 'predictive policing', which may help to reduce crime, but in ways that entail surveillance activities that impinge on individual liberty and privacy. Furthermore, AI systems' overall benefits should substantially exceed the foreseeable individual risks. While the above principles certainly offer guidance towards solutions, they remain abstract ethical prescriptions. AI practitioners can hence not be expected to find the right solution based on the principles above, yet they should approach ethical dilemmas and trade-offs via reasoned, evidence-based reflection rather than intuition or random discretion.

There may be situations, however, where no ethically acceptable trade-offs can be identified. Certain fundamental rights and correlated principles are absolute and cannot be subject to a balancing exercise (e.g. human dignity).

Key guidance derived from Chapter I:

- ✓ Develop, deploy and use AI systems in a way that adheres to the ethical principles of: *respect for human autonomy, prevention of harm, fairness and explicability.* Acknowledge and address the potential tensions between these principles.
- ✓ Pay particular attention to situations involving more vulnerable groups such as children, persons with disabilities and others that have historically been disadvantaged or are at risk of exclusion, and to situations which are characterised by asymmetries of power or information, such as between employers and workers, or between businesses and consumers.[34]

---

31 This is relates to the principle of proportionality (reflected in the maxim that one should not 'use a sledge hammer to crack a nut'). Measures taken to achieve an end (e.g. the data extraction measures implemented to realise the AI optimisation function) should be limited to what is strictly necessary. It also entails that when several measures compete for the satisfaction of a n end, preference should be given to the one that is least adverse to fundamental rights and ethical norms (e.g. AI developers should always prefer public sector data to personal data). Reference can also be made to the proportionality between user and deployer, considering the rights of companies (including intellectual property and confidentiality) on the one hand, and the rights of the user on the other.

32 Including by using their right of association and to join a trade union in a working environment , as provided for by Article 12 of the EU Charter of fundamental rights.

33 For example, little ethical concern may flow from inaccurate shopping recommendations generated by an AI system, in contrast to AI systems that evaluate whether an individual convicted of a criminal offence should be released on par ole.

34 See articles 24 to 27 of the EU Charter, dealing with the rights of the child and the elderly, the integration of persons wit h disabilities and workers' rights. See also article 38 dealing with consumer protection.

- ✓ Acknowledge that, while bringing substantial benefits to individuals and society, AI systems also pose certain risks and may have a negative impact, including impacts which may be difficult to anticipate, identify or measure (e.g. on democracy, the rule of law and distributive justice, or on the human mind itself.) Adopt adequate measures to mitigate these risks when appropriate, and proportionately to the magnitude of the risk.

II. Chapter II: Realising Trustworthy AI

This Chapter offers guidance on the implementation and realisation of Trustworthy AI, via a list of seven requirements that should be met, building on the principles outlined in Chapter I. In addition, available technical and non-technical methods are introduced for the implementation of these requirements throughout the AI system's life cycle.

1. Requirements of Trustworthy AI

The principles outlined in Chapter I must be translated into concrete requirements to achieve Trustworthy AI. These requirements are applicable to different stakeholders partaking in AI systems' life cycle: developers, deployers and end-users, as well as the broader society. By developers, we refer to those who research, design and/or develop AI systems. By deployers, we refer to public or private organisations that use AI systems within their business processes and to offer products and services to others. End-users are those engaging with the AI system, directly or indirectly. Finally, the broader society encompasses all others that are directly or indirectly affected by AI systems.

Different groups of stakeholders have different roles to play in ensuring that the requirements are met:

a. Developers should implement and apply the requirements to design and development processes;
b. Deployers should ensure that the systems they use and the products and services they offer meet the requirements;
c. End-users and the broader society should be informed about these requirements and able to request that they are upheld.

The below list of requirements is non-exhaustive.[35] It includes systemic, individual and societal aspects:

1 Human agency and oversight

Including fundamental rights, human agency and human oversight

2 Technical robustness and safety

Including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility

3 Privacy and data governance

Including respect for privacy, quality and integrity of data, and access to data

4 Transparency

Including traceability, explainability and communication

5 Diversity, non-discrimination and fairness

Including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation

6 Societal and environmental wellbeing

Including sustainability and environmental friendliness, social impact, society and democracy

7 Accountability

Including auditability, minimisation and reporting of negative impact, trade-offs and redress.

[35] Without imposing a hierarchy, we list the principles here below in manner that mirrors the order of appearance of the principles and rights to which they relate in the EU Charter.

Figure 2: Interrelationship of the seven requirements: all are of equal importance, support each other, and should be implemented and evaluated throughout the AI system's lifecycle

While all requirements are of equal importance, context and potential tensions between them will need to be taken into account when applying them across different domains and industries. Implementation of these requirements should occur throughout an AI system's entire life cycle and depends on the specific application. While most requirements apply to all AI systems, special attention is given to those directly or indirectly affecting individuals. Therefore, for some applications (for instance in industrial settings), they may be of lesser relevance.

The above requirements include elements that are in some cases already reflected in existing laws. We reiterate that – in line with Trustworthy AI's first component – it is the responsibility of AI practitioners to ensure that they comply with their legal obligations, both as regards horizontally applicable rules as well as domain-specific regulation.

In the following paragraphs, each requirement is explained in more detail.

1.1 Human agency and oversight

AI systems should support human autonomy and decision-making, as prescribed by the principle of *respect for human autonomy*. This requires that AI systems should both act as enablers to a democratic, flourishing and equitable society by supporting the user's agency and foster fundamental rights, and allow for human oversight.

Fundamental rights. Like many technologies, AI systems can equally enable and hamper fundamental rights. They can benefit people for instance by helping them track their personal data, or by increasing the accessibility of education, hence supporting their right to education. However, 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

regarding AI systems that potentially infringe on fundamental rights.

Human agency. 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. Key to this is 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.[36]

Human oversight. Human oversight helps ensuring that an AI system does not undermine human autonomy or causes other adverse effects. Oversight may be achieved through governance mechanisms such as a human-in-the-loop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC) approach. HITL refers to the capability for human intervention in every decision cycle of the system, which in many cases is neither possible nor desirable. HOTL refers to the capability for human intervention during the design cycle of the system and monitoring the system's operation. 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 a system. Moreover, 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.

1.2 Technical robustness and safety

A crucial component of achieving Trustworthy AI is technical robustness, which is closely linked to the *principle of prevention of harm*. Technical robustness requires that AI systems be developed with a preventative approach to risks and in a manner such that they reliably behave as intended while minimising unintentional and unexpected harm, and preventing unacceptable harm. This should also apply to potential changes in their operating environment or the presence of other agents (human and artificial) that may interact with the system in an adversarial manner. In addition, the physical and mental integrity of humans should be ensured.

Resilience to attack and security. AI systems, like all software systems, should be protected against vulnerabilities that can allow them to be exploited by adversaries, e.g. hacking. Attacks may target the data (data poisoning), the model (model leakage) or the underlying infrastructure, both software and hardware. If an AI system is attacked, e.g. in adversarial attacks, the data as well as system behaviour can be changed, leading the system to make different decisions, or causing it to shut down altogether. Systems and data can also become corrupted by malicious intention or by exposure to unexpected situations. Insufficient security processes can also result in erroneous decisions or even physical harm. For AI systems to be considered secure,[37] possible unintended applications of the AI system (e.g. dual-use applications) and potential abuse of the system by malicious actors should be taken into account, and steps should be taken to prevent and mitigate these.[38]

Fallback plan and general safety. AI systems should have safeguards that enable a fallback plan in case of problems.

---

36 Reference can be made to Article 22 of the GDPR where this right is already enshrined.

37 See e.g. considerations under 2.7 of the European Union's Coordinated Plan on Artificial Intelligence.

38 There may be a strong imperative to develop a virtuous circle in research and development between understanding of attacks, development of adequate protection, and improvement of evaluation methodologies. To achieve this, convergence between the AI community and the security community should be promoted. In addition, it is the responsibility of all relevant actors to create common cross-border safety and security norms and to establish an environment of mutual trust, fostering international collaboration. For possible measures, see *Malicious Use of AI*, Avin S., Brundage M. et. al., 2018.

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.[39] 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.

Accuracy. Accuracy pertains to an AI system's ability to make correct judgements, for example to correctly classify information into the proper categories, or its ability to make correct predictions, recommendations, or decisions based on data or models. An explicit and well-formed development and evaluation process can support, mitigate and correct unintended risks from inaccurate predictions. When occasional inaccurate predictions cannot be avoided, it is important that the system can indicate how likely these errors are. A high level of accuracy is especially crucial in situations where the AI system directly affects human lives.

Reliability and Reproducibility. 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[40] can facilitate the process of testing and reproducing behaviours.

1.3 Privacy and data governance

Closely linked to the *principle of prevention of harm* is privacy, a fundamental right particularly affected by AI systems. Prevention of harm to privacy also necessitates adequate data governance that covers the quality and integrity of the data used, its relevance in light of the domain in which the AI systems will be deployed, its access protocols and the capability to process data in a manner that protects privacy.

Privacy and data protection. AI systems must guarantee privacy and data protection throughout a system's entire lifecycle.[41] 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.

Quality and integrity of data. The quality of the data sets used is paramount to the performance of AI systems. When data is gathered, it may contain socially constructed biases, inaccuracies, errors and mistakes. This needs to be addressed prior to training with any given data set. In addition, the integrity of the data must be ensured. Feeding malicious data into an AI system may change its behaviour, particularly with self-learning systems. 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.

Access to data. In any given organisation that handles individuals' data (whether someone is a user of the system or not), data protocols governing data access should be put in place. These protocols should outline who can access data and under which circumstances. Only duly qualified personnel with the competence and need to access individual's data should be allowed to do so.

---

39 Scenarios where human intervention would not immediately be possible should also be considered.

40 This concerns files that will replicate each step of the AI system's development process, from research and initial data collection to the results.

41 Reference can be made to existing privacy laws, such as the GDPR or the forthcoming ePrivacy Regulation.

1.4 Transparency

This requirement is closely linked with the *principle of explicability* and encompasses transparency of elements relevant to an AI system: the data, the system and the business models.

Traceability. 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.

***Explainability*.** 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).

***Communication*.** 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. Beyond this, 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.

1.5 Diversity, non-discrimination and fairness

In order to achieve Trustworthy AI, we must enable inclusion and diversity throughout the entire AI system's life cycle. Besides the consideration and involvement of all affected stakeholders throughout the process, this also entails ensuring equal access through inclusive design processes as well as equal treatment. This requirement is closely linked with *the principle of fairness*.

Avoidance of unfair bias. 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 unintended (in)direct prejudice and discrimination[42] against certain groups or people, potentially exacerbating prejudice and marginalisation. Harm can also result from the intentional exploitation of (consumer) biases or by engaging in unfair competition, such as the homogenisation of prices by means of collusion or a non-transparent market.[43] 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. Moreover, hiring from diverse backgrounds, cultures and disciplines can ensure diversity of opinions and should be encouraged.

Accessibility and universal design. Particularly in business-to-consumer domains, systems should be user-centric and designed in a way that allows all people to use AI products or services, regardless of their age, gender, abilities or characteristics. Accessibility to this technology for persons with disabilities, which are present in all societal

[42] For a definition of direct and indirect discrimination, see for instance Article 2 of Council Directive 2000/78/EC of 27 Nove mber 2000 establishing a general framework for equal treatment in employment and occupation. See also Article 21 of the Charter of Fundamental Rights of the EU.

[43] See the EU Agency for Fundamental Rights' paper: "BigData: Discrimination in data-supported decision making", 2018, http://fra.europa.eu/en/publication/2018/big-data-discrimination.

groups, is of particular importance. AI systems should not have a one-size-fits-all approach and should consider Universal Design[44] principles addressing the widest possible range of users, following relevant accessibility standards.[45] This will enable equitable access and active participation of all people in existing and emerging computer-mediated human activities and with regard to assistive technologies.[46]

Stakeholder Participation. In order to develop AI systems that are trustworthy, it is advisable to consult stakeholders who may directly or indirectly be affected by the system throughout its life cycle. It is beneficial to solicit regular feedback even after deployment and set up longer term mechanisms for stakeholder participation, for example by ensuring workers information, consultation and participation throughout the whole process of implementing AI systems at organisations.

1.6 Societal and environmental well-being

In line with the *principles of fairness* and *prevention of harm*, the broader society, other sentient beings and the environment should be also considered as stakeholders throughout the AI system's life cycle. 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. Ideally, AI systems should be used to benefit all human beings, including future generations.

Sustainable and environmentally friendly AI. AI systems promise to help tackling some of the most pressing societal concerns, yet it must be ensured that this occurs in the most environmentally friendly way possible. The system's development, deployment and use process, as well as its entire supply chain, should be assessed in this regard, e.g. via a critical examination of the resource usage and energy consumption during training, opting for less harmful choices. Measures securing the environmental friendliness of AI systems' entire supply chain should be encouraged.

Social impact. Ubiquitous exposure to social AI systems[47] in all areas of our lives (be it in education, work, care or entertainment) may alter our conception of social agency, or impact our social relationships and attachment. While AI systems can be used to enhance social skills,[48] they can equally contribute to their deterioration. This could also affect people's physical and mental wellbeing. The effects of these systems must therefore be carefully monitored and considered.

Society and Democracy. 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.

1.7 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.

Auditability. 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

---

44 Article 42 of the Public Procurement Directive requires technical specifications to consider accessibility and 'design for all'.

45 For instance EN 301 549.

46 This requirement links to the United Nations Convention on the Rights of Persons with Disabilities.

47 This denotes AI systems communicating and interacting with humans by simulating sociality in human robot interaction (embodied AI) or as avatars in virtual reality. By doing so, those systems have the potential to change our socio-cultural practices and the fabric of our social life.

48 See for instance the EU-funded project developing AI-based software that enables robots to interact more effectively with autistic children in human-led therapy sessions, helping to improve their social and communication skills: http://ec.europa.eu/research/infocentre/article_en.cfm?id=/research/headlines/news/article_19_03_12_en.html?infocentre&item=Infocentre&artid=49968

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.

Minimisation and reporting of negative impacts. 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. Due protection must be available for whistle-blowers, NGOs, trade unions or other entities when reporting legitimate concerns about an AI system. The use of impact assessments (e.g. red teaming or forms of Algorithmic Impact Assessment) both prior to and during the development, deployment and use of AI systems can be helpful to minimise negative impact. These assessments must be proportionate to the risk that the AI systems pose.

Trade-offs. When implementing the above requirements, tensions may arise between them, which may lead to inevitable trade-offs. Such trade-offs should be addressed in a rational and methodological manner within the state of the art. This entails that relevant interests and values implicated by the AI system should be identified and that, if conflict arises, trade-offs should be explicitly acknowledged and evaluated in terms of their risk to ethical principles, including fundamental rights. In situations in which no ethically acceptable trade-offs can be identified, the development, deployment and use of the AI system should not proceed in that form. Any decision about which trade-off to make should be reasoned and properly documented. The decision-maker must be accountable for the manner in which the appropriate trade-off is being made, and should continually review the appropriateness of the resulting decision to ensure that necessary changes can be made to the system where needed.[49]

Redress. When unjust adverse impact occurs, accessible mechanisms should be foreseen that ensure adequate redress.[50] Knowing that redress is possible when things go wrong is key to ensure trust. Particular attention should be paid to vulnerable persons or groups.

2. Technical and non-technical methods to realise Trustworthy AI

To implement the above requirements, both technical and non-technical methods can be employed. These encompass all stages of an AI system's life cycle. An evaluation of the methods employed to implement the requirements, as well as reporting and justifying[51] changes to the implementation processes, should occur on an ongoing basis. AI systems are continuously evolving and acting in a dynamic environment. The realisation of Trustworthy AI is therefore a continuous process, as depicted in Figure 3 here below.

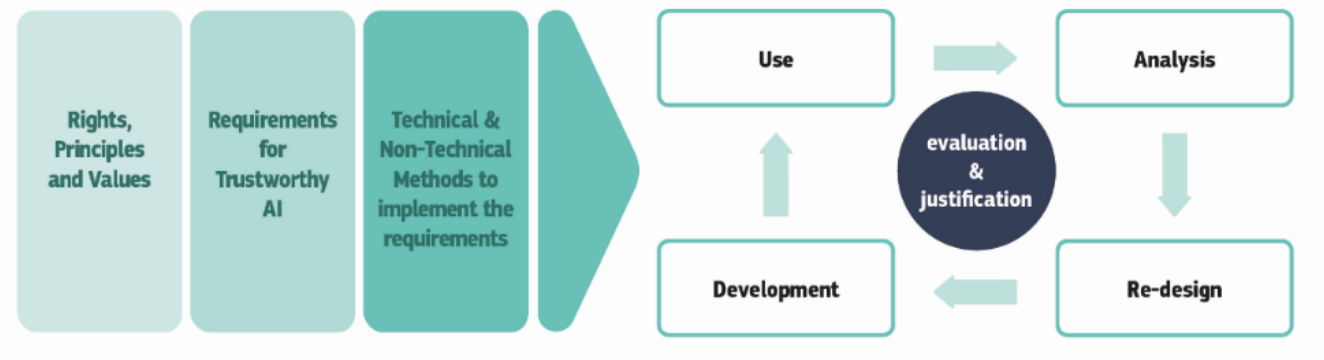

Figure 3: Realising Trustworthy AI throughout the system's entire life cycle

[49] Different governance models can help achieving this. E.g. the presence of an internal and/or external ethical (and sector specific) expert or board might be useful to highlight areas of potential conflict and suggest ways in which that conflict might best be resolved. Meaningful consultation and discussion with stakeholders, including those at risk of being adversely affected by an AI system is useful too. European universities should take a leading role in training the ethics experts needed.

[50] See also the European Union Agency for Fundamental Rights' Opinion on 'Improving access to remedy in the area of business and human rights at the EU level', 2017, https://fra.europa.eu/en/opinion/2017/business-human-rights.

[51] This entails e.g. justification of the choices in the system's design, development and deployment to implement the requirements.

The following methods can be either complementary or alternative to each other, since different requirements – and different sensitivities – may raise the need for different methods of implementation. This overview is neither meant to be comprehensive or exhaustive, nor mandatory. Rather, its aim is to offer a list of suggested methods that may help to implement Trustworthy AI.

2.1. Technical methods

This section describes technical methods to ensure Trustworthy AI that can be incorporated in the design, development and use phases of an AI system. The methods listed below vary in level of maturity.[52]

- *Architectures for Trustworthy AI*

Requirements for Trustworthy AI should be "translated" into procedures and/or constraints on procedures, which should be anchored in the AI system's architecture. This could be accomplished through a set of "white list" rules (behaviours or states) that the system should always follow, "black list" restrictions on behaviours or states that the system should never transgress, and mixtures of those or more complex provable guarantees regarding the system's behaviour. Monitoring of the system's compliance with these restrictions during operations may be achieved by a separate process.

AI systems with learning capabilities that can dynamically adapt their behaviour can be understood as non-deterministic systems possibly exhibiting unexpected behaviour. These are often considered through the theoretical lens of a "sense-plan-act" cycle. Adapting this architecture to ensure Trustworthy AI requires the requirements' integration at all three steps of the cycle: (i) at the "sense"-step, the system should be developed such that it recognises all environmental elements necessary to ensure adherence to the requirements; (ii) at the "plan"-step, the system should only consider plans that adhere to the requirements; (iii) at the "act"-step, the system's actions should be restricted to behaviours that realise the requirements.

The architecture as sketched above is generic and only provides an imperfect description for most AI systems. Nevertheless, it gives anchor points for constraints and policies that should be reflected in specific modules to result in an overall system that is trustworthy and perceived as such.

- *Ethics and rule of law by design (X-by-design)*

Methods to ensure values-by-design provide precise and explicit links between the abstract principles which the system is required to respect and the specific implementation decisions. The idea that compliance with norms can be implemented into the design of the AI system is key to this method. Companies are responsible for identifying the impact of their AI systems from the very start, as well as the norms their AI system ought to comply with to avert negative impacts. Different "by-design" concepts are already widely used, e.g. *privacy-by-design* and *security-by-design.* As indicated above, to earn trust AI needs to be secure in its processes, data and outcomes, and should be designed to be robust to adversarial data and attacks. It should implement a mechanism for fail-safe shutdown and enable resumed operation after a forced shut-down (such as an attack).

- *Explanation methods*

For a system to be trustworthy, we must be able to understand why it behaved a certain way and why it provided a given interpretation. A whole field of research, Explainable AI (XAI) tries to address this issue to better understand the system's underlying mechanisms and find solutions. Today, this is still an open challenge for AI systems based on neural networks. Training processes with neural nets can result in network parameters set to numerical values that are difficult to correlate with results. Moreover, sometimes small changes in data values might result in dramatic changes in interpretation, leading the system to e.g. confuse a school bus with an ostrich. This vulnerability can also be exploited during attacks on the system. Methods involving XAI research are vital not only to explain the system's

[52] While some of these methods are already available today, others still require more research. Those areas where further research is needed will also inform the AI HLEG's second deliverable, i.e. the Policy and Investment Recommendations.

behaviour to users, but also to deploy reliable technology.

- *Testing and validating*

Due to the non-deterministic and context-specific nature of AI systems, traditional testing is not enough. Failures of the concepts and representations used by the system may only manifest when a programme is applied to sufficiently realistic data. Consequently, to verify and validate processing of data, the underlying model must be carefully monitored during both training and deployment for its stability, robustness and operation within well-understood and predictable bounds. It must be ensured that the outcome of the planning process is consistent with the input, and that the decisions are made in a way allowing validation of the underlying process.

Testing and validation of the system should occur as early as possible, ensuring that the system behaves as intended throughout its entire life cycle and especially after deployment. It should include all components of an AI system, including data, pre-trained models, environments and the behaviour of the system as a whole. The testing processes should be designed and performed by an as diverse group of people as possible. Multiple metrics should be developed to cover the categories that are being tested for different perspectives. Adversarial testing by trusted and diverse "red teams" deliberately attempting to "break" the system to find vulnerabilities, and "bug bounties" that incentivise outsiders to detect and responsibly report system errors and weaknesses, can be considered. Finally, it must be ensured that the outputs or actions are consistent with the results of the preceding processes, comparing them to the previously defined policies to ensure that they are not violated.

- *Quality of Service Indicators*

Appropriate quality of service indicators can be defined for AI systems to ensure that there is a baseline understanding as to whether they have been tested and developed with security and safety considerations in mind. These indicators could include measures to evaluate the testing and training of algorithms as well as traditional software metrics of functionality, performance, usability, reliability, security and maintainability.

2.2. Non-technical methods

This section describes a variety of non-technical methods that can serve a valuable role in securing and maintaining Trustworthy AI. These too should be evaluated on an **ongoing basis**.

- *Regulation*

As mentioned above, regulation to support AI's trustworthiness already exists today – think of product safety legislation and liability frameworks. To the extent we consider that regulation may need to be revised, adapted or introduced, both as a safeguard and as an enabler, this will be raised in our second deliverable, consisting of AI Policy and Investment Recommendations.

- *Codes of conduct*

Organisations and stakeholders can sign up to the Guidelines and adapt their charter of corporate responsibility, Key Performance Indicators ("KPIs"), their codes of conduct or internal policy documents to add the striving towards Trustworthy AI. An organisation working on or with AI systems can, more generally, document its intentions, as well as underwrite them with standards of certain desirable values such as fundamental rights, transparency and the avoidance of harm.

- *Standardisation*

Standards, for example for design, manufacturing and business practices, can function as a quality management system for AI users, consumers, organisations, research institutions and governments by offering the ability to recognise and encourage ethical conduct through their purchasing decisions. Beyond conventional standards, co-regulatory approaches exist: accreditation systems, professional codes of ethics or standards for fundamental rights compliant design. Current examples are e.g. ISO Standards or the IEEE P7000 standards series, but in the future a possible 'Trustworthy AI' label might be suitable, confirming by reference to specific technical standards that the system, for instance, adheres to safety, technical robustness and transparency.

- *Certification*

As it cannot be expected that everyone is able to fully understand the workings and effects of AI systems, consideration can be given to organisations that can attest to the broader public that an AI system is transparent, accountable and fair.[53] These certifications would apply standards developed for different application domains and AI techniques, appropriately aligned with the industrial and societal standards of different contexts. Certification can however never replace responsibility. It should hence be complemented by accountability frameworks, including disclaimers as well as review and redress mechanisms.[54]

- *Accountability via governance frameworks*

Organisations should set up governance frameworks, both internal and external, ensuring accountability for the ethical dimensions of decisions associated with the development, deployment and use of AI systems. This can, for instance, include the appointment of a person in charge of ethics issues relating to AI systems, or an internal/external ethics panel or board. Amongst the possible roles of such a person, panel or board, is to provide oversight and advice. As set out above, certification specifications and bodies can also play a role to this end. Communication channels should be ensured with industry and/or public oversight groups, sharing best practices, discussing dilemmas or reporting emerging issues of ethical concerns. Such mechanisms can complement but cannot replace legal oversight (e.g. in the form of the appointment of a data protection officer or equivalent measures, legally required under data protection law).

- *Education and awareness to foster an ethical mind-set*

Trustworthy AI encourages the informed participation of all stakeholders. Communication, education and training play an important role, both to ensure that knowledge of the potential impact of AI systems is widespread, and to make people aware that they can participate in shaping the societal development. This includes all stakeholders, e.g. those involved in making the products (the designers and developers), the users (companies or individuals) and other impacted groups (those who may not purchase or use an AI system but for whom decisions are made by an AI system, and society at large). Basic AI literacy should be fostered across society. A prerequisite for educating the public is to ensure the proper skills and training of ethicists in this space.

- *Stakeholder participation and social dialogue*

The benefits of AI systems are many, and Europe needs to ensure that they are available to all. This requires an open discussion and the involvement of social partners and stakeholders, including the general public. Many organisations already rely on stakeholder panels to discuss the use of AI systems and data analytics. These panels include various members, such as legal experts, technical experts, ethicists, consumer representatives and workers. Actively seeking participation and dialogue on the use and impact of AI systems supports the evaluation of results and approaches, and can particularly be helpful in complex cases.

- *Diversity and inclusive design teams*

Diversity and inclusion play an essential role when developing AI systems that will be employed in the real world. It is critical that, as AI systems perform more tasks on their own, the teams that design, develop, test and maintain, deploy and procure these systems reflect the diversity of users and of society in general. This contributes to objectivity and consideration of different perspectives, needs and objectives. Ideally, teams are not only diverse in terms of gender, culture, age, but also in terms of professional backgrounds and skill sets.

---

[53] As advocated by e.g. the IEEE Ethically Aligned Design Initiative: https://standards.ieee.org/industry-connections/ec/autonomous-systems.html.

[54] For more on the limitations of certification, see: https://ainowinstitute.org/AI_Now_2018_Report.pdf.

Key guidance derived from Chapter II:

- ✓ Ensure that the AI system's entire life cycle meets the seven key requirements for Trustworthy AI: (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination and fairness, (6) environmental and societal well-being and (7) accountability.
- ✓ Consider technical and non-technical methods to ensure the implementation of those requirements.
- ✓ Foster research and innovation to help assessing AI systems and to further the achievement of the requirements; disseminate results and open questions to the wider public, and systematically train a new generation of experts in AI ethics.
- ✓ Communicate, in a clear and proactive manner, information to stakeholders about the AI system's capabilities and limitations, enabling realistic expectation setting, and about the manner in which the requirements are implemented. Be transparent about the fact that they are dealing with an AI system.
- ✓ Facilitate the traceability and auditability of AI systems, particularly in critical contexts and situations.
- ✓ Involve stakeholders throughout the AI system's life cycle. Foster training and education so that all stakeholders are aware of and trained in Trustworthy AI.
- ✓ Be mindful that there might be fundamental tensions between different principles and requirements. Continuously identify, evaluate, document and communicate these trade-offs and their solutions.

III. Chapter III: Assessing Trustworthy AI

Based on the key requirements of Chapter II, this Chapter sets out a non-exhaustive **Trustworthy AI assessment list** (pilot version) to **operationalise Trustworthy AI**. It particularly applies to AI systems that directly interact with users, and is primarily addressed to developers and deployers of AI systems (whether self-developed or acquired from third parties). This assessment list does not address the operationalisation of the first component of Trustworthy AI (lawful AI). Compliance with this assessment list is not evidence of legal compliance, nor is it intended as guidance to ensure compliance with applicable law. Given the application-specificity of AI systems, the assessment list will need to be tailored to the specific use case and context in which the system operates. In addition, this chapter offers a general recommendation on how to implement the assessment list for Trustworthy AI though a governance structure embracing both operational and management level.

The assessment list and governance structure will be developed in close collaboration with stakeholders across the public and private sector. The process will be driven as a piloting process, allowing for extensive feedback from two parallel processes:

a) a qualitative process, ensuring representability, where a small selection of companies, organisations and institutions (from different sectors and of different sizes) will sign up to pilot the assessment list and the governance structure in practice and to provide in-depth feedback;

b) a quantitative process where all interested stakeholders can sign up to pilot the assessment list and provide feedback through an open consultation.

After the piloting phase, we will integrate the results from the feedback process into the assessment list and prepare a revised version in early 2020. The aim is to achieve a framework that can be horizontally used across all applications and hence offer a foundation for ensuring Trustworthy AI in all domains. Once such foundation has been established, a sectorial or application-specific framework could be developed.

- *Governance*

Stakeholders may wish to consider how the Trustworthy AI assessment list can be implemented in their organisation. This can be done by incorporating the assessment process into existing governance mechanisms, or by implementing new processes. This choice will depend on the internal structure of the organisation as well as its size and available resources.

Research demonstrates that management attention at the highest level is essential to achieve change.[55] It also demonstrates that involving all stakeholders in a company, organisation or institution fosters the acceptance and the relevance of the introduction of any new process (whether or not technological).[56] Therefore, we recommend implementing a process that embraces both the involvement of operational level as well as top management level.

Level	Relevant roles (depending on the organisation)
Management and Board	Top management discusses and evaluates the AI systems' development, deployment or procurement and serves as an escalation board for evaluating all AI innovations and uses, when critical concerns are detected. It involves those impacted by the possible introduction of AI systems (e.g. workers) and their representatives throughout the process via information, consultation and participation procedures.
Compliance/Legal department/Corporate responsibility department	The responsibility department monitors the use of the assessment list and its necessary evolution to meet the technological or regulatory changes. It updates the standards or internal policies on AI systems and ensures that the use of such systems complies with the current legal and regulatory framework and to the values of the organisation.
Product and Service Development or equivalent	The Product and Service Development department uses the assessment list to evaluate AI-based products and services and logs all the results. These results are discussed at management level, which ultimately approves the new or revised AI-based applications.
Quality Assurance	The Quality Assurance department (or equivalent) ensures and checks the results of the assessment list and takes action to escalate an issue higher up if the result is not satisfactory or if unforeseen results are detected.
HR	The HR department ensures the right mix of competences and diversity of profiles for developers of AI systems. It ensures that the appropriate level of training is delivered on Trustworthy AI inside the organisation.
Procurement	The procurement department ensures that the process to procure AI-based products or services includes a check of Trustworthy AI.
Day-to-day Operations	Developers and project managers include the assessment list in their daily work and document the results and outcomes of the assessment.

- *Using the Trustworthy AI assessment list*

When using the assessment list in practice, we recommend paying attention not only to the areas of concern but also to the questions that cannot be (easily) answered. One potential problem might be the lack of diversity of skills and competences in the team developing and testing the AI system, and therefore it might be necessary to involve other stakeholders inside or outside the organisation. It is strongly recommended to log all results both in technical terms and in management terms, ensuring that the problem solving can be understood at all levels in the governance structure.

This assessment list is meant to guide AI practitioners to achieve Trustworthy AI. The assessment should be tailored to the specific use case in a proportionate way. During the piloting phase, specific sensitive areas might be revealed and the need for further specifications in such cases will be evaluated in the next steps. While this

[55] https://www.mckinsey.com/business-functions/operations/our-insights/secrets-of-successful-change-implementation

[56] See for instance A. Bryson, E. Barth and H. Dale-Olsen, The Effects of Organisational change on worker well-being and the moderating role of trade unions, *ILRReview*, 66(4), July 2013; Jirjahn, U. and Smith, S.C. (2006). 'What Factors Lead Management to Support or Oppose Employee Participation—With and Without Works Councils? Hypotheses and Evidence from Germany's Industrial Relations, 45(4), 650–680; Michie, J. and Sheehan, M. (2003). 'Labour market deregulation, "flexibility" and innovation', *Cambridge Journal of Economics*, 27(1), 123–143.

assessment list does not provide concrete answers to address the raised questions, it encourages reflection on how Trustworthy AI can be operationalised, and on the potential steps that should be taken in this regard.

- *Relation to existing law and processes*

It is also important for AI practitioners to recognise that there are various existing laws mandating particular processes or prohibiting particular outcomes, which may overlap and coincide with some of the measures listed in the assessment list. For example, data protection law sets out a series of legal requirements that must be met by those engaged in the collection and processing of personal data. Yet, because Trustworthy AI also requires the ethical handling of data, internal procedures and policies aimed at securing compliance with data protection laws might also help to facilitate ethical data handling and can hence complement existing legal processes. Compliance with this assessment list is *not*, however, evidence of legal compliance, nor is it intended as guidance to ensure compliance with applicable laws.

Moreover, many AI practitioners already have existing assessment tools and software development processes in place to ensure compliance also with non-legal standards. The below assessment should not necessarily be carried out as a stand-alone exercise, but can be incorporated into such existing practices.

TRUSTWORTHY AI ASSESSMENT LIST (PILOT VERSION)

1. Human agency and oversight

Fundamental rights:

✓ Did you carry out a fundamental rights impact assessment where there could be a negative impact on fundamental rights? Did you identify and document potential trade-offs made between the different principles and rights?

✓ Does the AI system interact with decisions by human (end) users (e.g. recommended actions or decisions to take, presenting of options)?
- Could the AI system affect human autonomy by interfering with the (end) user's decision-making process in an unintended way?
- Did you consider whether the AI system should communicate to (end) users that a decision, content, advice or outcome is the result of an algorithmic decision?
- In case of a chat bot or other conversational system, are the human end users made aware that they are interacting with a non-human agent?

Human agency:

✓ Is the AI system implemented in work and labour process? If so, did you consider the task allocation between the AI system and humans for meaningful interactions and appropriate human oversight and control?
- Does the AI system enhance or augment human capabilities?
- Did you take safeguards to prevent overconfidence in or overreliance on the AI system for work processes?

Human oversight:

✓ Did you consider the appropriate level of human control for the particular AI system and use case?
- Can you describe the level of human control or involvement?
- Who is the "human in control" and what are the moments or tools for human intervention?
- Did you put in place mechanisms and measures to ensure human control or oversight?
- Did you take any measures to enable audit and to remedy issues related to governing AI autonomy?

✓ Is there is a self-learning or autonomous AI system or use case? If so, did you put in place more specific mechanisms of control and oversight?
- Which detection and response mechanisms did you establish to assess whether something could go wrong?

- Did you ensure a stop button or procedure to safely abort an operation where needed? Does this procedure abort the process entirely, in part, or delegate control to a human?

2. Technical robustness and safety

Resilience to attack and security:

- ✓ Did you assess potential forms of attacks to which the AI system could be vulnerable?
  - Did you consider different types and natures of vulnerabilities, such as data pollution, physical infrastructure, cyber-attacks?
- ✓ Did you put measures or systems in place to ensure the integrity and resilience of the AI system against potential attacks?
- ✓ Did you verify how your system behaves in unexpected situations and environments?
- ✓ Did you consider to what degree your system could be dual-use? If so, did you take suitable preventative measures against this case (including for instance not publishing the research or deploying the system)?

Fallback plan and general safety:

- ✓ Did you ensure that your system has a sufficient fallback plan if it encounters adversarial attacks or other unexpected situations (for example technical switching procedures or asking for a human operator before proceeding)?
- ✓ Did you consider the level of risk raised by the AI system in this specific use case?
  - Did you put any process in place to measure and assess risks and safety?
  - Did you provide the necessary information in case of a risk for human physical integrity?
  - Did you consider an insurance policy to deal with potential damage from the AI system?
  - Did you identify potential safety risks of (other) foreseeable uses of the technology, including accidental or malicious misuse? Is there a plan to mitigate or manage these risks?
- ✓ Did you assess whether there is a probable chance that the AI system may cause damage or harm to users or third parties? Did you assess the likelihood, potential damage, impacted audience and severity?
  - Did you consider the liability and consumer protection rules, and take them into account?
  - Did you consider the potential impact or safety risk to the environment or to animals?
  - Did your risk analysis include whether security or network problems such as cybersecurity hazards could pose safety risks or damage due to unintentional behaviour of the AI system?
- ✓ Did you estimate the likely impact of a failure of your AI system when it provides wrong results, becomes unavailable, or provides societally unacceptable results (for example discrimination)?
  - Did you define thresholds and did you put governance procedures in place to trigger alternative/fallback plans?
  - Did you define and test fallback plans?

Accuracy

- ✓ Did you assess what level and definition of accuracy would be required in the context of the AI system and use case?
  - Did you assess how accuracy is measured and assured?
  - Did you put in place measures to ensure that the data used is comprehensive and up to date?
  - Did you put in place measures in place to assess whether there is a need for additional data, for example to improve accuracy or to eliminate bias?
- ✓ Did you verify what harm would be caused if the AI system makes inaccurate predictions?
- ✓ Did you put in place ways to measure whether your system is making an unacceptable amount of inaccurate predictions?
- ✓ Did you put in place a series of steps to increase the system's accuracy?

Reliability and reproducibility:

- ✓ Did you put in place a strategy to monitor and test if the AI system is meeting the goals, purposes and intended applications?
  - Did you test whether specific contexts or particular conditions need to be taken into account to ensure reproducibility?
  - Did you put in place verification methods to measure and ensure different aspects of the system's reliability and reproducibility?
  - Did you put in place processes to describe when an AI system fails in certain types of settings?
  - Did you clearly document and operationalise these processes for the testing and verification of the reliability of AI systems?
  - Did you establish mechanisms of communication to assure (end-)users of the system's reliability?

3. Privacy and data governance

Respect for privacy and data Protection:

- ✓ Depending on the use case, did you establish a mechanism allowing others to flag issues related to privacy or data protection in the AI system's processes of data collection (for training and operation) and data processing?
- ✓ Did you assess the type and scope of data in your data sets (for example whether they contain personal data)?
- ✓ Did you consider ways to develop the AI system or train the model without or with minimal use of potentially sensitive or personal data?
- ✓ Did you build in mechanisms for notice and control over personal data depending on the use case (such as valid consent and possibility to revoke, when applicable)?
- ✓ Did you take measures to enhance privacy, such as via encryption, anonymisation and aggregation?
- ✓ Where a Data Privacy Officer (DPO) exists, did you involve this person at an early stage in the process?

Quality and integrity of data:

- ✓ Did you align your system with relevant standards (for example ISO, IEEE) or widely adopted protocols for daily data management and governance?
- ✓ Did you establish oversight mechanisms for data collection, storage, processing and use?
- ✓ Did you assess the extent to which you are in control of the quality of the external data sources used?
- ✓ Did you put in place processes to ensure the quality and integrity of your data? Did you consider other processes? How are you verifying that your data sets have not been compromised or hacked?

Access to data:

- ✓ What protocols, processes and procedures did you follow to manage and ensure proper data governance?
  - Did you assess who can access users' data, and under what circumstances?
  - Did you ensure that these persons are qualified and required to access the data, and that they have the necessary competences to understand the details of data protection policy?
  - Did you ensure an oversight mechanism to log when, where, how, by whom and for what purpose data was accessed?

4. Transparency

Traceability:

- ✓ Did you establish measures that can ensure traceability? This could entail documenting the following methods:
  - ➢ Methods used for designing and developing the algorithmic system:
    - Rule-based AI systems: the method of programming or how the model was built;
    - Learning-based AI systems; the method of training the algorithm, including which input data was gathered and selected, and how this occurred.

- Methods used to test and validate the algorithmic system:
  - Rule-based AI systems; the scenarios or cases used in order to test and validate;
  - Learning-based model: information about the data used to test and validate.
- Outcomes of the algorithmic system:
  - The outcomes of or decisions taken by the algorithm, as well as potential other decisions that would result from different cases (for example, for other subgroups of users).

Explainability:

✓ Did you assess:
- to what extent the decisions and hence the outcome made by the AI system can be understood?
- to what degree the system's decision influences the organisation's decision-making processes?
- why this particular system was deployed in this specific area?
- what the system's business model is (for example, how does it create value for the organisation)?

✓ Did you ensure an explanation as to why the system took a certain choice resulting in a certain outcome that all users can understand?

✓ Did you design the AI system with interpretability in mind from the start?
- Did you research and try to use the simplest and most interpretable model possible for the application in question?
- Did you assess whether you can analyse your training and testing data? Can you change and update this over time?
- Did you assess whether you can examine interpretability after the model's training and development, or whether you have access to the internal workflow of the model?

Communication:

✓ Did you communicate to (end-)users – through a disclaimer or any other means – that they are interacting with an AI system and not with another human? Did you label your AI system as such?

✓ Did you establish mechanisms to inform (end-)users on the reasons and criteria behind the AI system's outcomes?
- Did you communicate this clearly and intelligibly to the intended audience?
- Did you establish processes that consider users' feedback and use this to adapt the system?
- Did you communicate around potential or perceived risks, such as bias?
- Depending on the use case, did you consider communication and transparency towards other audiences, third parties or the general public?

✓ Did you clarify the purpose of the AI system and who or what may benefit from the product/service?
- Did you specify usage scenarios for the product and clearly communicate these to ensure that it is understandable and appropriate for the intended audience?
- Depending on the use case, did you think about human psychology and potential limitations, such as risk of confusion, confirmation bias or cognitive fatigue?

✓ Did you clearly communicate characteristics, limitations and potential shortcomings of the AI system?
- In case of the system's development: to whoever is deploying it into a product or service?
- In case of the system's deployment: to the (end-)user or consumer?

5. Diversity, non-discrimination and fairness

Unfair bias avoidance:

✓ Did you establish a strategy or a set of procedures to avoid creating or reinforcing unfair bias in the AI system, both regarding the use of input data as well as for the algorithm design?
- Did you assess and acknowledge the possible limitations stemming from the composition of the used data sets?
- Did you consider diversity and representativeness of users in the data? Did you test for specific populations or problematic use cases?
- Did you research and use available technical tools to improve your understanding of the data, model and performance?

- Did you put in place processes to test and monitor for potential biases during the development, deployment and use phase of the system?

✓ Depending on the use case, did you ensure a mechanism that allows others to flag issues related to bias, discrimination or poor performance of the AI system?
- Did you establish clear steps and ways of communicating on how and to whom such issues can be raised?
- Did you consider others, potentially indirectly affected by the AI system, in addition to the (end)-users?

✓ Did you assess whether there is any possible decision variability that can occur under the same conditions?
- If so, did you consider what the possible causes of this could be?
- In case of variability, did you establish a measurement or assessment mechanism of the potential impact of such variability on fundamental rights?

✓ Did you ensure an adequate working definition of "fairness" that you apply in designing AI systems?
- Is your definition commonly used? Did you consider other definitions before choosing this one?
- Did you ensure a quantitative analysis or metrics to measure and test the applied definition of fairness?
- Did you establish mechanisms to ensure fairness in your AI systems? Did you consider other potential mechanisms?

Accessibility and universal design:

✓ Did you ensure that the AI system accommodates a wide range of individual preferences and abilities?
- Did you assess whether the AI system usable by those with special needs or disabilities or those at risk of exclusion? How was this designed into the system and how is it verified?
- Did you ensure that information about the AI system is accessible also to users of assistive technologies?
- Did you involve or consult this community during the development phase of the AI system?

✓ Did you take the impact of your AI system on the potential user audience into account?
- Did you assess whether the team involved in building the AI system is representative of your target user audience? Is it representative of the wider population, considering also of other groups who might tangentially be impacted?
- Did you assess whether there could be persons or groups who might be disproportionately affected by negative implications?
- Did you get feedback from other teams or groups that represent different backgrounds and experiences?

Stakeholder participation:

✓ Did you consider a mechanism to include the participation of different stakeholders in the AI system's development and use?

✓ Did you pave the way for the introduction of the AI system in your organisation by informing and involving impacted workers and their representatives in advance?

6. Societal and environmental well-being

Sustainable and environmentally friendly AI:

✓ Did you establish mechanisms to measure the environmental impact of the AI system's development, deployment and use (for example the type of energy used by the data centres)?

✓ Did you ensure measures to reduce the environmental impact of your AI system's life cycle?

Social impact:

✓ In case the AI system interacts directly with humans:
- Did you assess whether the AI system encourages humans to develop attachment and empathy towards the system?
- Did you ensure that the AI system clearly signals that its social interaction is simulated and that it

has no capacities of "understanding" and "feeling"?

- ✓ Did you ensure that the social impacts of the AI system are well understood? For example, did you assess whether there is a risk of job loss or de-skilling of the workforce? What steps have been taken to counteract such risks?

Society and democracy:

- ✓ Did you assess the broader societal impact of the AI system's use beyond the individual (end-)user, such as potentially indirectly affected stakeholders?

7. Accountability

Auditability:

- ✓ Did you establish mechanisms that facilitate the system's auditability, such as ensuring traceability and logging of the AI system's processes and outcomes?
- ✓ Did you ensure, in applications affecting fundamental rights (including safety-critical applications) that the AI system can be audited independently?

Minimising and reporting negative Impact:

- ✓ Did you carry out a risk or impact assessment of the AI system, which takes into account different stakeholders that are (in)directly affected?
- ✓ Did you provide training and education to help developing accountability practices?
  - Which workers or branches of the team are involved? Does it go beyond the development phase?
  - Do these trainings also teach the potential legal framework applicable to the AI system?
  - Did you consider establishing an 'ethical AI review board' or a similar mechanism to discuss overall accountability and ethics practices, including potentially unclear grey areas?
- ✓ Did you foresee any kind of external guidance or put in place auditing processes to oversee ethics and accountability, in addition to internal initiatives?
- ✓ Did you establish processes for third parties (e.g. suppliers, consumers, distributors/vendors) or workers to report potential vulnerabilities, risks or biases in the AI system?

Documenting trade-offs:

- ✓ Did you establish a mechanism to identify relevant interests and values implicated by the AI system and potential trade-offs between them?
- ✓ How do you decide on such trade-offs? Did you ensure that the trade-off decision was documented?

Ability to redress:

- ✓ Did you establish an adequate set of mechanisms that allows for redress in case of the occurrence of any harm or adverse impact?
- ✓ Did you put mechanisms in place both to provide information to (end-)users/third parties about opportunities for redress?

We invite all stakeholders to pilot this Assessment List in practice and to provide feedback on its implementability, completeness, relevance for the specific AI application or domain, as well as overlap or complementarity with existing compliance or assessment processes. Based on this feedback, a revised version of the Trustworthy AI assessment list will be proposed to the Commission in early 2020

Key guidance derived from Chapter III:

- ✓ Adopt a Trustworthy AI **assessment list** when developing, deploying or using AI systems, and adapt it to the specific use case in which the system is being applied.
- ✓ Keep in mind that such assessment list will **never be exhaustive**. Ensuring Trustworthy AI is not about ticking boxes, but about continuously identifying requirements, evaluating solutions and ensuring improved outcomes throughout the AI system's lifecycle, and involving stakeholders therein.

C. EXAMPLES OF OPPORTUNITIES AND CRITICAL CONCERNS RAISED BY AI

In the following section, we provide examples of AI development and use that should be encouraged, as well as examples of where AI development, deployment or use can run counter to our values and may raise specific concerns. A balance must be struck between what should and what can be done with AI, and due care must be given to what should not be done with AI.

1. Examples of Trustworthy AI's opportunities

Trustworthy AI can represent a great opportunity to support the mitigation of pressing challenges facing society such as an ageing population, growing social inequality and environmental pollution. This potential is also reflected globally, such as with the UN Sustainable Development Goals.[57] The following section looks at how to encourage a European AI strategy that tackles some of these challenges.

- *Climate action and sustainable infrastructure*

While tackling climate change should be a top priority for policy-makers across the world, digital transformation and Trustworthy AI have a great potential to reduce humans' impact on the environment and enable the efficient and effective use of energy and natural resources.[58] Trustworthy AI can, for instance, be coupled to big data in order to detect energy needs more accurately, resulting in more efficient energy infrastructure and consumption.[59]

Looking at sectors like public transportation, AI systems for intelligent transport systems[60] can be used to minimise queuing, optimise routing, allow vision impaired people to be more independent,[61] optimise energy efficient engines and thereby enhance decarbonisation efforts and reduce the environmental footprint, for a greener society. Currently, worldwide, one human dies every 23 seconds in a car accident.[62] AI systems could help to reduce the number fatalities significantly, for instance through better reaction times and better adherence to rules.[63]

- *Health and well-being*

Trustworthy AI technologies can be used – and are already being used – to render treatment smarter and more targeted, and to help preventing life-threatening diseases.[64] Doctors and medical professionals can potentially perform a more accurate and detailed analysis of a patient's complex health data, even before people get sick, and provide tailored preventive treatment.[65] In the context of Europe's ageing population, AI technologies and robotics can be valuable tools to assist caregivers, support elderly care,[66] and monitor patients' conditions on a real time

---

57 https://sustainabledevelopment.un.org/?menu=1300

58 A number of EU projects aim for the development of Smart Grids and Energy Storage, which have the potential to contribute to a successful digitally supported energy transition, including through AI-based and other digital solutions. To complement the work of those individual projects, the Commission has launched the BRIDGE initiative, allowing ongoing Horizon 2020 Smart Grid and Energy Storage projects to create a common view on cross cutting issues: https://www.h2020-bridge.eu/.

59 See for instance the Encompass project: http://www.encompass-project.eu/.

60 New AI-based solutions help prepare cities for the future of mobility. See for instance the EU funded project called Fabulos: https://fabulos.eu/.

61 See for instance the PRO4VIP project, which is part of the European Vision 2020 strategy to combat preventable blindness, especially due to old age. Mobility and orientation was one of the project's priority areas.

62 https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.

63 The European UP-Drive project for instance aims to address the outlined transport-related challenges by providing contributions enabling gradual automation of and collaboration among vehicles , facilitating a safer, more inclusive and more affordable transportation system. https://up-drive.eu/.

64 See for instance the REVOLVER (Repeated Evolution of Cancer) project: https://www.healtheuropa.eu/personalised-cancer-treatment/87958/, or the Murab project which conducts more accurate biopsies, and which aims at diagnosing cancer and other illnesses faster: https://ec.europa.eu/digital-single-market/en/news/murab-eu-funded-project-success-story.

65 See for instance the Live INCITE project: www.karolinska.se/en/live-incite. This consortium of healthcare procurers challenges the industry to develop smart AI and other ICT solutions that enable lifestyle interventions in the perioperative process. The ta rget concerns new innovative eHealth solutions that can influence patients in a personalised way to take the necessary actions both prior and after surgery in their lifestyle to optimise the healthcare outcome.

66 The EU-funded project CARESSES deals with robots for elderly care, focusing on their cultural sensitivity: they adapt their way of acting and speaking to match the culture and habits of the elderly person they are assisting: http://caressesrobot.org/en/project/. See also the AI application called Alfred, a virtual assistant helping older people stay active:

basis, thus saving lives.[67]

Trustworthy AI can also assist on a broader scale. For example, it can examine and identify general trends in the healthcare and treatment sector,[68] leading to earlier detection of diseases, more efficient development of medicines, more targeted treatments[69] and ultimately more lives saved.

- *Quality education and digital transformation*

New technological, economic and environmental changes mean that society needs to become more proactive. Governments, industry leaders, educational institutions and unions face a responsibility to bring the citizens into the new digital era ensuring they have the right skills to fill the future jobs. Trustworthy AI technologies could assist in more accurately forecasting which jobs and professions will be disrupted by technology, which new roles will be created and which skills will be needed. This could help governments, unions and industry with planning the (re)skilling of workers. It could also give citizens who may fear redundancy a path of development into a new role.

In addition, AI can be a great tool to fight educational inequalities and create personalised and adaptable education programmes that could help everyone acquire new qualifications, skills and competences according to his or her own ability to learn.[70] It could increase both the learning speed and the quality of education – reaching from primary school to university.

2. Examples of critical concerns raised by AI

A critical AI concern arises one of the components of Trustworthy AI is violated. Many of the concerns listed below will already fall within the scope of existing legal requirements, which are mandatory and must therefore be complied with. Yet even in circumstances where compliance with legal requirements has been demonstrated, these may not address the full range of ethical concerns that may arise. As our understanding of the adequacy of rules and ethical principles invariably evolves and may change over time, the following non-exhaustive list of concerns may be shortened, expanded, edited or updated in the future.

- *Identifying and tracking individuals with AI*

AI enables the ever more efficient identification of individual persons by both public and private entities. Noteworthy examples of a scalable AI identification technology are face recognition and other involuntary methods of identification using biometric data (i.e. lie detection, personality assessment through micro expressions, and automatic voice detection). Identification of individuals is sometimes the desirable outcome, aligned with ethical principles (for example in detecting fraud, money laundering, or terrorist financing). However, automatic identification raises strong concerns of both a legal and ethical nature, as it may have an unexpected impact on many psychological and sociocultural levels. A proportionate use of control techniques in AI is needed to uphold the autonomy of European citizens. Clearly defining if, when and how AI can be used for automated identification of individuals and differentiating between the identification of an individual vs the tracing and tracking of an individual,

---

https://ec.europa.eu/digital-single-market/en/news/alfred-virtual-assistant-helping-older-people-stay-active. Moreover, the EMPATTICS project (EMpowering PAtients for a BeTTer Information and improvement of the Communication Systems) will research and define how health care professionals and patients use ICT technologies including AI systems to plan interventions with patients and to monitor the progression of their physical and mental state: www.empattics.eu.

67 See for instance the MyHealth Avatar (www.myhealthavatar.eu), which offers a digital representation of a patient's health status. The research project launched an app and an online platform that collects, and gives access to, your digital long-term health-status information. This takes on the form of a life-long health companion ('avatar'). MyHealthAvatar also predicts your risk for stroke, diabetes, cardiovascular disease and hypertension.

68 See for instance the ENRICHME project (www.enrichme.eu), which tackles the progressive decline of cognitive capacity in the ageing population. An integrated platform for Ambient Assisted Living (AAL) and a mobile service robot for long-term monitoring and interaction will help the elderly to remain independent and active for longer.

69 See for instance the use of AI by Sophia Genetics, which leverages statistical inference, pattern recognition and machine learning to maximize the value of genomics and radiomics data: https://www.sophiagenetics.com/home.html.

70 See for instance the MaTHiSiS project, aimed at providing a solution for affect-based learning in a comfortable learning environment, comprising of high-end technological devices and algorithms: (http://mathisis-project.eu/). See also IBM's Watson Classroom or Century Tech's platform.

and between targeted surveillance and mass surveillance, will be crucial for the achievement of Trustworthy AI. The application of such technologies must be clearly warranted in existing law.[71] Where the legal basis for such activity is "consent", practical means[72] must be developed which allow meaningful and verified consent to be given to being automatically identified by AI or equivalent technologies. This also applies to the usage of "anonymous" personal data that can be re-personalised.

- *Covert AI systems*

Human beings should always know if they are directly interacting with another human being or a machine, and it is the responsibility of AI practitioners that this is reliably achieved. AI practitioners should therefore ensure that humans are made aware of – or able to request and validate the fact that – they interact with an AI system (for instance, by issuing clear and transparent disclaimers). Note that borderline cases exist and complicate the matter (e.g. an AI-filtered voice spoken by a human). It should be borne in mind that the confusion between humans and machines could have multiple consequences such as attachment, influence, or reduction of the value of being human.[73] The development of human-like robots[74] should therefore undergo careful ethical assessment.

- *AI enabled citizen scoring in violation of fundamental rights*

Societies should strive to protect the freedom and autonomy of all citizens. Any form of citizen scoring can lead to the loss of this autonomy and endanger the principle of non-discrimination. Scoring should only be used if there is a clear justification, and where measures are proportionate and fair. Normative citizen scoring (general assessment of "moral personality" or "ethical integrity") in *all* aspects and on a large scale by public authorities or private actors endangers these values, especially when used not in accordance with fundamental rights, and when used disproportionately and without a delineated and communicated legitimate purpose.

Today, citizen scoring – on a large or smaller scale – is already often used in purely descriptive and domain-specific scorings (e.g. school systems, e-learning, and driver licences). Even in those more narrow applications, a fully transparent procedure should be made available to citizens, including information on the process, purpose and methodology of the scoring. Note that transparency cannot prevent non-discrimination or ensure fairness, and is not the panacea against the problem of scoring. Ideally the possibility of opting out of the scoring mechanism when possible without detriment should be provided – otherwise mechanisms for challenging and rectifying the scores must be given. This is particularly important in situations where an asymmetry of power exists between the parties. Such opt-out options should be ensured in the technology's design in circumstances where this is necessary to ensure compliance with fundamental rights and is necessary in a democratic society.

- *Lethal autonomous weapon systems (LAWS)*

Currently, an unknown number of countries and industries are researching and developing lethal autonomous weapon systems, ranging from missiles capable of selective targeting to learning machines with cognitive skills to decide whom, when and where to fight without human intervention. This raises fundamental ethical concerns, such as the fact that it could lead to an uncontrollable arms race on a historically unprecedented level, and create military contexts in which human control is almost entirely relinquished and the risks of malfunction are not addressed. The European Parliament has called for the urgent development of a common, legally binding position addressing ethical and legal questions of human control, oversight, accountability and implementation of international human rights law, international humanitarian law and military strategies.[75] Recalling the European Union's aim to promote peace as enshrined in Article 3 of the Treaty of the European Union, we stand with, and look to support, the Parliament's resolution of 12 September 2018 and all related efforts on LAWS.

---

71 In this regard, Article 6 of the GDPR can be recalled, which provides, among other things, that processing of data shall only be lawful if it has a valid legal basis.

72 As current mechanisms for giving informed consent in the internet show, consumers typically give consent without meaningful consideration. Hence, they can hardly be classified as practical.

73 Madary & Metzinger (2016). Real Virtuality: A Code of Ethical Conduct. Recommendations for Good Scientific Practice and the Consumers of VR-Technology. Frontiers in Robotics and AI, 3(3).

74 This also applies to AI-driven avatars.

75 European Parliament's Resolution 2018/2752(RSP).

- *Potential longer-term concerns*

AI development is still domain-specific and requires well-trained human scientists and engineers to precisely specify its targets. However, extrapolating into the future with a longer time horizon, certain critical long-term concerns can be hypothesized.[76] A risk-based approach suggests that these concerns should be kept into consideration in view of possible unknown unknowns and "black swans."[77] The high-impact nature of these concerns, combined with the current uncertainty in corresponding developments, calls for regular assessments of these topics.

D. CONCLUSION

This document constitutes the AI Ethics Guidelines produced by the High-Level Expert Group on Artificial Intelligence (AI HLEG).

We recognise the positive impact that AI systems already have and will continue having, both commercially and societally. However, we are equally concerned to ensure that the risks and other adverse impacts with which these technologies are associated are properly and proportionately handled. AI is a technology that is both transformative and disruptive, and its evolution over the last several years has been facilitated by the availability of enormous amounts of digital data, major technological advances in computational power and storage capacity, as well as significant scientific and engineering innovation in AI methods and tools. AI systems will continue to impact society and citizens in ways that we cannot yet imagine.

In this context, it is important to build AI systems that are worthy of trust, since human beings will only be able to confidently and fully reap its benefits when the technology, including the processes and people behind the technology, are trustworthy. When drafting these Guidelines, Trustworthy AI has, therefore, been our foundational ambition.

Trustworthy AI has three components: (1) it should be lawful, ensuring compliance with all applicable laws and regulations, (2) it should be ethical, ensuring adherence to ethical principles and values and (3) it should be robust, both from a technical and social perspective since to ensure that, even with good intentions, AI systems do not cause any unintentional harm. Each component is necessary but not sufficient to achieve Trustworthy AI. Ideally, all three components work in harmony and overlap in their operation. Where tensions arise, we should endeavour to align them.

In Chapter I, we articulated the fundamental rights and a corresponding set of ethical principles that are crucial in an AI-context. In Chapter II, we listed seven key requirements that AI systems should meet in order to realise Trustworthy AI. We proposed technical and non-technical methods that can help with their implementation. Finally, in Chapter III we provided a Trustworthy AI assessment list that can help operationalising the seven requirements. In a final section, we provided examples of beneficial opportunities and critical concerns raised by AI systems, on which we hope to stimulate further discussion.

Europe has a unique vantage point based on its focus on placing the citizen at the heart of its endeavours. This focus is written into the very DNA of the European Union through the Treaties upon which it is built. The current document forms part of a vision that promotes Trustworthy AI which we believe should be the foundation upon which Europe can build leadership in innovative, cutting-edge AI systems. This ambitious vision will help securing human flourishing of European citizens, both individually and collectively. Our goal is to create a culture of "Trustworthy AI for Europe", whereby the benefits of AI can be reaped by all in a manner that ensures respect for our foundational values: fundamental rights, democracy and the rule of law.

---

[76] While some consider that Artificial General Intelligence, Artificial Consciousness, Artificial Moral Agents, Super-intelligence or Transformative AI can be examples of such long-term concerns (currently non-existent), many others believe these to be unrealistic.

[77] A black swan event is a very rare, yet high impact, event – so rare, that it might not have been observed. Hence, probability of occurrence typically can only be estimated with high uncertainty.

GLOSSARY

This glossary pertains to the Guidelines and is meant to help in the understanding of the terms used in this document.

Artificial Intelligence or AI systems

Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans[78] that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions.

As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems).

A separate document prepared by the AI HLEG and elaborating on the definition of AI used for the purpose of this document is published in parallel, titled "A definition of AI: Main capabilities and scientific disciplines".

AI Practitioners

By AI practitioners we denote all individuals or organisations that develop (including research, design or provide data for) deploy (including implement) or use AI systems, excluding those that use AI systems in the capacity of end-user or consumer.

AI system's life cycle

An AI system's life cycle encompasses its development (including research, design, data provision, and limited trials), deployment (including implementation) and use phase.

Auditability

Auditability refers to the ability of an AI system to undergo the assessment of the system's 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. Ensuring traceability and logging mechanisms from the early design phase of the AI system can help enabling the system's auditability.

Bias

Bias is an inclination of prejudice towards or against a person, object, or position. Bias can arise in many ways in AI systems. For example, in data-drive AI systems, such as those produced through machine learning, bias in data collection and training can result in an AI system demonstrating bias. In logic-based AI, such as rule-based systems, bias can arise due to how a knowledge engineer might view the rules that apply in a particular setting. Bias can also arise due to online learning and adaptation through interaction. It can also arise through personalisation whereby users are presented with recommendations or information feeds that are tailored to the user's tastes. It does not necessarily relate to human bias or human-driven data collection. It can arise, for example, through the limited contexts in which a system in used, in which case there is no opportunity to generalise it to other contexts. Bias can be good or bad, intentional or unintentional. In certain cases, bias can result in discriminatory and/or unfair outcomes, indicated in this document as unfair bias.

[78] Humans design AI systems directly, but they may also use AI techniques to optimise their design.

Ethics

Ethics is an academic discipline which is a subfield of philosophy. In general terms, it deals with questions like "What is a good action?", "What is the value of a human life?", "What is justice?", or "What is the good life?". In academic ethics, there are four major fields of research: (i) Meta-ethics, mostly concerning the meaning and reference of normative sentence, and the question how their truth values can be determined (if they have any); (ii) normative ethics, the practical means of determining a moral course of action by examining the standards for right and wrong action and assigning a value to specific actions; (iii) descriptive ethics, which aims at an empirical investigation of people's moral behaviour and beliefs; and (iv) applied ethics, concerning what we are obligated (or permitted) to do in a specific (often historically new) situation or a particular domain of (often historically unprecedented) possibilities for action. Applied ethics deals with real-life situations, where decisions have to be made under time-pressure, and often limited rationality. AI Ethics is generally viewed as an example of applied ethics and focuses on the normative issues raised by the design, development, implementation and use of AI.

Within ethical discussions, the terms "moral" and "ethical" are often used. The term "moral" refers to the concrete, factual patterns of behaviour, the customs, and conventions that can be found in specific cultures, groups, or individuals at a certain time. The term "ethical" refers to an evaluative assessment of such concrete actions and behaviours from a systematic, academic perspective.

Ethical AI

In this document, ethical AI is used to indicate the development, deployment and use of AI that ensures compliance with ethical norms, including fundamental rights as special moral entitlements, ethical principles and related core values. It is the second of the three core elements necessary for achieving Trustworthy AI.

Human-Centric AI

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 enjoy 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.

Red Teaming

Red teaming is the practice whereby a "red team" or independent group challenges an organisation to improve its effectiveness by assuming an adversarial role or point of view. It is particularly used to help identifying and addressing potential security vulnerabilities.

Reproducibility

Reproducibility describes whether an AI experiment exhibits the same behaviour when repeated under the same conditions.

Robust AI

Robustness of an AI system encompasses both its technical robustness (appropriate in a given context, such as the application domain or life cycle phase) and as well as its robustness from a social perspective (ensuring that the AI system duly takes into account the context and environment in which the system operates). This is crucial to ensure that, even with good intentions, no unintentional harm can occur. Robustness is the third of the three components necessary for achieving Trustworthy AI.

Stakeholders

By stakeholders we denote all those that research develop, design, deploy or use AI, as well as those that are (directly or indirectly) affected by AI – including but not limited to companies, organisations, researchers, public services, institutions, civil society organisations, governments, regulators, social partners, individuals, citizens,

workers and consumers.

Traceability

Traceability of an AI system refers to the capability to keep track of the system's data, development and deployment processes, typically by means of documented recorded identification.

Trust

We take the following definition from the literature: "Trust is viewed as: (1) a set of specific beliefs dealing with benevolence, competence, integrity, and predictability (trusting beliefs); (2) the willingness of one party to depend on another in a risky situation (trusting intention); or (3) the combination of these elements."[79] While "Trust" is usually not a property ascribed to machines, this document aims to stress the importance of being able to trust not only in the fact that AI systems are legally compliant, ethically adherent and robust, but also that such trust can be ascribed to all people and processes involved in the AI system's life cycle.

Trustworthy AI

Trustworthy AI has three components: (1) it should be lawful, ensuring compliance with all applicable laws and regulations (2) it should be ethical, demonstrating respect for, and ensure adherence to, ethical principles and values and (3) it should be robust, both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm. Trustworthy AI concerns not only the trustworthiness of the AI system itself but also comprises the trustworthiness of all processes and actors that are part of the system's life cycle.

Vulnerable Persons and Groups

No commonly accepted or widely agreed legal definition of vulnerable persons exists, due to their heterogeneity. What constitutes a vulnerable person or group is often context-specific. Temporary life events (such as childhood or illness), market factors (such as information asymmetry or market power), economic factors (such as poverty), factors linked to one's identity (such as gender, religion or culture) or other factors can play a role. The Charter of Fundamental Rights of the EU encompasses under Article 21 on non-discrimination the following grounds, which can be a reference point amongst others: namely sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age and sexual orientation. Other articles of law address the rights of specific groups, in addition to those listed above. Any such list is not exhaustive, and may change over time. A vulnerable group is a group of persons who share one or several characteristics of vulnerability.

[79] Siau, K., Wang, W. (2018), Building Trust in Artificial Intelligence, Machine Learning, and Robotics, *CUTTER BUSINESS TECHNOLOGY JOURNAL* (31), S. 47–53.

This Document was prepared by the members of the High-Level Expert Group on AI

listed here below in alphabetical order

Pekka Ala-Pietilä, Chair of the AI HLEG
AI Finland, Huhtamaki, Sanoma
Wilhelm Bauer
Fraunhofer
Urs Bergmann – Co-Rapporteur
Zalando
Mária Bieliková
Slovak University of Technology in Bratislava
Cecilia Bonefeld-Dahl – Co-Rapporteur
DigitalEurope
Yann Bonnet
ANSSI
Loubna Bouarfa
OKRA
Stéphan Brunessaux
Airbus
Raja Chatila
IEEE Initiative Ethics of Intelligent/Autonomous Systems &
Sorbonne University
Mark Coeckelbergh
University of Vienna
Virginia Dignum – Co-Rapporteur
Umea University
Luciano Floridi
University of Oxford
Jean-Francois Gagné – Co-Rapporteur
Element AI
Chiara Giovannini
ANEC
Joanna Goodey
Fundamental Rights Agency
Sami Haddadin
Munich School of Robotics and MI
Gry Hasselbalch
The thinkdotank DataEthics & Copenhagen University
Fredrik Heintz
Linköping University
Fanny Hidvegi
Access Now
Eric Hilgendorf
University of Würzburg
Klaus Höckner
Hilfsgemeinschaft der Blinden und Sehschwachen
Mari-Noëlle Jégo-Laveissière
Orange
Leo Kärkkäinen
Nokia Bell Labs
Sabine Theresia Köszegi
TU Wien
Robert Kroplewski
Solicitor & Advisor to Polish Government
Elisabeth Ling
RELX
Pierre Lucas
Orgalim – Europe's technology industries
Ieva Martinkenaite
Telenor
Thomas Metzinger – Co-Rapporteur
JGU Mainz & European University Association
Catelijne Muller
ALLAI Netherlands & EESC
Markus Noga
SAP
Barry O'Sullivan, Vice-Chair of the AI HLEG
University College Cork
Ursula Pachl
BEUC
Nicolas Petit – Co-Rapporteur
University of Liège
Christoph Peylo
Bosch
Iris Plöger
BDI
Stefano Quintarelli
Garden Ventures
Andrea Renda
College of Europe Faculty & CEPS
Francesca Rossi
IBM
Cristina San José
European Banking Federation
George Sharkov
Digital SME Alliance
Philipp Slusallek
German Research Centre for AI (DFKI)
Françoise Soulié Fogelman
AI Consultant
Saskia Steinacker – Co-Rapporteur
Bayer
Jaan Tallinn
Ambient Sound Investment
Thierry Tingaud
STMicroelectronics
Jakob Uszkoreit
Google
Aimee Van Wynsberghe – Co-Rapporteur
TU Delft
Thiébaut Weber
ETUC
Cecile Wendling
AXA
Karen Yeung – Co-Rapporteur
The University of Birmingham

Urs Bergmann, Cecilia Bonefeld-Dahl, Virginia Dignum, Jean-François Gagné, Thomas Metzinger, Nicolas Petit, Saskia Steinacker, Aimee Van Wynsberghe and Karen Yeung acted as rapporteurs for this document.

Pekka Ala-Pietilä is Chairing the AI HLEG. Barry O'Sullivan is Vice-Chair, coordinating the AI HLEG's second deliverable. Nozha Boujemaa, Vice-Chair until 1 February 2019 coordinating the first deliverable, also contributed to the content of this document.

Nathalie Smuha provided editorial support.