European Commission

The European Commission's

HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE

DRAFT ETHICS GUIDELINES FOR TRUSTWORTHY AI

Working Document for stakeholders' consultation

Brussels, 18 December 2018

High-Level Expert Group on Artificial Intelligence
Draft Ethics Guidelines for Trustworthy AI

European Commission
Directorate-General for Communication

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

European Commission
B-1049 Brussels

Document made public on 18 December 2018.

This working document was produced by the AI HLEG without prejudice to the individual position of its members on specific points, and without prejudice to the final version of the document. This document will still be further developed and a final version thereof will be presented in March 2019 following the stakeholder consultation through the European AI Alliance.

Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use which might be made of the following information. The contents of this working document are the sole responsibility of the High-Level Expert Group on Artificial Intelligence (AI HLEG). Although staff of the Commission services facilitated the preparation of the Guidelines, the views expressed in this document reflect the opinion of the AI HLEG, and may not in any circumstances be regarded as stating an official position of the European Commission. This is a draft of the first Deliverable of the AI HLEG. A final version thereof will be presented to the Commission in March 2019. A final version of the second Deliverable – the AI Policy and Investment Recommendations – will be presented mid-2019.

More information on the High-Level Expert Group on Artificial Intelligence is available online (https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence).

DRAFT ETHICS GUIDELINES

FOR TRUSTWORTHY AI

TABLE OF CONTENTS

EXECUTIVE SUMMARY

This working document constitutes a draft of the AI Ethics Guidelines produced by the European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG), of which a final version is due in March 2019.

Artificial Intelligence (AI) is one of the most transformative forces of our time, and is bound to alter the fabric of society. It presents a great opportunity to increase prosperity and growth, which Europe must strive to achieve. Over the last decade, major advances were realised due to the availability of vast amounts of digital data, powerful computing architectures, and advances in AI techniques such as machine learning. Major AI-enabled developments in autonomous vehicles, healthcare, home/service robots, education or cybersecurity are improving the quality of our lives every day. Furthermore, AI is key for addressing many of the grand challenges facing the world, such as global health and wellbeing, climate change, reliable legal and democratic systems and others expressed in the United Nations Sustainable Development Goals.

Having the capability to generate tremendous benefits for individuals and society, AI also gives rise to certain risks that should be properly managed. Given that, on the whole, AI's benefits outweigh its risks, we must ensure to follow the road that **maximises the benefits of AI while minimising its risks**. To ensure that we stay on the right track, a **human-centric approach to AI is needed,** forcing us to keep in mind that the development and use of AI should not be seen as a means in itself, but as having the goal to increase human well-being. **Trustworthy AI will be our north star**, since human beings will only be able to confidently and fully reap the benefits of AI if they can trust the technology.

Trustworthy AI has **two components:** (1) it should respect fundamental rights, applicable regulation and core principles and values, ensuring an "**ethical purpose**" and (2) it should be **technically robust** and reliable since, even with good intentions, a lack of technological mastery can cause unintentional harm.

These Guidelines therefore set out a **framework for Trustworthy AI:**

- **Chapter I** deals with **ensuring AI's ethical purpose**, by setting out the fundamental rights, principles and values that it should comply with.
- From those principles, **Chapter II** derives **guidance on the realisation** of Trustworthy AI, tackling both ethical purpose and technical robustness. This is done by listing the requirements for Trustworthy AI and offering an overview of technical and non-technical methods that can be used for its implementation.
- **Chapter III** subsequently **operationalises** the requirements by providing a concrete but non-exhaustive assessment list for Trustworthy AI. This list is then adapted to specific use cases.

In contrast to other documents dealing with ethical AI, the Guidelines hence do not aim to provide yet another list of core values and principles for AI, but rather offer guidance on the concrete implementation and operationalisation thereof into AI systems. Such guidance is provided in three layers of abstraction, from most abstract in Chapter I (fundamental rights, principles and values), to most concrete in Chapter III (assessment list).

The Guidelines are addressed to all **relevant stakeholders developing, deploying or using AI**, encompassing companies, organisations, researchers, public services, institutions, individuals or other entities. In the final version of these Guidelines, a mechanism will be put forward to allow stakeholders to voluntarily endorse them.

Importantly, these Guidelines are not intended as a substitute to any form of policymaking or regulation (to be dealt with in the AI HLEG's second deliverable: the Policy & Investment Recommendations, due in May 2019), nor do they aim to deter the introduction thereof. Moreover, the Guidelines should be seen as a living document that needs to be regularly updated over time to ensure continuous relevance as the technology and our knowledge thereof, evolves. This document should therefore be a starting point for the discussion on "**Trustworthy AI made in Europe"**.

While Europe can only broadcast its ethical approach to AI when competitive at global level, an **ethical approach to AI is key to enable responsible competitiveness**, as it will generate user trust and facilitate broader uptake of AI. These Guidelines are not meant to stifle AI innovation in Europe, but instead aim to use ethics as inspiration to develop a unique brand of AI, one that aims at protecting and benefiting both individuals and the common good. This allows Europe to position itself as a leader in cutting-edge, secure and ethical AI. Only by ensuring trustworthiness will European citizens fully reap AI's benefits.

Finally, beyond Europe, these Guidelines also aim to **foster reflection and discussion** on an ethical framework for AI at **global level**.

EXECUTIVE GUIDANCE

Each Chapter of the Guidelines offers guidance on achieving Trustworthy AI, addressed to all relevant stakeholders developing, deploying or using AI, summarised here below:

Chapter I: Key Guidance for Ensuring Ethical Purpose:

- Ensure that AI is **human-centric**: AI should be developed, deployed and used with an "**ethical purpose"**, grounded in, and reflective of, fundamental rights, societal values and the ethical principles of *Beneficence* (do good), *Non-Maleficence* (do no harm), *Autonomy of humans, Justice,* and *Explicability.* This is crucial to work towards **Trustworthy AI**.
- Rely on fundamental rights, ethical principles and values to prospectively evaluate possible effects of AI on human beings and the common good. Pay **particular attention** to situations involving more **vulnerable groups** such as children, persons with disabilities or minorities, or to situations with **asymmetries of power or information**, such as between employers and employees, or businesses and consumers.
- Acknowledge and be aware of the fact that, while bringing substantive benefits to individuals and society, AI can also have a negative impact. Remain vigilant for areas of critical concern.

Chapter II: Key Guidance for Realising Trustworthy AI:

- Incorporate the **requirements for Trustworthy** AI **from the earliest design phase**: Accountability, Data Governance, Design for all, Governance of AI Autonomy (Human oversight), Non-Discrimination, Respect for Human Autonomy, Respect for Privacy, Robustness, Safety, Transparency.
- Consider technical and non-technical methods to ensure the implementation of those requirements into the AI system. Moreover, keep those requirements in mind when building the team to work on the system, the system itself, the testing environment and the potential applications of the system.

- Provide, in a clear and proactive manner, **information to stakeholders** (customers, employees, etc.) about the AI system's capabilities and limitations, allowing them to set realistic expectations. Ensuring **Traceability** of the AI system is key in this regard.
- Make Trustworthy AI **part of the organisation's culture**, and provide information to stakeholders on how Trustworthy AI is implemented into the design and use of AI systems. Trustworthy AI can also be included in organisations' deontology charters or codes of conduct.
- Ensure participation and **inclusion of stakeholders** in the design and development of the AI system. Moreover, ensure **diversity** when setting up the teams developing, implementing and testing the product.
- Strive to **facilitate the auditability** of AI systems, particularly in critical contexts or situations. To the extent possible, design your system to enable tracing individual decisions to your various inputs; data, pre-trained models, etc. Moreover, define **explanation methods** of the AI system.
- Ensure a specific process for **accountability governance**.
- Foresee **training and education**, and ensure that managers, developers, users and employers are aware of and are trained in Trustworthy AI.
- Be mindful that there might be fundamental tensions between different objectives (transparency can open the door to misuse; identifying and correcting bias might contrast with privacy protections). Communicate and document these trade-offs.
- Foster research and innovation to further the achievement of the requirements for Trustworthy AI.

Chapter III: Key Guidance for Assessing Trustworthy AI

- Adopt an **assessment list** for Trustworthy AI when developing, deploying or using AI, and adapt it to the specific use case in which the system is being used.
- Keep in mind that an assessment list will **never be exhaustive**, and that ensuring Trustworthy AI is not about ticking boxes, but about a continuous process of identifying requirements, evaluating solutions and ensuring improved outcomes throughout the entire lifecycle of the AI system.

This guidance forms part of a vision embracing a human-centric approach to Artificial Intelligence, which will enable Europe to become a globally leading innovator in ethical, secure and cutting-edge AI. It strives to facilitate and enable **"Trustworthy AI made in Europe"** which will enhance the well-being of European citizens.

GLOSSARY

This glossary is still incomplete and will be further complemented in the final version of the Document.

Artificial Intelligence or AI:

Artificial intelligence (AI) refers to systems designed by humans that, given a complex goal, act in the physical or digital world by perceiving their environment, interpreting the collected structured or unstructured data, reasoning on the knowledge derived from this data and deciding the best action(s) to take (according to pre-defined parameters) to achieve the given goal. AI systems can also be designed to learn to 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 elaborating on the definition of AI that is used for the purpose of this working document is published in parallel to this draft.

Bias:

Bias is a prejudice for or against something or somebody, that may result in unfair decisions. It is known that humans are biased in their decision making. Since AI systems are designed by humans, it is possible that humans inject their bias into them, even in an unintended way. Many current AI systems are based on machine learning data-driven techniques. Therefore a predominant way to inject bias can be in the collection and selection of training data. If the training data is not inclusive and balanced enough, the system could learn to make unfair decisions. At the same time, AI can help humans to identify their biases, and assist them in making less biased decisions.

Ethical Purpose:

In this document, ethical purpose is used to indicate the development, deployment and use of AI which ensures compliance with fundamental rights and applicable regulation, as well as respecting core principles and values. This is one of the two core elements to achieve Trustworthy AI.

Human-Centric AI:

The human-centric approach to AI strives to ensure that human values are always the primary consideration, and forces us to keep in mind that the development and use of AI should not be seen as a means in itself, but with the goal of increasing citizen's well-being.

Trustworthy AI:

Trustworthy AI has two components: (1) its development, deployment and use should comply with fundamental rights and applicable regulation as well as respecting core principles and values, ensuring "**ethical purpose**", and (2) it should be **technically robust** and reliable.

A. RATIONALE AND FORESIGHT OF THE GUIDELINES

In its Communications of 25 April 2018 and 7 December 2018, the European Commission (the Commission) set out its vision for Artificial Intelligence (AI), which supports ethical, secure and cutting-edge AI "made in Europe". 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 thereof, the Commission established the High-Level Expert Group on Artificial Intelligence (AI HLEG) and mandated it with the drafting of two deliverables: (1) AI Ethics Guidelines and (2) Policy and Investment Recommendations. This **working document** constitutes the first draft of the AI Ethics Guidelines prepared by the AI HLEG.

Over the past months, the 52 of us met, discussed and interacted at various meetings, committed to the European motto: united in diversity.

Numerous academic and journalistic publications have shown the positives and negatives related to the design, development, use, and implementation of AI in the last year. The AI HLEG is convinced that AI holds the promise to increase human wellbeing and the common good but to do this it needs to be **human-centric** and respectful of fundamental rights. In a context of rapid technological change, we believe it is essential that trust remains the cement of societies, communities, economies and sustainable development. We therefore **set Trustworthy AI as our north star.**

This working document articulates a framework for Trustworthy AI that requires **ethical purpose** and **technical robustness**. Those two components are critical to enable responsible competitiveness, as it will generate user trust and, hence, facilitate AI's uptake.

This is the path that we believe Europe should follow to position itself as a home and leader to cutting-edge, secure and ethical technology.

And this is how, as European citizens, we will fully reap the benefits of AI.

Trustworthy AI

Artificial Intelligence helps improving our quality of life through personalised medicine or more efficient delivery of healthcare services. It can help achieving the sustainable development goals such as promoting gender balance, tackling climate change, and helping us make better use of natural resources. It helps optimising our transportation infrastructures and mobility as well as supporting our ability to monitor progress against indicators of sustainability and social coherence. AI is thus not an end in itself, but rather a means to increase individual and societal well-being.

In Europe, we want to achieve such ends through Trustworthy AI. Trust is a prerequisite for people and societies to develop, deploy and use Artificial Intelligence. Without AI being demonstrably worthy of trust, subversive consequences may ensue and its uptake by citizens and consumers might be hindered, hence undermining the realisation of AI's vast economic and social benefits. To ensure those benefits, our vision is to use ethics to inspire trustworthy development, deployment and use of AI. The aim is to foster a climate most favourable to AI's beneficial innovation and uptake.

Trust in AI includes: trust in the **technology,** through the way it is built and used by humans beings; trust in the **rules, laws and norms** that govern AI – it should be noted that no legal vacuum currently exists, as Europe already has regulation in place that applies to AI – or trust in the **business** and **public governance models** of AI services, products and manufacturers.

Trustworthy AI has two components: **(1)** its development, deployment and use should **respect fundamental rights and applicable regulation, as well as core principles and values, ensuring an "ethical purpose",** and **(2)** it should be **technically robust and reliable.** Indeed, even with good intentions or purpose, the lack of technological mastery can cause unintentional harm. Moreover, compliance with fundamental rights, principles and values entails that these are duly operationalised by implementing them throughout the AI technology's design, development, and deployment. Such implementation can be addressed both by technical and non-technical methods.

The Guidelines therefore offer a **framework for Trustworthy AI** that tackles all those aspects.

The Role of AI Ethics

The achievement of Trustworthy AI draws heavily on the field of ethics. Ethics as a field of study is centuries old and centres on questions like '*what is a good*' action, '*what is right'*, and in some instances '*what is the good life*'. AI Ethics is a sub-field of applied ethics and technology, and focuses on the ethical issues raised by the design, development, implementation and use of AI. The goal of AI ethics is to identify how AI can advance or raise concerns to the good life of individuals, whether this be in terms of quality of life, mental autonomy or freedom to live in a democratic society. It concerns itself with issues of diversity and inclusion (with regards to training data and the ends to which AI serves) as well as issues of distributive justice (who will benefit from AI and who will not).

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 and implementational details that cannot be captured in general Guidelines. This document should thus not be seen as an end point, but rather as the *beginning* of a new and open-ended process of discussion. We therefore assert that our European AI Ethics Guidelines should be read as a starting point for the debate on Trustworthy AI. The discussion begins here but by no means ends here.

Purpose and Target Audience of the Guidelines

These Guidelines offer guidance to stakeholders on how Trustworthy AI can be achieved. **All relevant stakeholders that develop, deploy or use AI** – companies, organisations, researchers, public services, institutions, individuals or other entities – are addressees. In addition to playing a regulatory role, governments can also develop, deploy or use AI and thus be considered as addressees.

A mechanism will be put in place that enables all stakeholders to formally endorse and sign up to the Guidelines on a voluntary basis. This will be set out in the final version of the document.

Scope of the Guidelines

A primordial and underlying assumption of this working document is that AI developers, deployers and users comply with fundamental rights and with all applicable regulations. Compliance with these Guidelines in no way replaces compliance with the former, but merely offers a complement thereto.

The Guidelines are not an official document from the European Commission and are not legally binding. They are **neither intended as a substitute to any form of policy-making or regulation**, nor are they intended to deter from the creation thereof.

While the Guidelines' scope covers AI applications in general, it should be borne in mind that **different situations raise different challenges**. AI systems recommending songs to citizens do not raise the same sensitivities as AI systems recommending a critical medical treatment. Likewise, different opportunities and challenges arise from AI systems used in the context of business-to-consumer, business-to-business or public-to-citizen relationships, or – more generally – in different sectors or use cases. It is, therefore, explicitly acknowledged that a **tailored approach is needed given AI's context-specificity**.

B. A FRAMEWORK FOR TRUSTWORTHY AI

These draft AI Ethics Guidelines **consist of three chapters** – each offering guidance on a further level of abstraction – together constituting a **framework for achieving Trustworthy AI**:

(I) **Ethical Purpose**. This Chapter focuses on the core values and principles that all those dealing with AI should comply with. These are based on international human rights law, which at EU level is enshrined in the values and rights prescribed in the EU Treaties and in the Charter of Fundamental Rights of the European Union. Together, this section can be coined as governing the "**ethical purpose**" of developers, deployers and users of AI, which should consist of **respect for the rights, principles and values** laid out therein. In addition, a number of **areas of specific concern** are listed, where it is considered that the use of AI may breach such ethical purpose.

(II) **Realisation of Trustworthy AI**. Mere good intentions are not enough. It is important that AI developers, deployers and users also take actions and responsibility to **actually implement these principles** and values into the technology and its use. Moreover, they should take precautions that the systems are as robust as possible from a technical point of view, to ensure that – even if the ethical purpose is respected – AI does not cause unintentional harm. Chapter II therefore identifies the requirements for Trustworthy AI and offers guidance on the potential methods – both technical and non-technical – that can be used to realise it.

(III) **Assessment List & Use Cases.** Based on the ethical purpose set out in Chapter I, and the implementation methods of Chapter II, Chapter III sets out a preliminary and **non-exhaustive assessment list for** AI developers, deployers and users to **operationalise Trustworthy AI**. Given the application-specificity of AI, the assessment list will need to be tailored to specific applications, contexts or sectors. We selected number of use cases to provide an example of such context-specific assessment list, which will be developed in the final version of the document.

This Guidelines' structure is illustrated in *Figure 1* below.

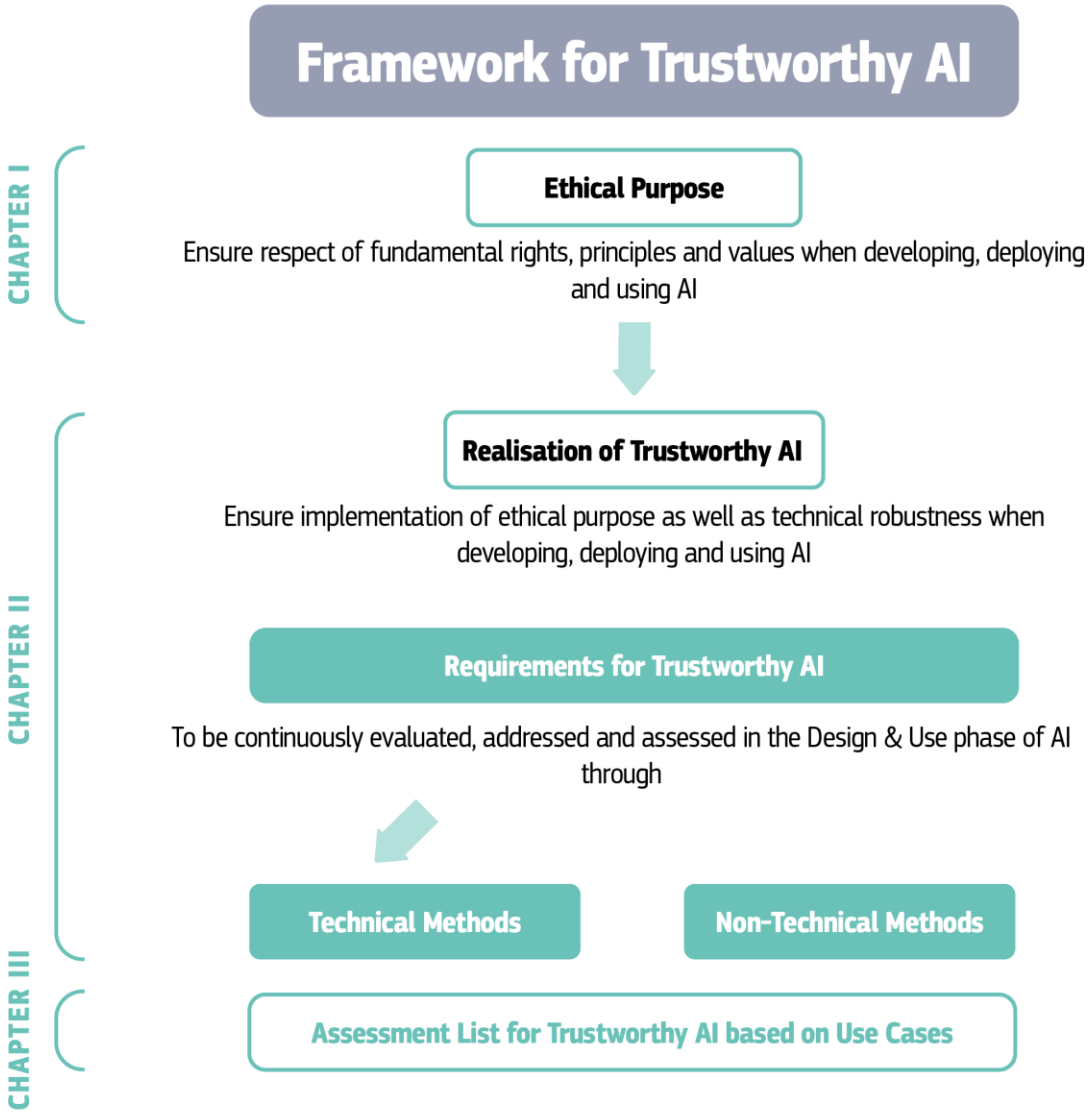

Figure 1: The Guidelines as a framework for Trustworthy AI

I. Respecting Fundamental Rights, Principles and Values - Ethical Purpose

1. The EU's Rights' Based Approach to AI Ethics

The High-Level Expert Group on AI ("AI HLEG") believes in an approach to AI ethics that uses the fundamental *rights* commitment of the EU Treaties and Charter of Fundamental Rights as the stepping stone to identify abstract ethical *principles,* and to specify how concrete ethical *values* can be operationalised in the context of AI. The EU is based on a constitutional commitment to protect the fundamental and indivisible rights of human beings[1], ensure respect for rule of law, foster democratic freedom and promote the common good. Other legal instruments further specify this commitment, like the European Social Charter or specific legislative acts like the General Data Protection Regulation (GDPR). Fundamental rights cannot only inspire new and specific regulatory instruments, they can also guide the rationale for AI systems' development, use and implementation – hence being dynamic.

The EU Treaties and the Charter prescribe the rights that apply when implementing EU law; which fall under the following chapters in the Charter: dignity, freedoms, equality and solidarity, citizens' rights and justice. The common thread to all of them is that in the EU a **human-centric approach** is upheld, whereby the human being enjoys a unique status of primacy in the civil, political, economic and social fields.

The field of ethics is also aimed at protecting individual rights and freedoms, while maximizing wellbeing and the common good. Ethical insights help us in understanding how technologies may give rise to different fundamental rights considerations in the development and application of AI, as well as finer grained guidance on what we *should* do with technology for the common good rather than what we (currently) *can* do with technology. A commitment to fundamental rights in the context of AI therefore requires an account of the ethical principles to be protected. In that vein, ethics is the foundation for, as well as a complement to, fundamental rights endorsed by humans.

The AI HLEG considers that a rights-based approach to AI ethics brings the additional benefit of limiting regulatory uncertainty. Building on the basis of decades of consensual application of fundamental rights in the EU provides clarity, readability and prospectivity for users, investors and innovators.

2. From Fundamental rights to Principles and Values

To give an example of the relationship between fundamental rights, principles, and values let us consider the fundamental **right** conceptualised as 'respect for human dignity'. This right involves recognition of the inherent value of humans (i.e. a human being does not need to look a certain way, have a certain job, or live in a certain country to be valuable, we are all valuable by virtue of being human). This leads to the ethical **principle** of autonomy which prescribes that individuals are free to make choices about their own lives, be it about their physical, emotional or mental wellbeing (i.e. since humans are valuable, they should be free to make choices about their own lives). In turn, informed consent is a **value** needed to operationalise the principle of autonomy in practice. Informed consent requires that individuals are given enough information to make an educated decision as to whether or not they will develop, use, or invest in an AI system at experimental or commercial stages (i.e. by ensuring that people are given the opportunity to consent to products or services, they can make choices about their lives and thus their value as humans is protected).

[1] These rights are for instance reflected in Articles 2 and 3 of the Treaty on European Union, and in the Charter of Fundamental Rights of the EU.

While this relationship appears to be linear, in reality values may often precede fundamental rights and/or principles.[2]

In short, **fundamental rights provide the bedrock for the formulation of ethical principles**. Those principles are abstract high-level norms that developers, deployers, users and regulators should follow in order to uphold the purpose of human-centric and Trustworthy AI. **Values, in turn, provide more concrete guidance on how to uphold ethical principles, while also underpinning fundamental rights.** The relationship between all three is illustrated in the following diagram (see Figure 2).

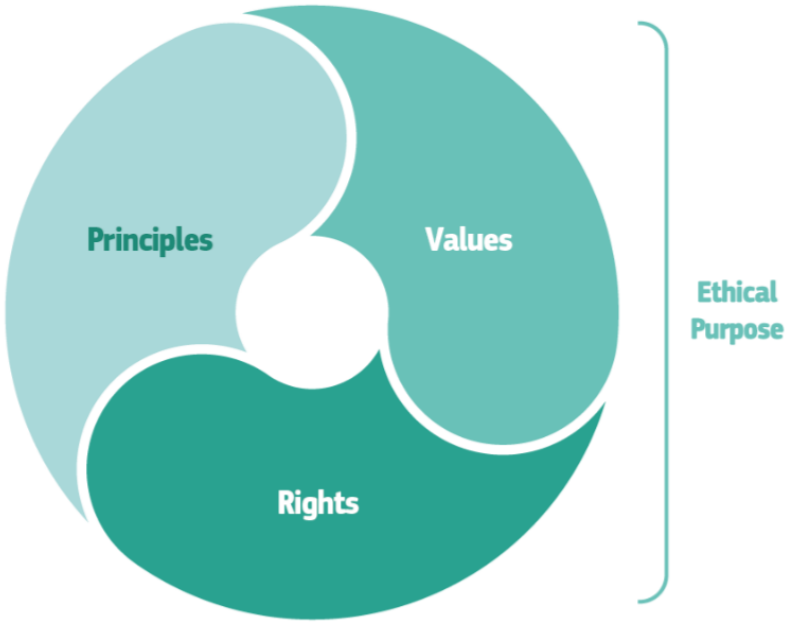

Figure 2: Relationship between Rights, Principles and Values – respect for which constitute Ethical Purpose

The AI HLEG is not the first to use fundamental rights to derive ethical principles and values. In 1997, the members of the Council of Europe adopted an instrument called the "*Convention for the Protection of Human Rights and Dignity of the Human Being with regard to the Application of Biology and Medicine*" (the "Oviedo Convention").[3] The Oviedo convention made it unambiguously clear that fundamental rights are the basic foundation to ensure the "primacy of the human being" in a context of technological change.

Respect for fundamental rights, principles and values – and ensuring that AI systems comply therewith – is coined here as ensuring "**ethical purpose**", and constitutes a key element to achieve Trustworthy AI.

[2] Additionally, values can be things we find good in themselves (i.e. intrinsic values) or good as a way of achieving another value (i.e. instrumental values). Our the use of values here (following the principles) is a specification of how these values can be impacted by AI rather than implying that these values are the result of, or derived from, the principles.

[3] This can be found at: https://rm.coe.int/168007cf98.

3. Fundamental Rights of Human Beings

Amongst the comprehensive set of indivisible rights set out in international human rights law, the EU Treaties and the Charter, the following families of rights are particularly apt to cover the AI field:

3.1 Respect for human dignity. Human dignity encompasses the idea that every human being possesses an "intrinsic worth", which can never be diminished, compromised or repressed by others – nor by new technologies like AI systems.[4] In the context of AI, respect for human dignity entails that all people are treated with respect due to them as individuals, rather than merely as data subjects. To specify the development or application of AI in line with human dignity, one can further articulate that AI systems are developed in a manner which serves and protects humans' physical and moral integrity, personal and cultural sense of identity as well as the satisfaction of their essential needs.

3.2 Freedom of the individual. This right refers to the idea that human beings should remain free to make life decisions for themselves. It does not only entail freedom from sovereign intrusion, but also requires intervention from government and non-governmental organizations to ensure that individuals or minorities benefit from equal opportunities. In an AI context, freedom of the individual requires protection from direct or indirect coercion, surveillance, deception or manipulation. In fact, freedom of the individual means a commitment to enable individuals to wield even higher control over their lives, including by protecting the freedom to conduct a business, the freedom of the arts and science, and the freedom of assembly and association**.**

3.3 Respect for democracy, justice and the rule of law. This entails that political power is human centric and bounded. AI systems must not interfere with democratic processes or undermine the plurality of values and life choices central to a democratic society. AI systems must also embed a commitment to abide by mandatory laws and regulation, and provide for due process by design, meaning a right to a human-centric appeal, review and/or scrutiny of decisions made by AI systems.

3.4 Equality, non-discrimination and solidarity including the rights of persons belonging to minorities. Equality means equal treatment of all human beings, regardless of whether they are in a similar situation. Equality of human beings 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 same rules should apply for everyone to access to information, data, knowledge, markets and a fair distribution of the value added being generated by technologies. Equality also requires adequate respect of inclusion of minorities, traditionally excluded, especially workers and consumers.

3.5. Citizens rights. In their interaction with the public sector, citizens benefit from a wide array of rights, including the right to a good administration, access to public documents, and the right to petition the administration. AI systems hold potential to improve the scale and efficiency of government in the provision of public goods and services to society. At the same time, citizens should enjoy a right to be informed of any automated treatment of their data by government bodies, and systematically be offered to express opt out. Citizens should never be subject to systematic scoring by government. Citizens should enjoy a right to vote and to be elected in democratic assemblies and institutions. To safeguard citizens' vote, governments shall take every possible measure to ensure full security of democratic processes.

[4] C. McCrudden, Human Dignity and Judicial Interpretation of Human Rights. European Journal of International Law, 19(4), 2008.

4. Ethical Principles in the Context of AI and Correlating Values

Many public, private, and civil organizations have drawn inspiration from fundamental rights to produce ethical frameworks for AI. 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 in the EU Charter of Fundamental Rights. More recently, the AI4People's project[5] has surveyed the aforementioned EGE principles as well as 36 other ethical principles put forward to date[6] and subsumed them under four overarching principles. These include: beneficence (defined as 'do good'), non-maleficence (defined as 'do no harm'), autonomy (defined as 'respect for self-determination and choice of individuals'), and justice (defined as 'fair and equitable treatment for all')[7]. These four principles have been updated by that same group to fit the AI context with the inclusion of a fifth principle: the principle of explicability. The AI HLEG believes in the benefits of convergence, as it allows for a recognition of most of the principles put forward by the variety of groups to date while at the same time clarifying the ends which all of the principles are aiming towards. Most importantly, these overarching principles provide guidance towards the operationalisation of core values[8].

Building on the above work, this section lists **five principles and correlated values that must be observed** to ensure that AI is developed in a human-centric manner. These have been proposed and justified by the abovementioned project[9].

It should also be noted that, in particular situations, tensions may arise between the principles when considered from the point of view of an individual compared with the point of view of society, and vice versa. There is no set way to deal with such trade-offs. In such contexts, it may however help to return to the principles and overarching values and rights protected by the EU Treaties and Charter. Given the potential of unknown and unintended consequences of AI, the presence of an internal and external (ethical) expert is advised to accompany the design, development and deployment of AI. Such expert could also raise further awareness of the unique ethical issues that may arise in the coming years.

We introduce and illustrate the principles and values in the context of AI below.

- <u>The Principle of Beneficence: "Do Good"</u>

AI systems should be designed and developed to improve individual and collective wellbeing. AI systems can do so by generating prosperity, value creation and wealth maximization and sustainability. At the same

---

5 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.

6 The principles analysed were: the Asilomar AI Principles, developed under the auspices of the Future of Life Institute (2017); the Montreal Declaration for Responsible AI, developed under the auspices of the University of Montreal (2017), the General Principles of the IEEE's second version of Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (2017), the Ethical Principles put forward by the European Commission's European Group on Ethics in Science and New Technologies (2018); the "five overarching principles for an AI code" of §417 of the UK House of Lords Artificial Intelligence Committee's report (2018); and the Tenets of the Partnership on AI (2018).

7 These principles were originally proposed in a medical context by T Beauchamp and J Childress, for more on this please refer to Beauchamp TL, Childress JF. Principles of biomedical ethics. 5th. New York: Oxford University Press; 2001.

8 We draw on the framework proposed by Ibo van de Poel for translating values into design requirements. This comprises two main phases; value specification and value operationalisation. For more on this see Van de Poel, I. (2013). Translating values into design requirements. In Philosophy and engineering: Reflections on practice, principles and process (pp. 253-266). Springer, Dordrecht.

9 L. Floridi et al. (2018), "AI4People —An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations", *Minds and Machines* 28(4): 689-707.

time, beneficent AI systems can contribute to wellbeing by seeking achievement of a fair, inclusive and peaceful society, by helping to increase citizen's mental autonomy, with equal distribution of economic, social and political opportunity. AI systems can be a force for collective good when deployed towards objectives like: the protection of democratic process and rule of law; the provision of common goods and services at low cost and high quality; data literacy and representativeness; damage mitigation and trust optimization towards users; achievement of the UN Sustainable Development Goals or *sustainability* understood more broadly, according to the pillars of economic development, social equity, and environmental protection[10]. In other words, AI can be a tool to bring more good into the world and/or to help with the world's greatest challenges.

- The Principle of Non maleficence: "Do no Harm"

AI systems should not harm human beings. By design, AI systems should protect the dignity, integrity, liberty, privacy, safety, and security of human beings in society and at work. AI systems should not threaten the democratic process, freedom of expression, freedoms of identify, or the possibility to refuse AI services. At the very least, AI systems should not be designed in a way that enhances existing harms or creates new harms for individuals. Harms can be physical, psychological, financial or social. AI specific harms may stem from the treatment of data on individuals (i.e. how it is collected, stored, used, etc.). To avoid harm, data collected and used for training of AI algorithms must be done in a way that avoids discrimination, manipulation, or negative profiling. Of equal importance, AI systems should be developed and implemented in a way that protects societies from ideological polarization and algorithmic determinism.

Vulnerable demographics (e.g. children, minorities, disabled persons, elderly persons, or immigrants) should receive greater attention to the prevention of harm, given their unique status in society. Inclusion and diversity are key ingredients for the prevention of harm to ensure suitability of these systems across cultures, genders, ages, life choices, etc. Therefore not only should AI be designed with the impact on various vulnerable demographics in mind but the above mentioned demographics should have a place in the design process (rather through testing, validating, or other).

Avoiding harm may also be viewed in terms of harm to the environment and animals, thus the development of *environmentally friendly*[11] AI may be considered part of the principle of avoiding harm. The Earth's resources can be valued in and of themselves or as a resource for humans to consume. In either case it is necessary to ensure that the research, development, and use of AI are done with an eye towards environmental awareness.[12]

- The Principle of Autonomy: "Preserve Human Agency"

Autonomy of human beings in the context of AI development means freedom from subordination to, or coercion by, AI systems. Human beings interacting with AI systems must keep full and effective self-

[10] For more information on the three pillars see Drexhage, J., & Murphy, D. (2010). Sustainable development: from Brundtland to Rio 2012. Background paper prepared for consideration by the High Level Panel on Global Sustainability at its first meeting 19 September 2010.

[11] The concept of "environmental friendliness" as stronger than that of sustainability is introduced in L. Floridi, et al. (2018), "AI4People —An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations", *Minds and Machines* 28(4): 689-707.

[12] Items to consider here are the impact of the large amounts of computing power to run AI systems, the data warehouses needed for storage of data, and the procurement of minerals to fuel the batteries needed for all devices involved in an AI system. For the latter, these minerals most often come from a mine without certification in an under-developed country and contribute to the inhumane treatment of individuals.

determination over themselves. If one is a consumer or user of an AI system this entails a right to decide to be subject to direct or indirect AI decision making, a right to knowledge of direct or indirect interaction with AI systems, a right to opt out and a right of withdrawal.[13]

Self-determination in many instances requires assistance from government or non-governmental organizations to ensure that individuals or minorities are afforded similar opportunities as the status quo. Furthermore, to ensure human agency, systems should be in place to ensure *responsibility* and *accountability*. It is paramount that AI does not undermine the necessity for human responsibility to ensure the protection of fundamental rights.

- The Principle of Justice: "Be Fair"

For the purposes of these Guidelines, the principle of justice imparts that the development, use, and regulation of AI systems must be fair. Developers and implementers need to ensure that individuals and minority groups maintain freedom from bias, stigmatisation and discrimination. Additionally, the positives and negatives resulting from AI should be evenly distributed, avoiding to place vulnerable demographics in a position of greater vulnerability and striving for equal opportunity in terms of access to education, goods, services and technology amongst human beings, without discrimination. Justice also means that AI systems must provide users with effective redress if harm occurs, or effective remedy if data practices are no longer aligned with human beings' individual or collective preferences. Lastly, the principle of justice also commands those developing or implementing AI to be held to high standards of accountability. Humans might benefit from procedures enabling the benchmarking of AI performance with (ethical) expectations.

- The Principle of Explicability: "Operate transparently"

Transparency is key to building and maintaining citizen's trust in the developers of AI systems and AI systems themselves. Both technological and business model transparency matter from an ethical standpoint. Technological transparency implies that AI systems be auditable,[14] comprehensible and intelligible by human beings at varying levels of comprehension and expertise. Business model transparency means that human beings are knowingly informed of the intention of developers and technology implementers of AI systems.

Explicability[15] is a precondition for achieving informed consent from individuals interacting with AI systems and in order to ensure that the principle of explicability and non-maleficence are achieved the requirement of informed consent should be sought. Explicability also requires accountability measures be put in place. Individuals and groups may request evidence of the baseline parameters and instructions given as inputs for AI decision making (the discovery or prediction sought by an AI system or the factors involved in the discovery or prediction made) by the organisations and developers of an AI system, the technology implementers, or another party in the supply chain.

---

13 This includes a right to individually and collectively decide on how AI systems operate in a working environment. This may also include provisions designed to ensure that anyone using AI as part of his/her employment enjoys protection for maintaining their own decision making capabilities and is not constrained by the use of an AI system.

14 We refer to both an IT audit of the algorithm as well as a procedural audit of the data supply chain.

15 The literature normally speaks of "explainability". The concept of "explicability" to refer both to "intelligibility" and to "explainability" and hence capture the need for transparency and for accountability is introduced in L. Floridi, et al. (2018), "AI4People —An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations", Minds and Machines 28(4): 689-707.

5. Critical concerns raised by AI

This section has sparked lively discussions between the AI HLEG members, and we did not reach agreement on the extent to which the areas as formulated here below raise concerns. We are therefore asking specific input on this point from those partaking in the stakeholder consultation.

Particular uses or applications, sectors or contexts of AI may raise specific concerns, as they run counter the rights and principles set out above. While AI can foster and enable our European values, like many other powerful technologies, its dual-use nature implies that AI can also be used to infringe these. A balance must thus be considered between what *should* and what *can* be done with AI, and due care should be given to what *should not* be done with AI. Of course, our understanding of rules and principles evolves over time and may change in the future. The following non-exhaustive list of critical concerns might therefore be shortened, edited, or updated in the future.

5.1 Identification without Consent

AI enables an ever more efficient identification of individual persons by either public or private entities. A proportionate use of control techniques in AI is needed to uphold the autonomy of European citizens. Differentiating between the identification of an individual vs. the tracing and tracking of an individual, and between targeted surveillance and mass surveillance, will be crucial for the achievement of Trustworthy AI. In this regard, Article 6 of the General Data Protection Regulation (GDPR) can be recalled, which provides that processing of data shall only be lawful if it has a valid legal basis.

As current mechanisms for giving informed consent in the internet show, consumers give consent without consideration. This involves an ethical obligation to develop entirely new and practical means by which citizens can give verified consent to being automatically identified by AI or equivalent technologies. Noteworthy examples of a scalable AI identification technology are face recognition or other involuntary methods of identification using biometric data (i.e. lie detection, personality assessment through micro expressions, automatic voice detection). Identification of individuals is sometimes the desirable outcome and aligned with ethical principles (for example in detecting fraud, money laundering, or terrorist financing, etc.). Where the application of such technologies is not clearly warranted by existing law or the protection of core values, automatic identification raises strong concerns of both legal and ethical nature, with the default assumption being that consent to identification has not been given. This also applies to the usage of "anonymous" personal data that can be re-personalized.

5.2 Covert AI systems

A human always has to know if she/he is interacting with a human being or a machine, and it is the responsibility of AI developers and deployers that this is reliably achieved. Otherwise, people with the power to control AI are potentially able to manipulate humans on an unprecedented scale. **AI developers and deployers should therefore ensure that humans are made aware of – or able to request and validate the fact that – they interact with an AI identity.** Note that border-cases exist and complicate the matter – e.g. an AI-filtered voice spoken by a human. Androids can be considered covert AI systems, as they are robots that are built to be as human-like as possible. Their inclusion in human society might change our perception of humans and humanity. It should be born in mind that the confusion between humans and machines has

multiple consequences such as attachment, influence, or reduction of the value of being human.[16] The development of humanoid and android robots should therefore undergo careful ethical assessment.

5.3 Normative & Mass Citizen Scoring without consent in deviation of Fundamental Rights

We value the freedom and autonomy of all citizens. Normative citizen scoring (e.g., general assessment of "moral personality" or "ethical integrity") in *all* aspects and on a large scale by public authorities endangers these values, especially when used not in accordance with fundamental rights, or when used disproportionately and without a delineated and communicated legitimate purpose. Today, citizen scoring – at large or smaller scale – is already often used in purely descriptive and domain-specific scorings (e.g. school systems, e-learning, or driver licenses). However, whenever citizen scoring is applied in a limited social domain, a fully transparent procedure should be available to citizens, providing them with information on the process, purpose and methodology of the scoring, and ideally providing them with the possibility to opt-out of the scoring mechanism. This is particularly important in situations where an asymmetry of power exists between the parties. Developers and deployers should therefore ensure such opt-out option in the technology's design, and make the necessary resources available for this purpose.

5.4 Lethal Autonomous Weapon Systems (LAWS)

LAWS can operate without meaningful human control over the critical functions of selecting and attacking individual targets. Ultimately, human beings are, and must remain, responsible and accountable for all casualties. 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 can lead to an uncontrollable arms race on a historically unprecedented level, and can create military contexts in which human control is almost entirely relinquished and risks of malfunction not addressed. Note that, on the other hand, in an armed conflict LAWS can reduce collateral damage, e.g. saving selectively children. 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.[17] Recalling the European Union's aim to promote peace as enshrined in Article 3 of the TEU, the AI HLEG stands with, and looks to support, the EU Parliament's resolution of 12 September 2018 and all related efforts on LAWS.

5.5 Potential longer-term concerns

This sub-section has proven to be highly controversial in discussions between the AI HLEG members, and we did not reach agreement on the extent to which the areas formulated below raise concerns. We therefore ask specific input on this point from those partaking in the stakeholder consultation.

All current AI 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, critical long-term concerns can be identified – which are by their very nature speculative. The probability of occurrence of such

[16] 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).

[17] European Parliament's Resolution 2018/2752(RSP).

scenarios may from today's perspective be very low, yet the potential harm associated with it could in some instances be very high (examples thereof are the development of *Artificial Consciousness, i.e.* AI systems that may have a subjective experience,[18] of Artificial Moral Agents[19] or of Unsupervised Recursively Self-Improving Artificial General Intelligence (AGI)[20] – which today still seem to belong to the very distant future). A risk-assessment approach therefore invites us to keep such areas into consideration and invest resources into minimizing epistemic indeterminacy about long-term risks, unknown unknowns and "black swans"[21]. We invite those partaking in the consultation to share their views thereon.

KEY GUIDANCE FOR ENSURING ETHICAL PURPOSE:

- Ensure that AI is **human-centric**: AI should be developed, deployed and used with an "**ethical purpose"** as set out above, grounded in and reflective of fundamental rights, societal values and the ethical principles of *Beneficence* (do good), *Non-Maleficence* (do no harm), *Autonomy of humans, Justice,* and *Explicability.* This is crucial to work towards **Trustworthy AI**.
- Rely on fundamental rights, ethical principles and values to prospectively evaluate possible effects of AI on human beings and the common good. Pay **particular attention** to situations involving more **vulnerable groups** such as children, persons with disabilities or minorities, or to situations with **asymmetries of power or information**, such as between employers and employees, or businesses and consumers.
- Acknowledge and be aware of the fact that – while bringing substantive benefits to individuals and society – AI can also have a negative impact. Remain **vigilant for areas of critical concern**.

[18] We currently lack a widely accepted theory of consciousness. However, should the development of artificial consciousness be possible, this would be highly problematic from an ethical, legal, and political perspective. It could create potentially large amounts of suffering on self-conscious non-biological carrier systems. Moreover, it would carry the risk that certain future types of self-conscious AI systems would need to be treated as ethical objects, having specific rights. It is in this regard noted that consciousness research labs already exist today in France, the USA and Japan, which have the proclaimed target to build artificial consciousness.

[19] A "moral agent" is a system that a) autonomously arrives at normative judgments and conclusions, and b) autonomously *acts* on the basis of such self-generated judgments and conclusions. Current systems are not able to do this. The development thereof, however, would potentially present a conflict with maintaining responsibility and accountability in the hands of humans, and would potentially threaten the values of autonomy and self-determination.

[20] As mentioned, current AI is domain specific and not general, yet the potential occurrence of the ability to develop unsupervised recursively self-improving AGI (an artificial general intelligence that can develop a subsequent, potentially more powerful, generation of artificial general intelligence) might lose alignment with human values, even if its designers carefully implemented them, as goal-permanence and value alignment would not be assured under such a complex self-improving process. This does not yet apply to current AI systems or systems that incrementally gather sensory experiences and thereby improve their internal models and possibly the structure of such models. Nevertheless, research in this domain should hence not only adhere to safety conditions, but also to the ethics of risk mentioned above.

[21] 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 is not computable using scientific methods.

III. Realising Trustworthy AI

This Chapter offers **guidance on the implementation and realisation of Trustworthy AI.** We set out what the **main requirements are for AI to be Trustworthy**, and the **methods** available in order to implement those requirements when developing, deploying and using AI, so as to enable full benefit from the opportunities created thereby.

1. Requirements of Trustworthy AI

Achieving Trustworthy AI means that the general and abstract principles need to be mapped into concrete requirements for AI systems and applications. The ten requirements listed below have been derived from the rights, principles and values of Chapter I. While they are all equally important, in different application domains and industries, the specific context needs to be taken into account for further handling thereof.

1. Accountability
2. Data Governance
3. Design for all
4. Governance of AI Autonomy (Human oversight)
5. Non-Discrimination
6. Respect for (& Enhancement of) Human Autonomy
7. Respect for Privacy
8. Robustness
9. Safety
10. Transparency

This list is non-exhaustive and introduces the requirements for Trustworthy AI in alphabetical order, to stress the equal importance of all requirements. In Chapter III, we provide an Assessment List to support the operationalisation on these requirements.

1. Accountability

Good AI governance should include accountability mechanisms, which could be very diverse in choice depending on the goals. Mechanisms can range from monetary compensation (no-fault insurance) to fault finding, to reconciliation without monetary compensations. The choice of accountability mechanisms may also depend on the nature and weight of the activity, as well as the level of autonomy at play. An instance in which a system misreads a medicine claim and wrongly decides not to reimburse may be compensated for with money. In a case of discrimination, however, an explanation and apology might be at least as important.

2. Data Governance

The quality of the data sets used is paramount for the performance of the trained machine learning solutions. Even if the data is handled in a privacy preserving way, there are requirements that have to be fulfilled in order to have high quality AI. The datasets gathered inevitably contain biases, and one has to be able to prune these away before engaging in training. This may also be done in the training itself by requiring a symmetric behaviour over known issues in the training set.

In addition, it must be ensured that the proper division of the data which is being set into training, as well as validation and testing of those sets, is carefully conducted in order to achieve a realistic picture of the performance of the AI system. It must particularly be ensured that anonymisation of the data is done in a way that enables the division of the data into sets to make sure that a certain data – for instance, images from same persons – do not end up into both the training and test sets, as this would disqualify the latter.

The integrity of the data gathering has to be ensured. Feeding malicious data into the system may change the behaviour of the AI solutions. This is especially important for self-learning systems. It is therefore advisable to always keep record of the data that is fed to the AI systems. When data is gathered from human behaviour, it may contain misjudgement, errors and mistakes. In large enough data sets these will be diluted since correct actions usually overrun the errors, yet a trace of thereof remains in the data.

To trust the data gathering process, it must be ensured that such data will not be used against the individuals who provided the data. Instead, the findings of bias should be used to look forward and lead to better processes and instructions – improving our decisions making and strengthening our institutions.

3. Design for all

Systems should be designed in a way that allows all citizens to use the products or services, regardless of their age, disability status or social status. It is particularly important to consider accessibility to AI products and services to people with disabilities, which are horizontal category of society, present in all societal groups independent from gender, age or nationality. AI applications should hence not have a one-size-fits-all approach, but be user-centric and consider the whole range of human abilities, skills and requirements. Design for all implies the accessibility and usability of technologies by anyone at any place and at any time, ensuring their inclusion in any living context[22], thus enabling equitable access and active participation of potentially all people in existing and emerging computer-mediated human activities. This requirement links to the United Nations Convention on the Rights of Persons with Disabilities.[23]

4. Governance of AI Autonomy (Human oversight)

The correct approach to assuring properties such as safety, accuracy, adaptability, privacy, explicability, compliance with the rule of law and ethical conformity heavily depends on specific details of the AI system, its area of application, its level of impact on individuals, communities or society and its level of autonomy. The level of autonomy[24] results from the use case and the degree of sophistication needed for a task. All other things being equal, the greater degree of autonomy that is given to an AI system, the more extensive

[22] ftp://ftp.cencenelec.eu/EN/EuropeanStandardization/HotTopics/Accessibility/ETSIGuide.pdf

[23] https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities.html

[24] AI systems often operate with some degree of autonomy, typically classified into 5 levels: (1) Domain model is implicitly implemented and part of the programme code. No intelligence implemented, interaction is based on stimulus-response basis. Responsibility for behaviour lies with the developer. (2) Machine can learn and adapt but works on implemented/ given domain model; responsibility has to be with the developer since basic assumptions are hard coded. (3) Machine correlates internal domain model with sensory perception & information. Behaviour is data driven with regard to a mission. Ethical behaviour can be modelled according to decision logic with a utility function. (4) Machine operates on a world model as perceived by sensors. Some degree of self-awareness could be created for stability and resilience; might be extended to act based on a deontic ethical model. (5) Machine operates on a world model and has to understand rules & conventions in a given world fragment. Capability of full moral judgement requires higher order reasoning, however, second order or modal logics are undecidable. Thus, some form of legal framework and and international conventions seem necessary and desirable. Systems that operate at level 4 can be said to have "Operational autonomy". I.e., given a (set of) goals, the system can set its actions or plans.

testing and stricter governance is required. It must be ensured that AI systems continue to behave as intended when feedback signals become sparser.

Depending on the area of application and/or the level of impact on individuals, communities or society of the AI-system, different levels or instances of governance (incl. human oversight) will be necessary. This is relevant for a large number of AI applications, and more particularly for the use of AI to suggest or take decisions concerning individuals or communities (algorithmic decision support). Good governance of AI autonomy in this respect includes for instance more or earlier human intervention depending on the level of societal impact of the AI-system. This also includes the predicament that a user of an AI system, particularly in a work or decision-making environment, is allowed to deviate from a path or decision chosen or recommended by the AI system.

5. Non-Discrimination

Discrimination concerns the variability of AI results between individuals or groups of people based on the exploitation of differences in their characteristics that can be considered either intentionally or unintentionally (such as ethnicity, gender, sexual orientation or age), which may negatively impact such individuals or groups.

Direct or indirect discrimination[25] through the use of AI can serve to exploit prejudice and marginalise certain groups. Those in control of algorithms may intentionally try to achieve unfair, discriminatory, or biased outcomes in order to exclude certain groups of persons. Intentional harm can, for instance, be achieved by explicit manipulation of the data to exclude certain groups. Harm may also result from exploitation of consumer biases or unfair competition, such as homogenisation of prices by means of collusion or non-transparent market[26].

Discrimination in an AI context can occur unintentionally due to, for example, problems with data such as bias, incompleteness and bad governance models. Machine learning algorithms identify patterns or regularities in data, and will therefore also follow the patterns resulting from biased and/or incomplete data sets. An incomplete data set may not reflect the target group it is intended to represent. While it might be possible to remove clearly identifiable and unwanted bias when collecting data, data always carries some kind of bias. Therefore, the upstream identification of possible bias, which later can be rectified, is important to build in to the development of AI.

Moreover, it is important to acknowledge that AI technology can be employed to identify this inherent bias, and hence to support awareness training on our own inherent bias. Accordingly, it can also assist us in making less biased decisions.

6. Respect for (& Enhancement of) Human Autonomy

AI systems should be designed not only to uphold rights, values and principles, but also to protect citizens in all their diversity from governmental and private abuses made possible by AI technology, ensuring a fair

[25] For a definition of direct and indirect discrimination, see for instance Article 2 of Council Directive 2000/78/EC of 27 November 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.

[26] Cf. Paper by the European Union Agency for Fundamental Rights: "BigData: Discrimination in data-supported decision making (2018)" http://fra.europa.eu/en/publication/2018/big-data-discrimination.

distribution of the benefits created by AI technologies, protect and enhance a plurality of human values, and enhance self-determination and autonomy of individual users and communities.

AI products and services, possibly through "extreme" personalisation approaches, may steer individual choice by potentially manipulative "nudging". At the same time, people are increasingly willing and expected to delegate decisions and actions to machines (e.g. recommender systems, search engines, navigation systems, virtual coaches and personal assistants). Systems that are tasked to help the user, must provide explicit support to the user to promote her/his own preferences, and set the limits for system intervention, ensuring that the overall wellbeing of the user as explicitly defined by the user her/himself is central to system functionality.

7. Respect for Privacy

Privacy and data protection must be guaranteed at all stages of the life cycle of the AI system. This includes all data provided by the user, but also all information generated about the user over the course of his or her interactions with the AI system (e.g. outputs that the AI system generated for specific users, how users responded to particular recommendations, etc.). Digital records of human behaviour can reveal highly sensitive data, not only in terms of preferences, but also regarding sexual orientation, age, gender, religious and political views. The person in control of such information could use this to his/her advantage. Organisations must be mindful of how data is used and might impact users, and ensure full compliance with the GDPR as well as other applicable regulation dealing with privacy and data protection.

8. Robustness

Trustworthy AI requires that algorithms are secure, reliable as well as robust enough to deal with errors or inconsistencies during the design, development, execution, deployment and use phase of the AI system, and to adequately cope with erroneous outcomes.

Reliability & Reproducibility. Trustworthiness requires that the accuracy of results can be confirmed and reproduced by independent evaluation. However, the complexity, non-determinism and opacity of many AI systems, together with sensitivity to training/model building conditions, can make it difficult to reproduce results. Currently there is an increased awareness within the AI research community that reproducibility is a critical requirement in the field. Reproducibility is essential to guarantee that results are consistent across different situations, computational frameworks and input data. The lack of reproducibility can lead to unintended discrimination in AI decisions.

Accuracy. Accuracy pertains to an AI's confidence and ability to correctly classify information into the correct 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.

Resilience to Attack. AI systems, like all software systems, can include vulnerabilities that can allow them to be exploited by adversaries. Hacking is an important case of intentional harm, by which the system will purposefully follow a different course of action than its original purpose. If an AI system is attacked, the data as well as system behaviour can be changed, leading the system to make different decisions, or causing the system to shut down altogether. Systems and/or data can also become corrupted, by malicious intention or by exposure to unexpected situations. Poor governance, by which it becomes possible to intentionally or

unintentionally tamper with the data, or grant access to the algorithms to unauthorised entities, can also result in discrimination, erroneous decisions, or even physical harm.

Fall back plan. A secure AI has safeguards that enable a fall-back plan in case of problems with the AI system. In some cases this can mean that the AI system switches from statistical to rule-based procedure, in other cases it means that the system asks for a human operator before continuing the action.

9. Safety

Safety is about ensuring that the system will indeed do what it is supposed to do, without harming users (human physical integrity), resources or the environment. It includes minimizing unintended consequences and errors in the operation of the system. Processes to clarify and assess potential risks associated with the use of AI products and services should be put in place. Moreover, formal mechanisms are needed to measure and guide the adaptability of AI systems.

10. Transparency

Transparency concerns the reduction of information asymmetry. Explainability – as a form of transparency – entails the capability to describe, inspect and reproduce the mechanisms through which AI systems make decisions and learn to adapt to their environments, as well as the provenance and dynamics of the data that is used and created by the system. Being explicit and open about choices and decisions concerning data sources, development processes, and stakeholders should be required from all models that use human data or affect human beings or can have other morally significant impact.

2. Technical and Non-Technical Methods to achieve Trustworthy AI

In order to address the requirements described in the previous section, both **technical** and **non-technical** methods can be employed, at all levels of the development processes - including analysis, design, development and use (cf. Figure 3). An evaluation of the requirements and the methods employed to implement these, as well as reporting and justifying changes to the processes, should occur on **an on-going basis**. In fact, given that AI systems are continuously evolving and acting in a dynamic environment, achieving Trustworthy AI is a **continuous process**.

While the list of methods below is not exhaustive, it aims to reflect the main approaches that are recommended to implement Trustworthy AI. To enhance the trustworthiness of an AI system, these methods should be grounded in the rights and principles defined in Chapter I.

Figure 3 depicts the impact of rights, principles and values on systems' development processes. These abstract principles and rights are concretized into requirements for the AI system, whose implementation and realisation is supported by different technical and non-technical methods. Moreover, given the adaptable and dynamic aspect of AI technology, continued adherence to principles and values requires that evaluation and justification[27] processes are central to the development process.

[27] This entails for instance justification of the choices made in the design, development and deployment of the system in order to incorporate the abovementioned requirements.

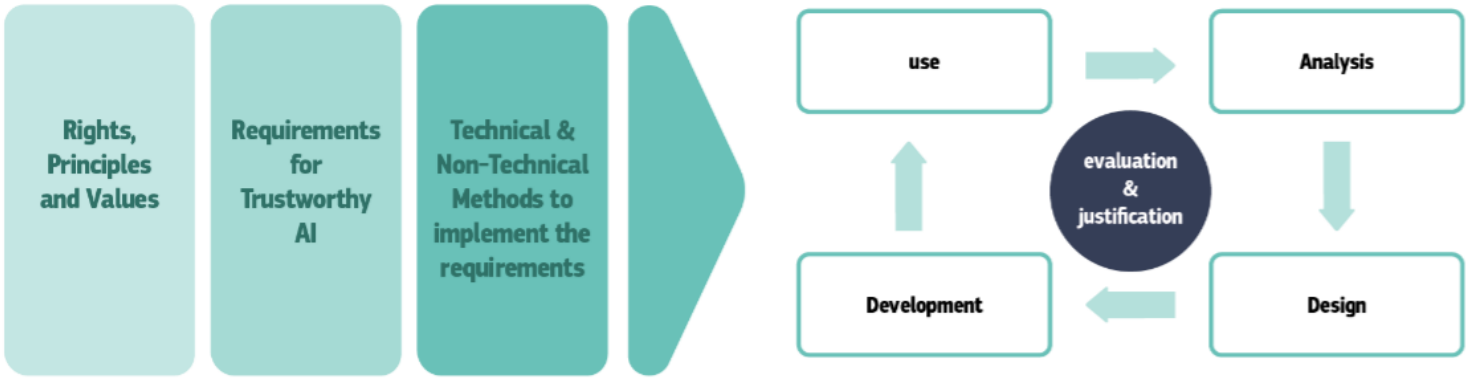

Figure 3: Realising Trustworthy AI throughout the entire life cycle of the system

1. Technical methods

This section describes technical methods to ensure trustworthy AI, which can be incorporated in the design, development and use phase of an AI system. Importantly, evaluating the requirements and implementing the methods should occur on **an on-going basis**. While the list of methods below is not exhaustive nor meant as mandatory, it aims to reflect the main technical approaches that can help to ensure the implementation of Trustworthy AI.

Some methods already exist today, others can still be much improved over time in light of research that is being undertaken in that area, while others do not yet exist today, necessitating further research. Those areas where further research is needed will also inform the second deliverable of the AI HLEG (for instance equity-by-design in supervised machine learning approaches, algorithmic repeatability, robustness to bias and corruption or development of causal models). Below, examples of existing solutions are presented.

- *Ethics & Rule of law by design (X-by-design)*

Methods to ensure values-by-design provide precise and explicit links between the abstract principles the system is required to adhere to and the specific implementation decisions, in ways that are accessible and justified by legal rules or societal norms. Central therein is the idea that compliance with law as well as with ethical values can be implemented, at least to a certain extent, into the design of the AI system itself.

This also entails a responsibility for companies to identify from the very beginning the ethical impact that an AI system can have, and the ethical and legal rules that the system should comply with. Different "by-design" concepts are already widely used, **two examples** of which are *Privacy-by-design or Security-by-design.* To earn trust, AI needs to be secure with its processes, data and outcomes and be able to take adversarial data and attacks into account. In addition, it should implement a mechanism for fail-safe shutdown and resume operation after a forced shut-down (e.g. after an attack).

- *Architectures for Trustworthy AI*

The requirements for Trustworthy AI need to be "translated" into procedures and/or constraints on procedures, which should be anchored in an intelligent system's architecture. This can either be accomplished by formulating rules, which control the behaviour of an intelligent agent, or as behaviour boundaries that must not be trespassed, and the monitoring of which is a separate process.

An intelligent system that will have the capabilities to learn and adapt its behaviour actively can be understood as a stochastic system and is often described by a "sense-plan-act" cycle. For such architecture to be adapted to ensure Trustworthy AI, ethical goals and requirements should be integrated at "sense"-level in a way that plans can be formulated that observe and ensure adherence to those principles. In this way, actions and decisions by the system reflect the observed principles.

The architecture as sketched above is generic and may be only partly implemented in most AI systems. Nevertheless, it gives anchor points for constraints and policies that have to be reflected in specific modules to result in an overall system that is perceived as trustworthy.

- *Testing & Validating*

Due to the non-deterministic nature of intelligent systems, traditional testing is not enough. Intelligence manifests itself on the semantic level, e.g. during program execution. Consequently, to verify and validate consistent and intended processing of data, the underlying model has to be carefully monitored regarding stability, robustness, and operation in 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 taken can be made plausible in a way allowing validation of the underlying process. Testing and validation of the system should thus occur as early as possible and be iterative, ensuring the system behaves as intended throughout its entire life cycle and especially after deployment.

Importantly, testing should not be limited to data, but include all inputs to the system (e.g. pre-trained models), and the behaviour of the system as a whole. Moreover, it should be performed by an as diverse a group of people as possible. Multiple metrics need to be developed to cover the categories that are being tested for different perspectives. The data used for testing should be carefully developed and updated regularly. To this end, adversarial test and bounty hunting can be considered, whenever feasible.

Finally, it has to be ensured that the commands to the "acting-module" are again consistent with the results of the preceding processes and have to be compared to the previously defined policies to ensure that they are not violated.

- *Traceability & Auditability*

To tackle the challenges of transparency and explainability, AI systems should document both the decisions they make and the whole process that yielded the decisions, to make decisions traceable. While traceability is not (always) able to tell us why a certain decision was reached, it can tell us how it came about – this enables reasoning as to why an AI-decision was erroneous and can help prevention of future mistakes. Traceability is thus a facilitator for auditability, which entails the enablement and facilitation of monitoring and verification of algorithms, data and design processes. To a certain extent, auditability of AI is reachable today, and will improve over time through further research.

Whenever an AI system has a significant impact on people's lives, laypersons should be able to understand the causality of the algorithmic decision-making process and how it is implemented by organisations that deploy the AI system. The development of human machine interfaces that provide mechanisms for understanding the system's behaviour can assist in this regard. Evaluation by internal and external auditors can contribute to the laymen's acceptance of the technology. Importantly, in order to enable regulatory bodies to undertake verification and auditing of AI systems where needed, they would need to undergo a digital transformation and develop the necessary tools to this end.

- *Explanation (XAI research)*

A known issue with learning systems based on neural nets is the difficulty to provide clear reasons for the interpretations and decisions of the system. This is due to the fact that the training process has resulted in setting the network parameters to numerical values that are difficult to correlate with the results. In addition, sometimes small changes in some values of the data might result in dramatic changes in the interpretation, leading the system to confuse a school bus with an ostrich for example. This specific issue might be used to deceive the system.

For a system to be trustworthy, it is necessary to be able to understand why it had a given behaviour and why it has provided a given interpretation. It is also necessary to limit these adversarial situations. As of today, this is still an open challenge to AI systems based on neural networks. A whole field of research, Explainable AI (XAI) is trying to address this issue, to better understand the underlying mechanisms and find solutions. The matter is of prime importance not only to explain AI's behaviour to the developer or the user, but also to simply deploy reliable AI systems.

2. Non-Technical Methods

This section describes non-technical methods to ensure trustworthy AI, which should likewise be evaluated on **an on-going basis**. The list of methods below is not exhaustive nor meant as mandatory, it aims to help to ensuring the implementation of Trustworthy AI.

- *Regulation*

Many regulations already exist today that increase AI's Trustworthiness, such as safety legislation or liability frameworks. To the extent the AI HLEG considers that regulation may need to be revised, adapted or introduced, this will be discussed in the second deliverable.

Trustworthy AI also requires responsibility mechanisms that, when harm does occur, ensure an appropriate remedy can be put in place. Knowing that redress is possible when things go wrong increases trust. Mechanisms can range from monetary compensation in circumstances where AI systems caused harm, to negligence or culpability-based mechanisms for liability, to reconciliation, rectification and apology without the need for monetary compensation. In this regard, applicable law comes into play – encompassing access to justice – in compliance with fundamental rights.[28]

- *Standardization*

Using agreed standards for design, manufacturing and business practices can function as a quality management system for AI offering consumers, actors and governments the ability to recognise and reward 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. Examples are ISO Standards, the Fair Trade mark or Made in Europe label.

[28] See for instance the Fundamental Rights Agency's opinion reflecting on business and human rights, including the concpet of due diligence in this context: http://fra.europa.eu/en/opinion/2017/business-human-rights.

- *Accountability Governance*

Organisations should set up an internal or external governance framework to ensure accountability. This can, for instance, include the appointment of a person in charge of ethics issues as they relate to AI, an internal ethics panel or board, and/or an external ethics panel or board. Amongst the possible roles of such a person, panel or board, is to provide oversight on issues that may arise and provide advice throughout the process. This can be in addition to, but cannot replace, legal oversight; for example, in the form of a data protection officer or equivalent.

- *Codes of Conduct*

Organisations and stakeholders can sign up to the Guidelines, and adapt their charter of corporate responsibility, Key Performance Indicators ("KPIs"), or their codes of conduct to add the striving towards Trustworthy AI. An organisation working on an AI system 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.

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

Trustworthy AI requires informed participation of all stakeholders. This necessitates that education plays an important role, both to ensure that knowledge of the potential impact of AI is widespread, and to make people aware that they can participate in shaping the societal development. Education here refers to the people 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, society at large). A pre-requisite for educating the public is to ensure the proper skills and training of ethicists in this space.

- *Stakeholder and social dialogue*

From better healthcare to safer transport, the benefits of AI are many and Europe needs to ensure that they are available to all Europeans. This requires an open discussion and the involvement of social partners, stakeholders and general public. Many organisations already rely on panels of stakeholders to discuss the use of AI and data analytics. These panels include different experts and stakeholders: legal experts, technical experts, ethicists, representatives of the customers and employees, etc. Actively seeking participation and dialogue on use and impact of AI supports the evaluation and review of results and approaches, and the discussion of complex cases.

- *Diversity and inclusive design teams*

Diversity and inclusion play an essential role in AI systems. It is therefore critical that as AI systems perform more tasks on their own, the teams that design, develop, test and maintain these systems reflect the diversity of users and of society in general. This contributes to objectivity and consideration of different perspectives, needs and objectives. It is not only necessary that teams are diverse in terms of gender, culture, age, but also in terms of professional backgrounds and skillsets.

We invite stakeholders partaking in the consultation of the Draft Guidelines to share their thoughts on additional technical or non-technical methods that can be considered in order to address the requirements of Trustworthy AI.

KEY GUIDANCE FOR REALISING TRUSTWORTHY AI

- Incorporate the **requirements for Trustworthy** AI **from the earliest design phase**: Accountability, Data Governance, Design for all, Governance of AI Autonomy (Human oversight), Non-Discrimination, Respect for Human Autonomy, Respect for Privacy, Robustness, Safety, Transparency.
- Consider technical and non-technical methods to ensure the implementation of those requirements into the AI system. Moreover, keep those requirements in mind when building the team working on the system, the system itself, the testing environment and the potential applications of the system.
- Provide in a clear and proactive manner **information to stakeholders** (customers, employees, etc.) about the AI system's capabilities and limitations, allowing them to set realistic expectations. Ensuring **Traceability** of the AI system is key in this regard.
- Make Trustworthy AI **part of the organisation's culture**, and provide information to stakeholders on how Trustworthy AI is implemented into the design and use of AI systems. Trustworthy AI can also be included in organisations' deontology charters or codes of conduct.
- Ensure participation and **inclusion of stakeholders** in the design and development of the AI system. Moreover, ensure **diversity** when setting up the teams developing, implementing and testing the product.
- Strive to **facilitate the auditability** of AI systems, particularly in critical contexts or situations. To the extent possible, design your system to enable tracing individual decisions to your various inputs; data, pre-trained models, etc. Moreover, define **explanation methods** of the AI system.
- Ensure a specific process for **accountability governance**.
- Foresee **training and education**, and ensure that managers, developers, users and employers are aware of, and trained in, Trustworthy AI.
- Be mindful that there might be fundamental tensions between different objectives (transparency can open the door to misuse; identifying and correcting bias might contrast with privacy protections). Communicate and document these trade-offs.
- Foster research and innovation to further the achievement of the requirements for Trustworthy AI.

V. Assessing Trustworthy AI

The Assessment List provided below is preliminary only, and we invite all stakeholders partaking in the consultation to share their thoughts and expertise on additional items to consider in order to ensure that the requirements for Trustworthy AI are implemented.

The objective of this chapter is to operationalise the implementation and assessment of the requirements of Trustworthy AI set out above, throughout the different stages of AI development and use. We propose the use of an Assessment List for this purpose that offers guidance to steer developers, deployers and other innovators towards ethical purpose and technical robustness.

The primary target audience of this chapter are those individuals or teams responsible for any aspect of the design, development and deployment of any AI-based system that interfaces directly or indirectly with humans, i.e. that will have an impact on decision-making processes of individuals or groups of individuals.

The list proposes questions that should be reflected upon by those leading the assessment. The list is should not be considered as exhaustive, and is only preliminary at this stage. Moreover, the precise questions will vary from use case to use case, and a tailored approach needs to be taken for each specific situation, given the context-specificity of AI.

In the next iteration of this document, which will be published after the consultation, this chapter will consider several use cases to illustrate how the Assessment List can work in practice in highly contextualised settings. Our expectation is that assessments of Trustworthy AI will be aligned with the spirit of the list we describe.

A circular model is envisaged, where the **assessment is continuous and no step is conclusive** (cfr. Figure 3 above). It will include specific metrics, and for each metric key questions and actions to assure Trustworthy AI will be identified. These metrics are subsequently used to **conduct an evaluation in every step of the AI process**: from the data gathering, the initial design phase, throughout its development and the training or implementation of the AI system, to its deployment and usage in practice. This is however not a strict, delineated and execute-once-only process: continuous **testing, validation, evaluation and justification** is needed to improve and (re-)build the AI system according to the assessment.

It should be born in mind that a list-based assessment is not meant as a stand-alone exercise and must be combined with the implementation of management processes embracing an ethical framework for AI.

An Assessment List that addresses the requirements for Trustworthy AI could look as follows:

1. Accountability:

- Who is accountable if things go wrong?
- Are the skills and knowledge present in order to take on the responsibility? (Responsible AI training? Ethical oath?)
- Can third parties or employees report potential vulnerabilities, risks or biases, and what processes are in place to handle these issues and reports? Do they have a single contact point to turn to?
- Is an (external) auditing of the AI system foreseen?
- Was a diversity and inclusiveness policy considered in relation to recruitment and retention of staff working on AI to ensure diversity of background?

- Has an Ethical AI review board been established? A mechanism to discuss grey areas? An internal or external panel of experts?

2. Data governance:

- Is proper governance of data and process ensured? What process and procedures were followed to ensure proper data governance?
- Is an oversight mechanism put in place? Who is ultimately responsible?
- What data governance regulation and legislation are applicable to the AI system?

3. Design for all:

- Is the system equitable in use?
- Does the system accommodate a wide range of individual preferences and abilities?
- Is the system usable by those with special needs or disabilities, and how was this designed into the system and how is it verified?
- What definition(s) of fairness is (are) applicable in the context of the system being developed and/or deployed?
- For each measure of fairness applicable, how is it measured and assured?

4. Governing AI autonomy:

- Is a process foreseen to allow human control, if needed, in each stage?
- Is a "stop button" foreseen in case of self-learning AI approaches? In case of prescriptive (autonomous decision making) AI approaches?
- In what ways might the AI system be regarded as autonomous in the sense that it does not rely on human oversight or control?
- What measures have been taken to ensure that an AI system always makes decisions that are under the overall responsibility of human beings?
- What measures are taken to audit and remedy issues related to governing AI autonomy?
- Within the organisation who is responsible for verifying that AI systems can and will be used in a manner in which they are properly governed and under the ultimate responsibility of human beings?

5. Non-discrimination:

- What are the sources of decision variability that occur in same execution conditions? Does such variability affect fundamental rights or ethical principals? How is it measured?
- Is there a clear basis for trade-offs between conflicting forms of discrimination, if relevant?
- Is a strategy in place to avoid creating or reinforcing bias in data and in algorithms?
- Are processes in place to continuously test for such biases during development and usage of the system?
- Is it clear, and is it clearly communicated, to whom or to what group issues related to discrimination can be raised, especially when these are raised by users of, or others affected by, the AI system?

6. Respect for Privacy:

- If applicable, is the system GDPR compliant?

- Is the personal data information flow in the system under control and compliant with existing privacy protection laws?
- How can users seek information about valid consent and how can such consent be revoked?
- Is it clear, and is it clearly communicated, to whom or to what group issues related to privacy violation can be raised, especially when these are raised by users of, or others affected by, the AI system?

7. Respect for (& Enhancement of) Human Autonomy:

- Is the user informed in case of risks on human mental integrity (nudging) by the product?
- Is useful and necessary information provided to the user of the service/product to enable the latter to take a decision in full self-determination?
- Does the AI system indicate to users that a decision, content, advice, or outcome, is the result of an algorithmic decision of any kind?
- Do users have the facility to interrogate algorithmic decisions in order to fully understand their purpose, provenance, the data relied on, etc.?

8. Robustness:

Resilience to Attack:

- What are the forms of attack to which the AI system is vulnerable? Which of these forms of attack can be mitigated against?
- What systems are in place to ensure data security and integrity?

Reliability & Reproducibility:

- Is a strategy in place to monitor and test that my products or services meet goals, purposes and intended applications?
- Are the used algorithms tested with regards to their reproducibility? Are reproducibility conditions under control? In which specific and sensitive contexts is it necessary to use a different approach?
- For each aspect of reliability and reproducibility that should be considered, how is it measured and assured?
- Are processes for the testing and verification of the reliability of AI systems clearly documented and operationalised to those tasked with developing and testing an AI system?
- What mechanisms can be used to assure users of the reliability of an AI system?

Accuracy through data usage and control:

- What definition(s) of accuracy is (are) applicable in the context of the system being developed and/or deployed?
- For each form of accuracy to be considered how is it measured and assured?
- Is the data comprehensive enough to complete the task in hand? Is the most recent data used (not out-dated)?
- What other data sources / models can be added to increase accuracy?
- What other data sources / models can be used to eliminate bias?
- What strategy was put in place to measure inclusiveness of the data? Is the data representative enough of the case to be solved?

Fall-back plan:

- What would be the impact of the AI system failing by: Providing wrong results? Being unavailable? Providing societally unacceptable results (e.g. bias)?
- In case of unacceptable impact - Have thresholds and governance for the above scenarios been defined to trigger alternative/fall-back plans?
- Have fall-back plans been defined and tested?

9. Safety:

- What definition(s) of safety is (are) applicable in the context of the system being developed and/or deployed?
- For each form of safety to be considered how is it measured and assured?
- Have the potential safety risks of (other) foreseeable uses of the technology, including accidental or malicious misuse thereof, been identified?
- Is information provided in case of a risk for human physical integrity?
- Is a process in place to classify and assess potential risks associated with use of the product or service?
- Has a plan been established to mitigate and/or manage the identified risks?

10. Transparency:

Purpose:

- Is it clear who or what may benefit from the product/service?
- Have the usage scenarios for the product been specified and clearly communicated?
- Have the limitations of the product been specified to its users?
- Have criteria for deployment for the product been set and made available to the user?

Traceability:

- What measures are put in place to inform on the product's accuracy? On the reasons/criteria behind outcomes of the product?
- Is the nature of the product or technology, and the potential risks or perceived risks (e.g. around biases) thereof, communicated in a way that the intended users, third parties and the general public can access and understand?
- Is a traceability mechanism in place to make my AI system auditable, particularly in critical situations? This entails documentation of:
    - *Method of building the algorithmic system*
        - In case of a rule-based AI system, the method of programming the AI system should be clarified (i.e. how they build their model)
        - In case of a learning-based AI system, the method of training the algorithm should be clarified. This requires information on the data used for this purpose, including: how the data used was gathered; how the data used was selected (for example if any inclusion or exclusion criteria applied); and was personal data used as an input to train the algorithm? Please specify what types of personal data were used.
    - *Method of testing the algorithmic system*
        - In case of a rule-based AI system, the scenario-selection or test cases used in order to test and validate their system should be provided

- In case of a learning based model, information about the data used to test the system should be provided, including: how the data used was gathered; how the data used was selected; and was personal data used as an input to train the algorithm? Please specify what types of personal data were used.

- *Outcomes of the algorithmic system*
  - The outcome(s) of or decision(s) taken by the algorithm should be provided, as well as potential other decisions that would result from different cases (e.g. for other subgroups).

Note: As stated above, in view of AI's context-specificity, any assessment list must be tailored to the specific use case in which the AI system is being deployed. To help with the practical operationalisation of the assessment list, we will therefore reflect on four particular use cases of AI, selected based on the input from the 52 AI HLEG experts and the members of the European AI Alliance: **(1) Healthcare Diagnose and Treatment, (2) Autonomous Driving/Moving, (3) Insurance Premiums and (4) Profiling and law enforcement.** These use case, and a tailored assessment list in each of those contexts, will be developed in the final version of the Guidelines.

We invite stakeholders partaking in the consultation of the Draft Guidelines to share their thoughts on how the assessment list can be construed for and applied to the four use cases listed above, and what particular sensitivities these use cases bring forth that should be taken into consideration.

KEY GUIDANCE FOR ASSESSING TRUSTWORTHY AI:

- Adopt an **assessment list** for Trustworthy AI when developing, deploying or using AI, and adapt it to the **specific use case** in which the system is being used.
- Keep in mind that an assessment list will **never be exhaustive**, and that ensuring Trustworthy AI is not about ticking boxes, but about a continuous process of identifying requirements, evaluating solutions and ensuring improved outcomes throughout the **entire lifecycle** of the AI system.

CONCLUSION

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

The AI HLEG recognises the enormous positive impact that AI already has globally, both commercially and societally. 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 will continue to impact society and citizens in ways that we cannot yet imagine. In this exciting context, it is important that due regard is given to ensuring an understanding and commitment to building AI that is worthy of trust, since only when the technology is trustworthy will human beings be able to confidently and fully reap its benefits. When drafting these Guidelines, **Trustworthy AI** has, therefore, been **our north star**.

Trustworthy AI has **two components:** (1) it should respect fundamental rights, applicable regulation, and core principles, ensuring "**ethical purpose**"; and (2) it should be **technically robust** and reliable. However, even with the best of intentions, the use of AI can result in unintentional harm. Therefore, in contrast to other groups, the AI HLEG has developed a framework to actually implement Trustworthy AI, offering concrete guidance on its achievement.

In Chapter I, we articulated the **fundamental rights** and a corresponding **set of principles** and values that underpin **ethical purposes for AI**. In Chapter II, we proposed both **technical and non-technical methods that can serve to help realising and implementing Trustworthy AI**. Finally, in Chapter III we provided **an assessment list** that helps operationalise the achievement of Trustworthy AI. A process is envisaged that will allow stakeholders to formally endorse the final version of these Guidelines, due in March 2019.

The AI HLEG welcomes input from all interested stakeholders through the European AI Alliance as part of the consultation process on this draft. It must be borne in mind that this draft represents the current working document of the AI HLEG, and should be treated in that context at this moment in time.

Europe has a unique vantage point based on its focus on placing the citizen at the heart of its endeavours. Indeed, this focus is written into the very DNA of Europe through the Treaties upon which the European Union in built. This document forms part of a vision that emphasises **human-centric artificial intelligence** which will enable Europe to become a globally leading innovator in AI, rooted in ethical purpose. This ambitious vision will facilitate a rising tide that will raise the boats of all European citizens. Our goal is to create a culture of **"Trustworthy AI made in Europe".**

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

listed here below in alphabetical order

Pekka Ala-Pietilä, Chair of the AI HLEG
AI Finland, Huhtamaki, Sanoma
Wilhelm Bauer
Fraunhofer
Urs Bergmann
Zalando
Mária Bieliková
Slovak University of Technology
Cecilia Bonefeld-Dahl
DigitalEurope
Yann Bonnet
ANSSI
Loubna Bouarfa
OKRA
Nozha Boujemaa, Vice-Chair of the AI HLEG
Inria & BVDA
Stéphan Brunessaux
Airbus
Raja Chatila
IEEE Initiative Ethics of Intelligent/Autonomous Systems & Sorbonne University
Mark Coeckelbergh
University of Vienna
Virginia Dignum
Umea University
Luciano Floridi
University of Oxford
Jean-Francois Gagné
Element AI
Chiara Giovannini
ANEC
Joanna Goodey
Fundamental Rights Agency
Sami Haddadin
Munich School of Robotics and MI
Gry Hasselbalch
The thinkdotank DataEthics & University of Copenhagen
Fredrik Heintz
Linköping University
Fanny Hidvegi
Access Now
Eric Hilgendorf
University of Würzburg
Klaus Höckner
Hilfsgemeinschaft Blinden & Sehschachen
Mari-Noëlle Jégo-Laveissière
Orange
Leo Mikko Johannes Kärkkäinen
Nokia Bell Labs
Sabine Theresia Köszegi
TU Wien
Robert Kroplewski
Solicitor & Advisor to Polish Government
Pierre Lucas
Orgalime
Ieva Martinkenaite
Telenor
Thomas Metzinger
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
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 Ivanov Sharkov
Digital SME Alliance
Philipp Slusallek
DFKI
Françoise Soulié Fogelman
AI Consultant
Saskia Steinacker
Bayer

Jaan Tallinn
Ambient Sound Investment
Thierry Tingaud
STMicroelectronics
Jakob Uszkoreit
Google
Aimee Van Wynsberghe
TU Delft
Thiébaut Weber
ETUC
Cecile Wendling
AXA
Karen Yeung
Birmingham University