Response on the draft ethical guidelines for trustworthy AI produced by the European Commission's High-Level Expert Group on Artificial Intelligence.

The AI and Robotics group at the Tilburg Institute for Law, Technology and Society

Contributors: Merel Noorman, Esther Keymolen, Maurice Schellekens, Aviva de Groot, Silvia da Conca, Robbert Coenmans (LL&SP department), Ronald Leenes, Bo Zhao, Lorenzo Dalla Corte, Emre Bayamlioglu, Robin Pierce, Linnet Taylor

31 January 2019

The HLEG has taken on the ambitious project of developing general guidelines for ethical AI and has, a first step in this process, published a draft document for stakeholders to review and comment upon. In this draft document, the group builds upon, and rightly so, a range of existing frameworks, principles and manifestos. It proposes to centre the guidelines on the concept of Trustworthy AI. They elaborate this concept in three sections that each address different levels of abstraction: ethical purpose rooted in fundamental rights, technical and non-technical methods, and an assessment list.

We would like to congratulate the HLEG on this first step in a complex and multifaceted process and complement the group on finding a shared basis to further build upon. In particular, we welcome the rights-based approach that HLEG chose to pursue, as it roots the guidelines in shared values and principles within Europe while at the same time aligning them with many of the existing guidelines.

Moreover, we were pleased to see the substantive definition of AI as it is outlined in the document published in parallel with the guidelines and summarized in the draft document. In particular, by distinguishing between AI as a technology and artefact designed and deployed by human beings on the one hand and AI as a scientific discipline on the other, the authors have managed to highlight the extensiveness and heterogeneity of AI. They have also signalled the human agency and work that is involved in making these AI systems function. The focus in the definition on the pre-determined goals and parameters provides regulators something to work with.

The HLEG also brings the discussions on ethical AI a step forward by not only focusing on rights, principles and values, but also on the implementation and embedding of the technology. The ambition to provide concrete tools and methods for policy makers, developers, and citizens is needed to bring ethical AI into practice and we encourage further work in this direction.

As the HLEG has explicitly asked for critical feedback we would like to offer a few suggestions and comments for the further improvement of the document. We will first provide some general comments and then go into more specific comments per section of the guidelines document.

General (conceptual) comments

Our general comments focus primarily on the conceptual elaboration of key terms and ideas in the document. In particular, we argue that the concept of trustworthiness needs to be centred on vulnerability and uncertainty; AI should be treated as embedded in a larger sociotechnical context and the benefits and risks of AI require a more nuanced consideration.

Trustworthiness
The HLEG has chosen to centre the guidelines on the concept of trustworthiness, which has the potential to provide a useful tool for the different audiences of the document to structure their thinking about how to proceed (or not) with AI. Trust, according to the HLEG, is the cement of societies, communities, economies and sustainable development. Early on in the document, the HLEG defines Trustworthy AI as consisting of 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." (p. 1)

However, the current use of the concept in the document runs the risks of providing another rhetorical tool for parties involved to carry on with business as usual, while claiming to have adopted an ethical approach to AI. The elaboration of trustworthy AI in the draft document suggests there is technological fix for a possible lack of trust, namely ensuring ethical purpose and technological robustness and reliability. Once these two criteria have been met, citizens and others will be able to maximize control and minimize risks and thus not have to worry about AI.

Yet, trust is fundamentally about vulnerability and uncertainty. We trust someone when we know there are uncertainties and potential risks, but we are nevertheless willing to work with or rely on them. Unfortunately, this uncertainty and vulnerability is underexposed in the current guidelines and should, in our opinion, be put centre stage. Trustworthy AI should be about how we deal with the uncertainty and vulnerability that come with the development, deployment and use of AI. Technological reliability and robustness are to be encouraged, but what happens when things go wrong? What gives us that trust that things will work out despite the uncertainties and vulnerabilities? Are citizens sufficiently informed about the potential risks of AI technologies? Is it only the AI system that should be reliable, or should the sociotechnical system in which it is embedded also be reliable and robust? Will AI systems afford trust in institutions? When should we cultivate a healthy distrust? The guidelines should emphasise that trust is a means of dealing with the unknown.

AI as embedded in a larger sociotechnical context
In part, the narrow conceptualization of trustworthiness, as reflected throughout the document, is the result of a tendency of the authors to treat AI as a monolithic autonomous thing, isolated from its context. On several occasions the authors attribute agency to AI in such a way that is obscures the work done by human beings. In the Executive Summary, the authors note that human beings will only be able to confidently and fully reap the benefits of AI if they trust the technology. However, it is not the technology that we need to trust, as the authors justly note in a later section of the document, when they argue:

> "Trust in AI includes: trust in the technology, *through the way it is built and used by human beings*; trust in the *rules, laws and norms that govern* AI […] or trust in the *business and public governance models* of AI services, products and manufacturers." [emphasis added] (p. 2)

In our view, the question then is how do (1) the human beings that build and use these technologies; (2) the rules, laws, and norms that govern these activities, and (3) the business and public governance models for these technologies deal with uncertainty and vulnerability in such a way that they foster trust? Although focusing on ethical purpose and technical robustness and reliability, are part of this, it is not sufficient. Ensuring that the design, development and deployment of AI respect rights and regulation, and adhere to core principles and values, and making technology robust and reliable, will not simply dissolve the uncertainties and vulnerabilities. Although, the authors address uncertainties and vulnerabilities in Chapter 2 - for instance through the requirement of accountability and data governance - the two elements should be at the heart of the definition of trustworthy AI.

The author's treatment of AI is also problematic in the operationalization of trustworthy AI. Despite the nuanced definition that it has been given in the separate document, the draft document does not position AI - as a system or scientific discipline - within the broader discussion on the ethics of digital and data-driven technologies. AI techniques are rarely developed or used as standalone systems, but are embedded in broader eco-systems. They are components of decision-making processes that extend throughout and beyond organizations and involve multiple human beings in different roles. AI techniques, are for example, used by platform companies to analyse data obtained through social media to optimize advertisement targeting for different companies. Many critical analyses have been made of the network dynamics involved in these developments, e.g. monopolization, power asymmetries, etc.. Yet, the problem of growing power asymmetries as a result of large scale datafication of society, partly enabled by AI technologies, is not discussed in the document. Or perhaps, one could argue, that it is addressed in the requirement of respect for human autonomy, but only in a very abstract sense without reference to other developments that fuel these asymmetries. By positioning AI as an isolated thing, we lose sight of all the elements in the sociotechnical systems that contribute to its trustworthiness, such as human relationships and stabilizing institutions.

Benefits and risks

Finally, the authors suggest in the Executive Summary that on the whole the benefits of AI outweigh its risks and we should therefore invest in maximizing the benefits while minimizing the risks. However, given the abstractness of AI as used in the document, this is an empty and misleading statement. It suggests that there is some way of measuring the benefits and risks of this very ambiguous and broad thing called AI and that there is a moral imperative to pursue the benefits. It places the critical citizen, company or organization in the awkward position of not being able to refuse the technology or opt-out of its use. AI has to be used, and therefore we should make it trustworthy. This precludes a lot of other options. Perhaps this is also a reason for the absence of a discussion on the precautionary principle in the document.

Comments on the Chapter on Ethical purpose

While applauding the rights-based approach of the HLEG, we would encourage the authors to avoid the conflation of fundamental rights and ethics in the concept of 'ethical purpose'. Law and ethics are two separate domains that need to be clearly distinguished with regard to their rationale and function. Not doing so, runs the risks of obscuring or down-playing the central role of law in the governance of the design, deployment and use of AI and to reinterpret ethics as 'industry self-regulation'. Yet, ethics should go above and beyond the law. Moreover, we would like to note that the authors spend some time on explaining the cycle from rights to principles to values. However, they let this cycle go in the following chapters.

A minor point to be made here, is that on page five the authors suggest that they derive the principle of autonomy from human dignity, but it seems to us that this principle should derive from freedom and liberty.

Our further comments on this section focus in particular on the interpretation of the fundamental rights, the selection and description of the principles and the reasoning behind the critical concerns on AI.

Fundamental rights

The Chapter on ethical purpose that elaborates on the fundamental rights requires re-thinking and re-structuring. In particular, the section on fundamental rights lacks a firm logical structure. It is not clear what the authors are basing their decisions on for their interpretations of these rights. We recommend that the HLEG links the description to existing frameworks in a more systematic and objective way.

The description of the fundamental rights currently appears to be an ad hoc creation. In particular, paragraph 3.4 seems peculiar and does not match with existing legal frameworks. The five sentences in this section contradict each other. Equality does not mean equal treatment of everyone regardless of the situation. Rather, it means people should be treated equal in equal situations and unequal in unequal situations according to their unequalness. Equality in AI should be about neutrality in access and how it applies to you and affects you. Combining equality and the rights of minorities, might make sense from a natural language perspective, but not necessarily from a legal perspective: if all situations are treated equally, it is impossible to protect minorities, which by definitions are in a different position than the majority and require a different treatment.

Another curious element in this section is the suggestion that consumers and workers are minorities. We assume that the authors intended to note that equality requires respect for the position of less powerful stakeholders, such as consumers and workers. At the same time, it is a bit odd to find the word consumer in a human rights framework instead of citizen.

Section 3.5 about citizen rights is suddenly very specifically aimed at the public sector, while corporations are completely left out. In our view, this is an omission. Citizens must be protected from preventive intervention by corporations. AI should not be employed to inhibit citizens' rights in their relations with public and commercial institutions alike.

What is also missing in this section - in fact it is mostly missing from the entire document - is the respect for human relations and the environment.

The fundamental rights section is a key part of the guidelines document. The authors note that "[c]ompliance with these Guidelines in no way replaces compliance with [fundamental rights and with all applicable regulations], but merely offers a complement thereto" (p.2). It therefore needs to provide a solid foundation and should be linked explicitly to existing fundamental rights. Particular interpretations should not depart from those that are accepted within these frameworks. Yet, in its current form, the document seems to argue for a lower standard then is currently set by the existing legal framework.

Ethical principles

The HLEG proposes five principles for AI: beneficence, non-maleficence, autonomy, justice and explicability. The first four principles are well known leading principles in the medical, care and bioethics domains. As such, these principles are in line with the principles put forward by existing ethical principles in various fields, including computer ethics, as well as principles offered by various proposed AI ethical guidelines.[1] The HLEG has added the fifth principle of explicability. These principles are intended to guide the operationalisation of core values derived from the fundamental rights.

It is curious that AI would turn to the four principles of Beauchamp and Childress given the rigorous controversy that surrounds both the usefulness and insufficiency of "The Four Principles" as a moral framework. In 1995, leading UK bioethicist, Soren Holm argued that "The theory [the four principles of bioethics] is developed as a common-morality theory, and the present paper attempts to show how this approach, starting from American common-morality, leads to an underdevelopment of beneficence and justice, and that the methods offered for specification and balancing of principles are inadequate."[2] The reference to these principles without incorporation of the fruits of the robust critique that has ensued in the past twenty years and has led to more nuanced and useful engagement with ethical principles and what they require is a missed opportunity. Greater nuance and attention to the widely-recognised limitations of the four principles is strongly encouraged.

The authors have chosen to highlight the principle of explicability for AI, which is understandable given the seeming complexity of AI systems. The authors define explicability in terms of transparency, where transparency is about the auditability, comprehensibility and intelligibility of AI systems as well as about the awareness of the intentions underlying particular business models. This kind of transparency, according to the HLEG, is needed for informed consent. Explicability, according to the document, is primarily about consent as a form of control for citizens. Yet, on its own consent is a very weak instrument, and should not form the primary basis for requiring these practices.

[1] In their pre-study, Cowls and Floridi argue how many or even most AI principles in other frameworks and guidelines that are currently being drafted could be categorized under these headers.

[2] Holm, S. (1995) Not Just Autonomy—The Principles of American Biomedical Ethics. *Journal of Medical Ethics* **21**, 332-338.

The strong connotation with the medical and bioethics domain also makes us doubt whether the addition of the fifth principle of 'explicability' is the right choice. The notions of explaining (and understanding) are very much already a part of the 'mother' framework, notably, the autonomy of patients/subjects is served through the informed consent paradigm. Where the other principles are 'ends in themselves', explicability is not. The side effect of singling out explicability is that it diminishes the importance of informed consent in the principle of autonomy. By separating it from the principle of autonomy, it narrows human autonomy to the ability to choose to opt in or out and not much more. Yet, the kind of transparency that the HLEG proposes should empower human beings to not only be able to choose to opt-in or out, but to understand why one might choose to do so, under what conditions and how to object to the framing of the offered choice.

Moreover, the principle of explicability does not reflect the current state of AI and our current limited understanding of what explicability *should* mean. It is key that we will find ways to make these technologies sufficiently understandable to serve responsible use, but as we have not yet defined how to serve this understanding, it is unclear what 'explicability' is or should be. Variants of explanation may thus be offered to comply with this principle, and it will be hard to argue why these will or will not serve the purposes as contained in the other principles.

At the same time, the principle of explicability seems to undermine the importance of trust, as trust is needed when things are not transparent; when you do not know the innerworkings of something. If the focus is on trust, the question is how do we deal with not knowing for sure? It is not enough to have an indication that AI will function as expected; trust also comes from knowing that there is a safety net when things do not work out as expected. This is one important reason to have a strong legal framework. If AI and the human beings behind it are able to explain what the systems does and why, then it helps citizens if they can legally hold those responsible to their explanation. If AI raises certain expectations, it is important that citizens can legally hold those people behind the AI accountable for the impression the AI made. The law functions here as a safety net. Citizens are not left to their own devices, if things go wrong and AI and the human beings behind prove not worthy of citizens’ trust.

In terms of human autonomy, we suggest that the authors reflect on the connection between human and AI autonomy. On the one hand to clearly distinguish between the two, but on the other hand to also signal the importance of human autonomy in meaningful human control (Human oversight of AI autonomy). Although citizens can be attributed responsibility for the behaviour of certain systems, they have to be in the position to exercise their autonomy in order to be able to live up to the responsibility. That is, they should not only be able to understand AI, but they should also have sufficient discretionary power and be able to intentionally affect the behaviour of the system directly or indirectly. Moreover, it should be noted that it is generally not one individual that is responsible. Responsibility should be appropriately and fairly distributed across stakeholders in accordance with the level of control or influence they have.

One final note on the principles. Although solidarity is mentioned in passing, it is underdeveloped in the description of the principles. The emphasis is on human individuals, but not on the relationships and common bonds between them. This disregards the influence of AI on the social whole(s), while placing too much burden on the individual as the actor that needs to know/trust/understand.

We look forward to seeing the further elaboration of the case-studies that the HLEG has announced for the following version of the document. This would be a valuable contribution to the document as it will help to demonstrate how certain rights and principles are relevant and applicable to particular AI designs and uses, as the HLEG notes. For example, the rule of law might not be that relevant for a smart-refrigerator, but key to law-enforcement AI systems. Similarly, not every system should have an opt-out option, but those that do not have an opt-out should adhere to much stricter requirements in terms of the rule of law. Moreover, it would allow the authors to address the issues of potential conflicts between principles as well as between rights. With regard to the domains chosen for case-studies, public administration systems seem to be conspicuously missing.

Critical concerns raised by AI
In the final section of this Chapter, the HLEG presents a list of possible concerns about AI and notes that this has proven to be a contentious part of the process. Although we appreciate the concerns about the future developments in AI, it is unclear what the connection of this list of somewhat generic scenario's is to trustworthy AI. Moreover, what are the reasons for choosing these scenario's and leaving out others? What is missing, for instance, are economic concerns, such as the influence of AI on the labour market or growing (income) inequality? We would have expected a more systematic evaluation of what the possible risks of the proposed trajectory towards trustworthy AI might be.

We would propose to start this section by establishing some criteria for elaborating the critical concerns. One such criterion could be the timeframe to look at. Subsequently, we suggest linking the concerns to the work already done in the previous chapters, and examining how developments in AI could threaten trustworthiness or what would happen if fundamental rights and principles are not respected. For instance, what happens when the principle of autonomy is ignored? A more systematic analysis would connect this section better to the preceding sections and would enrich the elaboration of the concept of trustworthiness.

Some minor comments on this section are perhaps also worth noting:
- There is no such thing as "“anonymous” personal data" (p. 11).
- Only one of the legal bases for personal data processing has been discussed, despite the reference to art. 6 GDPR (*idem*).

Comments on the realization Chapter

Chapter 2 provides an overview of possible methods to implement the ten requirements that have been derived from the principles in the previous Chapter, including accountability, data governance and respect for privacy. Although these requirements are certainly some that AI developers and users should adhere to, it is difficult to evaluate these principles at the given level of abstraction. Moreover, it is again unclear why the HLEG chose these particular requirements and left others out. For example, why is environmental sustainability not part of this list? Also, as part of the principle of explicability the authors mention that informed consent should be a requirement, yet it is absent from the list and does not come back as part of the requirement of transparency. Similarly, robustness is a requirement and thus elaborated, but reliability is not.

With regard to the requirement of accountability, it should be noted that accountability is not just about compensation, but also about learning and adjusting existing practices in order to prevent or minimize the risk of untoward events from occurring again. Here again, it would be helpful if the authors could provide some reasons for choosing particular interpretations of concepts. In the section on non-discrimination, they provide some references in support of particular definitions. They might want to do this for the other sections as well.

The relations and possible conflicts between the requirements also warrant further elaboration. For example, the authors illustrate the relation between other values such as non-discrimination. However, they do not mention the potential conflicts between privacy and identifying and correcting problematic bias. Nor do they address the potential conflict with transparency or the supplementary, mutual support between privacy and human autonomy or safety.

To implement the requirements, the authors provide a list of technical as well as non-technical methods. Although potentially helpful as a starting point, we offer a few suggestions for consideration:

- Although audibility is mentioned in the technical section, it is missing from the non-technical section. No mention is made of the kind of institutions or mechanisms that are necessary to audit these technologies. Moreover, the authors may want to further elaborate the notion of auditability.
- Democratic decision-making is missing from both the technical and non-technical methods.
- The support of interpersonal relationships to foster trust is missing.
- Learning and training with new systems is missing.
- Developing new protocols for the deployment and use of AI etc. could be added to the list of non-technical methods.

Comments on the assessment list

In the last part of the document the HLEG provides an assessment list “to operationalise the implementation and assessment of the requirements of Trustworthy AI set out above, throughout the different stages of AI development and use.” This is a potentially helpful way of providing a concrete tool for the intended audience to work with and there is definitely a demand for such a list. However, we feel that such a list or set of lists would be very context sensitive and it is therefore difficult to comment on this rather abstract list without the necessary context.

Nevertheless, we would like to point out a few issues that may inform future work on the assessment list(s). In particular, one question that is not in the assessment list is whether the use of AI is justifiable given the circumstances. Are there other better ways of solving a particular problem? In addition, the document does not discuss (unintended/unanticipated) interactions with other systems nor the embedding of the system in existing practices. For the successful adoption of AI systems, these are things that need to be taken into consideration. Finally, the assessment list is currently lacking an operationalization of the value-sensitive design approach (e.g. stakeholder inclusion, weighing of different values).

We would like to conclude by once again congratulating the HLEG on this first step in developing the ethical guidelines. We hope our feedback and suggestions will contribute to further fine-tuning of the guidelines document.

For questions or further information please send an email to m.e.noorman@uvt.nl