EUROPEAN HUMANIST FEDERATION

Stakeholders' Consultation on Draft AI Ethics Guidelines

Contribution of the European Humanist Federation

Introduction: Rationale and Foresight of the Guidelines

The introduction of the working document prepared by the HLEG AI sets out the basis for the construction of the next three chapters devoted to the ethical purpose, the realization of "trustworthy" AI and finally its assessment.

It is therefore important that all the assumptions, principles and objectives on which the rest of the document is built are properly identified. In this respect, the European Humanist Federation feels that the following key elements are missing from this methodological basis or the consideration given to them is too shallow. This echoes throughout the rest of the document.

Benefits vs. Risks

The document takes as an axiom the fact that "on the whole, AI's benefits outweigh its risks" without providing convincing evidence to back up this claim.

As humanists, we are committed to technological progress and we measure the extent of the economic and strategic potential of AI. However, the understandable race to reap the benefits of this very potential should not result in a lack of methodological rigor, especially when the matter at hand is to draft ethical guidelines.

Social acceptance of risk.

In highly formalized administrative systems such as financial loan decisions, web searches, online customer services, personalized marketing based on social data, financial speculation, etc. intelligent tools are booming and provide outstanding results. If the algorithm used in a specific application respects what would have been a human decision and if the database that is used to train it is sufficiently exhaustive, there is in principle no bad surprise. However, even if there cannot be any guarantee of control in the design phase, it has to be possible to correct potential biases or inconsistencies post hoc.

The document rightly suggests that the EU has to find a "road that maximises the benefits of AI while minimising its risks [and that] to ensure that we stay on the right track, a human-centric approach to AI is needed."

It however fails to clearly recognize and acknowledge that risk zero does not exist. This in turn means that we have to ask a second fundamental question: What level of risk are we willing to accept, socially speaking?

Trustworthy AI vs. societal validation of AI

The answer to the above question can only be given via democratic processes. It therefore becomes clear that many more efforts have to be invested in educating society about the risks represented by AI technologies. This becomes all the more critical since citizens are heavily impacted by the use of AI but are often not even conscious of the fact that other choices were theoretically possible.

This dimension is addressed to some extent in the document but only marginally, despite the fact that it is absolutely central as it will define the level to which society will, on the long run, trust AI technologies.

Beyond education – and this is also acknowledged in the document to some extent – end users should be involved at all levels of the design of AI services: from conception and design (ex-ante validation) to feedback after usage (post hoc validation and recourse).

Post hoc mechanisms should not only be put in place within design teams at the discretion of AI developers. On the contrary, they should be systematic. This would result in strengthened public debate and informed societal oversight. Therefore, we would propose to add a third component to the definition of Trustworthy AI:

> "it should ensure an **ethical purpose**, it should be **robust** and should be **socially controlled** on an ongoing basis."

We therefore propose the creation of a **European Observatory of AI Technologies and Services** in charge of implementing social control at any stage of the design, deployment or use, including post hoc end-user return of incidents.

In this sense, AI technologies would not only be trustworthy, they would actually be trusted.

An EU observatory would also address two other elements that the document rightfully captures: given the nature and pervasiveness of AI technologies and the necessarily unknown future developments:

- a one-size-fits-all approach does not apply
- ethical guidelines will have to be regularly re-debated and updated

Chapter I: Respecting Fundamental Rights, Principles and Values - Ethical Purpose

Chapter I provides and overview of the principles, rights and values that an ethical approach to AI should entail. While the structure of the chapter seems appropriate, key elements seem to be missing from the overall reasoning. Certain sections also need to be further refined.

Informed consent and societal control

As it was the case in the introduction, this chapter as well considers users and citizens as mostly passive actors of the development of AI systems. Section 2 takes "informed consent" as the basis for operating trustworthy AI whereby people are to be "given enough information". As humanists, we think that citizens should be much more empowered to become actors of the development of AI.

On the one hand, it is troubling that the acknowledgement that current practices - that clearly show that end users give consent without consideration despite being informed - is part of the section where the HLEG AI seems not to have reached consensus.

On the other hand, even if there was consensus within the HLEG AI, given the impact and the pervasiveness of AI technologies, mere consent is not enough, even if it is "informed". End users, citizens, workers and society as a whole have to have a much more active role in the entire life cycle of AI technologies: from design to usage, including ex ante but also post hoc validation. The relationship between users and developers has to be bidirectional and continuous.

This first of all entails that the concern about education should be central to the question of the future of AI. Without proper education citizens will not be able to reap the benefits of AI and minimize the risk of its usage. Their emancipation and their free will could be severely hampered, without them even realizing it. Current debates relating to the impact of social media on election outcomes clearly demonstrates this.

However, beyond education, public debate about AI should be actively fostered by systematic societal control and oversight. This is why we propose the creation of an EU Observatory of AI Technologies and Services, which would be in charge of implementing this societal control, including post hoc return on incidents for individual users.

Instead of weakening it, the results of increased civic engagement in the development of AI would help fostering the trust of society in AI technologies by deconstructing certain myths and providing substance to many of the complex issues outlined in the HLEG AI Working document. The proposal to involve people belonging to minorities and specific demographics to reduce the risk of reinforcing discriminative patterns present in society would be part of such a mechanism.

The principle of autonomy and human agency

The principle of autonomy and human agency are fundamental to the AI debate. The document rightly identifies the need to guarantee the right of people to know whether they are interacting directly or indirectly with AI systems, their right to know and reject being subject to direct or indirect AI decision making and their right to opt out and withdraw.

It would however be of utmost importance to complement this aspect with the concept of human supremacy over AI decision-making. Although the idea is expressed to some extent in other chapters, the principles of autonomy and "do no harm" have to fully encompass this idea. When situations become critical, e.g. when lives are in danger, when the risk element comes into question, when non-quantifiable moral dilemmas enter into play, humans have to retain control. It is therefore necessary that regulation and intervention by humans remains possible at all times.

The principle of explicability

The working paper rightly elevates the principle of explicability to one of the key principles upon which to base the development of AI in the future. We welcome putting stress on such an important dimension. However, this section as well considers "informed consent" as a basis for usage of AI services and as experience shows, this is not enough.

Furthermore, the document proposes that informed consent be based on the possibility for individuals or groups to request evidence about the instructions and inputs that lead to a certain output, the organisations involved, etc.

Instead of considering this an option, proper intelligible explanatory mechanisms on the main parameters, instructions and inputs, their correlation to the outputs, and the role and responsibility of all actors involved in the AI decision in question should become the rule. This would also ensure that the outcome serves the user rather than the commercial interests of certain actors, including third parties, at the expense of users.

Without the availability of such explanatory mechanisms, traceability will be undermined and responsibilities diluted. However, as discussed in later chapters of the document, research in explainable AI is in its infancy. This is why, once again, societal control is of utmost importance.

Long term risks

Considering the last section of this paragraph, it is highly alarming that the HLEG AI cannot reach a consensus on threats as basic as the ones listed in the text. Many of these threats are well documented and should not spark controversy, but rather to trigger the finding of responsible answers.

One has the intuition that many of these controversies are linked to the tremendous economic and strategic benefits that AI promises to those who manage to establish themselves in the global market.

While the economic incentive is understandable, it cannot overshadow the strict requirement to abide by our democratic principles, values and rights as described in the first 4 sections of this chapter.

Furthermore, certain threats seems to have been relegated to mere technical issues, to be dealt with in chapter 2, such as the issue of discriminative biases resulting from social data carrying discriminative tendencies present in society.

- **Reinforcing discriminative patterns present in society**

From a humanist point of view, one of the main risk concerns the possible reinforcement of forms of discrimination and the possible picking up by algorithms of reactionary and exclusionary social stereotypes.

An algorithm may be conceived biased from the beginning, as a conscious or unconscious consequence of bias held by its makers.

That was seemingly the case of a facial recognition software introduced by Google where a young African-American couple realised that one of their photos had been tagged under the “gorilla” tag. The explanation lied in the data with which the algorithm was trained to recognize people. In this case, it is likely that it mainly, if not exclusively, consisted of pictures of white people. As a result, the algorithm considered that a black person had more similarity to the “gorilla” object that it had been trained to recognize than to the “human” object.

In other cases, it may be unclear whether the bias and discrimination are the result of the algorithm itself or of its interaction with users.

That is the case of the gender bias revealed in the functioning of "Adsense", Google's advertising platform. In 2015, researchers from the Carnegie Mellon University and the International Computer Science Institute highlighted that it was biased at the expense of women. Using a software called "Adfisher", they created 17,000 profiles and simulated web browsing to conduct a series of experiments. They found out that women were systematically offered lower paid jobs than men with a similar level of qualification and experience. The precise causes are difficult to establish. It is of course conceivable that such a bias was the result of the will of the advertisers themselves: they would then deliberately choose to send different offers to men and women. It is however also possible that this phenomenon is the result of the algorithm's learning process. In this case, men may on average have been more inclined to click on ads advertising the highest paid jobs, whereas women would have resorted to self-restrain, following mechanisms that are well known and described in social sciences. Therefore, the sexist bias resulting from the functioning of the algorithm would be nothing more than the reproduction of a pre-existing bias in society.

In other cases, the discriminatory result may be totally unintentional.

In April 2016, it was revealed that Amazon had excluded from one of its new services (free home delivery in 24h) neighbourhoods mainly populated by disadvantaged people in Boston, Atlanta, Chicago, Dallas, New York and Washington. Initially, an algorithm from Amazon had found, by analyzing the data at its disposal, that the neighbourhoods in question offered little opportunity for profit to the company. Even though Amazon's objective was certainly not that of excluding any particular area from its services because of their predominantly black population, this proved to be the result of the use of this algorithm. It is therefore obvious that Amazon's algorithm had the effect of reproducing pre-existing discriminations, even if no intentional racism was here at work.

Even more evident of a non-intentional result was the case of Microsoft's Tay, a "learning" robot supposed to enter into conversations on Twitter. In less than 24 hours, Tay converted from its humanist and politically correct original attitude to a racist, sexist and xenophobic discourse, as a consequence of its interaction with what people were writing in their responses. Microsoft apologized and recalled that Tay had been built on the basis of "cleaned up" and "filtered" public data. This ex ante precaution clearly turned out not to be sufficient, once the algorithm was left to operate "autonomously" on Twitter and in interaction with other non-proprietary data.

This poses a real question: how to train algorithms and AI to use public data without incorporating the worst traits of humanity?

We should therefore be aware that the risk of AI becoming the vehicle for reinforced bias and discrimination may depend: 1) on the choices made by the programmers that create the algorithm; 2) on the data absorbed by the system in its interaction with the public; 3) on the simple circumstance that sometimes the "logical" choice is inconsistent with our ethical and constitutional values.

- **Commercial interests of third parties – the example of medicine**

Even if, in the future, final decisions will be (and should be) taken by people, a technical pre-structuring and influencing of these decisions will be possible, if not even likely. The opportunities to support medical decisions and therapies that AI offers are promising, and sometimes breath-taking. This concerns the future of clinical care as well as care. We can assume that AI systems will bring about relevant changes in

this area. However, especially with regard to patients' autonomy, we should be careful and avoid AI recommendations and decisions that are subject to bias.

AI should work for the benefit of human beings. Regarding medical ethics, it necessarily implies that the decision on the appropriate therapy must be based on knowledge and analysis and not depend on the potential benefits of third party interests. Given the existence of current unfair business practices, it is reasonable to highlight the danger of cases where therapeutic choices would be influenced by a selective use of data or (hidden) algorithms that would include the economic interests of health insurances or health care institutions.

In order to avoid this, the functioning of AI in the medical field must be of the utmost transparency and explainability.

This does not only apply to general therapeutic decisions or procedures, but also to situations at the end of life. For it is precisely here that the individual, autonomous and responsible will of the patient must be the decisive criterion. In this particularly vulnerable and complex ethical situation, the patient's will is to be respected in the widest possible sense. It is critical to guarantee that algorithms cannot hinder or make impossible the implementation of the will of the patient because ideological convictions of third parties or economic interests of institutions become decisive, perhaps without this even becoming apparent.

- **Other domains**

Further domains raise even more questions and seems to require more in-depth reflection, debate and deliberations. This is the case for instance of Lethal Autonomous Weapon Systems. The ethical challenges related to this field of application are enormous, especially when one considers the extreme economic and strategic benefits involved. The fact that the HLEG AI has not reached consensus on this question is very concerning. More importantly however, the critical nature of the question suggests that it requires a much wider societal debate. Such a debate could be steered by the EU Observatory on AI proposed by the EHF. A similar question is linked to the way terrorist organisations use existing AI algorithms used to track user preferences and tailor ad contents for purposes other than what they were originally designed for.

Chapter II: Realising Trustworthy AI

The fact that Chapter 1 left aside a number of issues or did not given them strong enough consideration results in these elements not being addressed enough throughout Chapter II.

While the fact that the list of requirements discussed is acknowledged by the paper itself as non-exhaustive is on the one hand laudable – as indeed it lacks key elements – it is difficult to see how the current list – even if enhanced – will not become some kind of baseline in the future. It has to be clear from the outset that the nature of AI and the necessarily yet unknown applications and services that it will bring about carry the fundamental need that this set of "requirements" be continuously reviewed and submitted to societal oversight and validation. The document rightly recognizes this. However, the way this will be guaranteed and the way its current contents will be debated in the wider society are not clear.

Creating a European Observatory of AI applications and services, as proposed by the European Humanist Federation, would definitely be a strong signal pointing in this direction.

Accountability, autonomy, safety and transparency:

In order for accountability to be implemented in practice, three elements are of utmost importance. First, users have to be provided with the tools to detect and understand anomalies and dysfunctions. This however presupposes that explainability issues are properly addressed. Second, procedures should be in place, allowing them to lodge complaints to specialized bodies creating a level playing field between them and the legal departments of private companies developing AI. Finally, thorough legal research has to prove that AI's specificity does not create loopholes that could be exploited at the expense of the consumer or user.

It appears from the section on explainable AI research (XAI) that explanability of AI systems has not yet reached a satisfactory level, not by a longshot. In turn, this means that a number of requirements expressed in this chapter remain at the level of laudable intentions which however, cannot be yet be followed by deeds.

This increases the importance of societal control and debate about the level of risk that we, as a society, are willing to accept. Indeed, as long as explainability remains poor, increasingly pressing questions will surface regarding the legitimacy of AI decisions, given their (sometimes very difficult to detect) impact on individuals and on society as a whole. The existing case of discriminative biases clearly demonstrate this.

Furthermore, as mentioned in the previous chapter already, human oversight and the possibility for humans to intervene is fundamental. This is all the more true in critical applications where uncertainty and risk or the presence of moral dilemmas that cannot be expressed in terms of quantifiable parameters require a decision based on human judgment.

Technical and non-technical ways to achieve trustworthy AI

The second part of chapter II concerns technical and non-technical ways to achieve trustworthy AI. Without diving into technical considerations, the lacking third dimension of the definition of "trustworthy AI" throughout the document – societal control - also echoes in this section.

Naturally, creating secure architectures with fallback mechanisms, testing of systems and their auditability are important. However, since the explainability and the traceability of AI systems is difficult to guarantee, technical approaches aimed at avoiding issues ex ante have to be complemented with post hoc evaluation of usage by consumers and a systematized integration of their feedback into an ongoing societal oversight and debate related to AI applications.

When it comes to the non-technical ways discussed in the working paper, the EHF believes that, in the sector of AI, the importance of safeguarding ethical and democratic principles, it is difficult to see how certain aspects can be guaranteed without regulation. We will follow with interest the second deliverable of the HLEG AI. The content of that deliverable will complement this one and a final opinion on their joint relevance will be possible when both documents are finalized.

In any case, responses to the exposed issues – whether these are codes of conduct or standardization – have to be prompt and be carried out at European level so as to make it possible to leverage the weight of the Single Market and impose a set of high ethical standards at global scale.

As expressed throughout our response to this consultation, we propose the creation of a European Observatory of the use of AI, including the design and management of feedback systems allowing to flag

incidents, in a similar way as it already exists in sectors of high risk technologies (e.g. nuclear, aeronautical).

Furthermore, the entire "algorithmic chain", from the algorithm designer to the professional user, including engineers, data scientists or coders must receive training on the ethical dimension of their sector. Such trainings should highlight the need for transparency, traceability and intelligibility of systems. As highlighted by the HLEG AI, programmes aiming at increasing diversity in in design teams would also have a positive contribution.

Finally, as recognized in the document, citizens must be aware of the functioning, the problems and the risks related to artificial intelligence. This implies that school curricula raise their awareness about the reality of algorithms and promote genuine education in terms of values, citizenship and critical thinking. Beyond school, public authorities must develop awareness programs on these issues and foster public debate on artificial intelligence in general. This should be a priority of EU policies in the domain of AI.

Chapter III: Assessing Trustworthy AI

Concerning Chapter III, the EHF does see the merits of the approach taken. It also welcomes the ambition of creating a number of use-case-specific sets of assessment questions.

However, it warns that these lists – as the document itself acknowledges – are by definition incomplete. Here as well, because of AI's pervasiveness and the diversity of its applications, ex ante measures have to be complemented with systematic post hoc procedures including the design and management of feedback systems allowing to flag incidents.

General Comments

The EHF welcomes the work done by the HLEG AI and is looking forward to see the results of this consultation included in the final version. However, we are also worried that despite what the document claims – that it is to become a living document – the final version will be used as a baseline to consider whether a specific AI application is deemed ethical by European standards – whether it is "trustworthy AI, made in Europe."

The EHF also expresses its concerns that the understandable race for reaping the benefits of AI have resulted, within the HLEG AI in a certain lack of methodological rigor. For instance, the claim that AI's benefits largely outweigh its challenges is not proven. The declaration that there is no legal vacuum when it comes to AI seems very hasty. Despite the fact that the document acknowledges that the explicability of AI is by far not guaranteed, it does build some of its reasoning on this concept.

The growing opacity of AI technologies and their extension to multiple domains of life pose less the problem of control over design or use - these have become almost impossible in some instances – than that of social impact and possible recourse in case of problems.

In this sense, the working document is not realistic enough. It sets itself the goal to guarantee "trustworthy AI" without acknowledging the fact that maybe, to some extent, this can only be an aspiration. It focuses

primarily on ex ante measures to minimize the risks of AI – and this is laudable. It however overlooks the importance of systematized feedback about the dysfunctions, threats and risks experienced by users. In disregarding the importance of societal control and validation, it hinders efficient detection of yet unknown threats and potentially undermines societal acceptance of AI, including the acceptance of the inherent risks that their usage might entail.

This is why, to complement the ex ante measures listed, the EHF's main proposal concerns the creation of a European Observatory of AI Technologies and Services in charge of implementing social control at any stage of the design, deployment or use, including post hoc end-user return of incidents.

Furthermore, empowerment of citizens via education, awareness raising on the one hand and massive improvement in user interfaces on the other is fundamental. Unidirectional informed consent is not enough if one wants to help citizens truly understand what parameters, data, inputs and processes influence the outcomes of the AI application they are using.

The EHF will follow with great interest the development of this document as well as the drafting of the HELG AI's other key deliverable concerning regulation.