Evangelische Kirche in Deutschland

Der Bevollmächtigte des Rates

Dienststelle Brüssel **AZ 2.18.14**

Stakeholders Consultation on draft AI Ethics Guidelines

Introduction: Rationale and Foresight of the Guidelines

page i, Executive Summary: The fact that the "draft ethics guidelines for trustworthy Artificial Intelligence" (draft) explicitly name opportunities concerning Artificial Intelligence (AI) by concrete examples while remaining unspecific concerning risks, shows a lack of awareness. The draft should be more explicit in naming risks. In its current form the guidelines still have some shortcomings as they seem quite application-oriented and business-friendly.

page ii, Executive Summary: The draft guidelines openly talk about "using ethics as an inspiration to develop a unique brand of AI". The wording is ambiguous: If it is understood in the way that ethics is supposed to establish a unique brand of AI this would be equivalent to a utilitarian approach instrumentalizing ethics for predefined goals. If the wording indicates that ethics should explicitly be taken into consideration during the development of AI and should contribute to the trustworthiness of AI (integrated AI-research) it is welcomed. An ethical superiority could pass as an advantage with regard to global competition. The given wording, which stresses the aspect of "competitiveness" in the same context as the "ethical approach to AI" gives reason for our concern that the draft guidelines are driven mainly by business interests and lack a balanced approach. We recommend to clarify the phrasing.

page ii, iii, Executive Summary, and page 29, Conclusion: The idea of the draft to establish a kind of European trademark named „Trustworthy AI made in Europe" hinders an open and transparent ethical debate about AI in general. The term and framing of "trustworthy AI" already entail a positive bias towards AI. As a first step a debate about the conditions and procedures how to get trustworthy AI was needed. The draft should put more effort into clarifying that an in-depth reflection is necessary to handle the issue of AI in a responsible way.

page iv, glossary "Ethical purpose": The glossary defines the term "ethical purpose" indicating 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 paragraph is problematic in several ways. Firstly, because an ethical review starts from a more general perspective and is not purpose-bound in the way that ethics might be (mis-) used as a justification/legitimization for a certain kind of research, development or application. Secondly, the ethical concerns with regard to the development, deployment and use might vary significantly and might even lead to certain contradictions. Therefore, the approach trying to lay out an ethical purpose with regard to "development, deployment and use" must by definition remain quite general and vague and could be misleading.

page iv, glossary "Human centric AI": We welcome the human-centric approach to AI. But not only "human values" as mentioned under this heading must be given primary consideration, but human dignity and human rights. This should be clearly mentioned.

Page 2, "A. Rationale and foresight of the Guidelines, Purpose and target Audience of the Guidelines: The draft states that „mechanism will be put in place that enable all stakeholders to formally endorse and sign up to the guidelines on a voluntary basis". On the one hand – given the dynamics in AI it is realistic to assume that the guidelines are "a living document" and work in progress, on the other hand

there is no clarity about the conclusion of the process. For the time being the guidelines are not legally binding. But what should be the final result of the consultations, ethical reflections etc.? Completely new regulations, an update of the current regulatory framework, a code of conduct, non-binding guidelines? Who are supposed to be the addressees? Who should be held accountable and liable?

We want to draw the attention to the risks inherent to an uncoordinated and non-transparent approach in the regulation of AI. A regulatory patchwork may give rise to unclear responsibilities and a lack of accountability leading to a state of bad/ non- governance. The draft should make clear that a ping-pong-effect with regard to accountability and responsibility would not fulfil the requirements of good governance. Moreover, accountability does not have any added value if there is no debate about a regulation of liability at the same time.

page 3, "B. A Framework For Trustworthy AI" The draft addresses „developers, deployers and users to comply with fundamental rights and with all applicable regulations". The draft's declared aim is to build guidelines for trustworthy AI. All confidence building processes require as relational acts clear reference objects. The summary and heterogeneous group of „developers, deployers and users" is not appropriate for that purpose. A more tailored, group-specific approach is needed. In addition, it is not clear how the sheer adherence alone to a regulatory framework which is a general obligation for companies and citizens anyway should especially increase the trust in AI compared to the current state of play.

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

page 7, "3. Fundamental Rights of Human Beings: The draft states: „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".

Nonetheless, this paragraph does not yet reveal anything about scoring in a private business context. From an ethical perspective it is always problematic to reduce a person's identity to digital data. The guidelines should clarify that it cannot make a difference if governments or business companies abuse artificial intelligence for scoring. The cooperation and links between companies and governments concerning these issues should also be addressed.

page 8, Ethical principles in the context of AI and correlating values: We welcome the five mentioned ethical principles and consider them to be of primary importance for an ethical approach to AI. Still there is a need for further clarifications: How do they relate to each other, how should they be balanced against each other when ethical principles are contradicting regarding a technical solution? How does the EU deal with the fact that different actors with different interests might understand ethical principles in a different way? EKD would assume that an ethical discourse based on the rules of fairness and equality needed to be established providing a balanced approach of diverging principles and a common understanding. We would also recommend a human rights impact assemement with regard to algorithms to be established.

Surely, public, private, and civil organizations have drawn inspiration from fundamental rights to produce ethical frameworks for AI. The work of European Group on Ethics in Science and New Technologies (EGE) on AI is named in the draft as one example. Nonetheless, the relation between the high-level expert group´s draft guidelines and the EGE opinion on the matter needs to be clarified in the draft. An explanation is needed in which way the AI HLEG "build on the above work".

page 10, "4. Ethical Principles in the Context of AI and Correlating Values, The Principle of Explicability": The draft explains: „Explicability 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“.

This statement ignores the fact that the consent in AI touches the principle of human dignity, understood as the recognition of the inherent human state of being worthy of respect. A relational conception of human dignity requires that we are aware of whether and when we are interacting with a machine or another human being for example.

We furthermore agree that explicability is a precondition for trustworthy AI. However, the model of „informed consent from individuals“ is not realistic for two reasons and therefore should not be applied in this context: on the one hand, the complexity of AI exceeds the capacity of understanding of most individuals. On the other hand, AI algorithms should be legitimately regarded as business secrets which excludes automatically any public explicability. Explicability should be guaranteed towards public authorities and Treuhandstellen (trusts) which are obliged to keep business secrets, dispose of sufficient expertise and which act in the interest of a user or consumer. The present draft does moreover not sufficiently consider the fact, that convergence of interests is not given in the relation between business companies profiting from AI and affected individual human beings. In our view, this aspect of (information) asymmetry must be reconsidered in the draft. Trust in AI will not emerge if AI developers and enterprises refuse to explain their algorithms towards public authorities or state agencies and rely exclusively on their own, interest driven „explications“ towards users who usually are lay persons in IT.

page 12, Lethal Autonomous Weapon Systems (LAWS): EKD welcomes the fact that regarding lethal autonomous weapon systems (LAWS) the draft clarifies, that “human beings are, and must remain, responsible and accountable for all casualties.” However, the draft claims that on the other hand LAWS could “reduce collateral damage, e.g. saving selectively children”. Such hypothetical assumptions blur the goal to raise awareness about a responsible handling of AI and seem to justify interests of the military industry which should not be part of the draft.

Chapter II: Realising Trustworthy AI

page 19, technical methods: The paragraphs on technical methods to realise “trustworthy AI” remain very vague and not specific enough also given the fact that they should apply to “the design, development and use” which is a very broad spectrum. A more differentiated and diligent approach would be needed to look into this important field.

page 18, technical and non-technical methods: EKD welcome the fact that the draft guideline stress the importance of an evaluation of the requirements and methods employed on “an on-going basis”.
page 22, education and awareness to foster an ethical mind-set: Given the fact that EKD considers education on the impact of AI on society as well as on the individuum being of primary importance we feel the draft guidelines should be more specific on this point clarifying that the education should enable the individuum not only to know how to apply AI but provide a broader orientation.

page 21, non-technial methods: A human rights impact assessment of AI and the applied algorithms should be added to the list**.**

Page 22, stakeholder and social dialogue: EKD welcomes that the AI HLEG sees the need for “an open discussion and an involvement of social partners, stakeholders and general”, but the paragraphs lacks any details about the How of the involvement and remains too vague and unclear with regard to the intended activities. As the involvement of the public is key to establishing a European AI strategy more diligence should be devoted to clarifying this point in EKD´s view also involving theological-ethical expertise.

Chapter III: Assessing Trustworthy AI

General Comments

The present draft constitutes a user-oriented paper addressing developers, deployers and users to comply with fundamental rights and with all applicable regulations. However, it would be necessary to develop a clear concept about the aims to be achieved at the end of the process. It needs to be clarified if there should be a set of regulations with the power to impose sanctions on governments or companies.

Ethics is always connected to proceedings and communication channels. Publishing a draft in the pre-Christmas-period remaining open for primarily a period of only one month does not go along with the high-level expert group's declared aim to build ethical guidelines for trustworthy AI in cooperation with the generally public and to allow for a substantive exchange. Moreover, this approach contradicts the ambition of the European Commission to make "the EU more transparent and accountable" through "consultations that are of a high quality and transparent, reach all relevant stakeholders and target the evidence needed to make sound decisions" as stated in the communication "Better regulation for better results – A European Agenda". The circumstances of this opaque procedure give reason for the suspicion that profound debates about this working document are not really desired by the European Commission.

We welcome the fact that a voice is given to renowned universities in the European's high-level expert group, but the fact that giant internet companies like *google*, *Zalando* or *SAP* are part of the high-level expert group as well as the biased approach to "a unique brand of AI" without labelling AI as "ethically responsible AI" as well as the content of the guidelines being very application- oriented lead to the impression that the draft is driven by business interests. We also deplore that no theological know-how was involved in the AI HLEG despite competent candidates.

Moreover, it should be the aim of such guidelines to avoid the appearance that competition is more important than preservation and – where necessary – enhancement of ethical standards. „Trustworthy AI" can only be successfully established as a brand if the autonomy of the users is strengthened and at the same time an ethically responsible AI is established in correlation with public (governmental or EU) regulation and control in the interest of the users and for the common good . In case this model of AI is refused explicitly by the EU or the enterprises such a concept would even undermine the aim of „trustworthy AI" and hamper the implementation.

AI will lead to in-depth societal changes which will go far beyond business-consumer relations and the questions addressed in these guidelines. The risks mentioned under section 5 are not only a question of how to design AI but examples of challenges our societies will have to cope with and find answers to. AI could change and query the current functioning of our societies, the perception of individuals and fundamental principles like human dignity and fundamental rights. Therefore, the EU should set in motion a broad and fundamental societal debate on opportunities and challenges of AI, the relationship of AI to human beings and the society in general.