Telefónica's Response
Stakeholders' Consultation on Draft AI Ethics Guidelines

(January 2019)

General Comments

Overall the Guidelines provide a good and thorough approach to ensure that AI will have much more good use than bad (intentional or unintentional) use. We agree with that, evidenced by the fact that, in October 2018, Telefonica has voluntarily published its Company AI principles to foster a trustworthy environment for our stakeholders regarding how we will develop and use AI, and by Telefónica's more general Business principles where transparency and acting in accordance with non-negotiable ethical standards are two core principles.

In this "general comments" part, we distinguish our feedback in two parts: 1) key concerns on the objectives and scope of the draft guidelines, and 2) general feedback on the content of the draft guidelines.

1) Key concerns on the objectives and scope of the Guidelines

- We think it is important to **explicitly state** that the stakeholders invited to voluntary endorse the Guidelines should **not only include European organizations**, but all organizations that serve EU citizens, businesses and governments, **wherever in the world they are based.**

- It is **difficult to assess** how these Guidelines will **increase the competitiveness of Europe** without having the opportunity to analyze how the AI Policy & Investment Recommendations promotes the development of AI capabilities, which is key for understanding how Europe plans to catch up with other regions around the world.

- In the same line, we are concerned about these voluntary **Guidelines turning into Regulation**, especially if that new Regulation would **only apply to, or be enforced on, European businesses**, and not to businesses in other regions of the world that are serving European customers. Since most of European citizens' personal data -a fundamental pillar for the development and improvement of AI- is controlled by non-EU businesses, having a framework imposing safeguards on AI that does only apply to EU based businesses will not benefit EU citizens nor EU business competitiveness.

- If these Guidelines are becoming regulation to follow the "GDPR" model approach, it would be wise to first assess what is the impact of the application of GDPR on **EU companies vs. non-EU companies** and their respective competitiveness. This will provide critical learning for any possible future regulation on AI.

2) Other general comments on the Guidelines:

- **Monopolization of data** is a critical, immediate threat to the development and implementation of AI and could be **an ethical question** in and of itself. Since access to data, and more relevantly, behavioral data, is a requirement for the development of AI, and mostly all this data is going to be in the hands of a few companies, AI will be controlled by such companies. Indeed, the ability to track the behavior of billions of users worldwide, through the provision of multiple

conglomerate services, is possible just and only for an extremely short list of global digital players; only they will be able to gather such an extensive and diverse amount of data, indispensable for the training of AI algorithms. No one will be able to compete as those **few players will dominate the end-to-end ecosystem** and decide how AI evolves.
This an ethical concern that should be raised in the guidelines, for example within the assessment list in Requirements of Trustworthy AI in Section II: the requirement of Respect for Human Autonomy refers to protecting citizens in all their diversity from private abuses made possible by AI technology, ensuring a fair distribution of the benefits created by AI technologies. Certainly, **in a monopolized data market** where a few companies control the development of AI, and no alternatives are feasible, **citizens could be subject to abuses from such companies**.
Unless abuses in the data gathering by the biggest digital players are avoided, allowing the emergence of alternative AI providers, benefits would not be fairly distributed among users and providers, neither geographies. At the end of this section the following paragraph could be added "Access to data and behavioral data is critical for the development and implementation of AI, and thus its monopolization in the hands of very few companies could limit the emergence of AI solutions from other alternative players, thus enabling potential abuses on citizens and an uneven distribution of AI benefits; situations where critical data for the development of AI is being monopolized, should be avoided".

- It would be interesting to know what **percentage of current AI applications complies and doesn't comply with the Guidelines** and give a few visible examples of each of those. This would also serve as a test case of how pragmatic and feasible the approach is.

- Due to the relevance of data governance and the business model for trust in AI, we would suggest to highlight this also in the Executive Summary, and not only in the Rationale and Foresight section: "**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, the data governance and the business model".

- A minor terminological issue: instead of speaking about a "rule-based AI system", it is better to call it "**symbolic AI system**" or "**knowledge-based AI system**". Rule-based has a specific connotation in the AI world (if-then-else rules), but there are many other approaches (not falling under "learning-based") that reason with explicit knowledge without rules.

Introduction: Rationale and Foresight of the Guidelines

- "Trust in the business model" is identified in the introduction as one of the three pillars for a trustworthy AI, but only referenced again in the **Principle of Explicability** (Transparency). Business models enabled by AI technology should not pursue an unethical purpose and thus be included within the Transparency requirement to realize a trustworthy AI (Section II) and also reflected by example questions in the Assessment List both on Transparency and Fairness (Section III).

- It is much welcomed that the Scope of the Guidelines acknowledges that different situations raise different challenges by referring to concrete examples of AI systems: recommendation of songs and of critical medical treatment. Along these lines, it should also be acknowledged that based on such different challenges raised by different situations, **guidelines and related obligations should be graded**, or applied with different intensity according to the impact a specific AI based system has throughout all the levels of AI system life cycle (development, deployment and usage). The same way cybersecurity requirements are different for a domestic watering IoT device vs nationwide energy grid, AI principles should apply differently depending

on its impact in order not to inhibit innovation of simpler, lower impact AI based systems. Therefore, it should be emphasized that requirements for trustworthy AI (Section II.1) and, even more so, the technical and non-technical methods to achieve trustworthy AI (Section II.2) should be domain and application specific.

- What is the evidence that fostering **a human-centric approach to European AI** will enable Europe to become a globally leading innovator in ethical, secure and cutting-edge AI? Current leading institutes for ethical AI are non-EU based, e.g. in NY (AI Now), the Singapore government has already created an AI ethics commission, and the UK is also setting up initiatives (e.g. Ada Lovelace institute). In fact**, Europe should act, as other countries/regions could put in practice an ethical AI approach much earlier**. Europe's advantage lays in its ability to drive a collective agreement and act as a block. This will be valuable in the future as the world grapples with the cross-border challenges of AI.

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

- On the **principle of Beneficence**, "Do good". We do not discuss the importance of doing good as a great ethical principle. What we challenge is the opportunity and applicability to AI (or to any other technology). Even more, in principle, rules (and these are rules) should not demand others (or the technology) **to do good but to do no harm**. Boldly applying the "Do Good" principle would restrict companies' freedom to innovate in or perform regular businesses to the extent that their primary objective might not be improving collective wellbeing. A relevant case among many would be the use of AI in advertising, which some might argue is not aligned with the beneficence principle.
  This would limit Europe's opportunity of learning to use AI through marketing, particularly advertising, while this activity is a low risk / highly profitable form of AI (as compared to other AI based decisions with greater societal impact and thus risk, such as a healthcare decision) which could in turn enable the funding of more AI research in the EU. Considering Europe is lagging behind other regions in the use of data for marketing purposes, restricting the use of AI for marketing purposes based on the "do good" principle would have the opposite desired effect of this guidelines. **The "do good" principle should be modified in order to provide room for these activities** (such as advertising).

- The "**Do not harm principle**" states negative profiling should be avoided. While it is already clear that the do not harm principle enshrines eliminating all negative actions, "profiling" is neutral from a normative perspective, and what makes it harmful is the purpose of the profiling, which relates to the business or public governance model. In fact, profiling is widely used and needed for whatever commercial activity. At the end of the day, we are profiled countless times by every digital interaction we have, with or without use of AI. Thus, unless it is clearly explained what is to be interpreted as "negative profiling", **we would ask the deletion of the reference to profiling in the "do not harm" principle.**

- **On "informed consent" as a value.** The Guidelines consider informed consent as **an ethical value**, which puts in practice the fundamental right of human dignity (sic) and makes a direct link with explicability. This does not take into account that even the GDPR considers five legal bases for processing other than consent which of course do not negatively impact human dignity. Additionally, transparency (explicability) should not be inextricably linked to consent but applied independently of the concrete ground for the processing. As such, we would recommend including a different example on **how to go from fundamental right to principles and values.**

- It is questionable that Covert AI systems by themselves represent a critical concern; it will depend upon the function the system provides. If for example AI system is used for speech recognition for a more advanced IVR system, it is not a really a concern. In fact, IVR does not announce itself as a machine and we don't question this now, should it? Does it need to since it is not AI? But is a synthetic voice AI? In this case, this is more a transparency related issue than a critical concern.

- Providing examples on Potential longer-term concerns, such as Artificial Consciousness, given that there is no accepted or consensus theory around the topic, just serves to increase unfounded concerns. As **the aim of the paper is to provide the foundation for trustworthy AI**, giving such futuristic science fiction like vision on AI just serves the opposite objective, raising alarms without providing any solution or mitigating effect. Guidelines should refrain from providing speculative views on AI. Though it is a natural research goal to work on AGI, whether AGI is possible (and how long it will take to reach it) or not is an opinion that can be argued against or in favor of, but currently the answer is unknown.

- Another long-term concern to be included is **the monopolization of our attention based on AI technologies**, with the unintended consequence of people getting addicted to some digital services. This may also result in the promotion of certain content based on AI-driven decisions which ends up in the reduction of quality of content being replaced with fake news and junk content. But this is more related to the business model of the service than with AI technology per se. Therefore, it is important to consider the business model in the trustworthiness assessment.

Chapter II: Realising Trustworthy AI

- **Data Governance** should add **labelling of data as a best practice** in order to assure accountability, explainability and improvement of AI training and validations tests, and thus be also able to assess the quality of the data itself.

- **Data Governance** is a known term in the area of Big Data and has a broader meaning than the intended in this section. We suggest changing to **Data requisites**. One of the main aspects mentioned here is about bias in data sets, and the importance of being aware of this and correcting it. The section should include a reference to **ethics around data gathering or the use of AI to coerce data collection**

- **Accountability Governance** as a form of non-technical method should include a reference to **auto-regulation**, **self-regulation** and the procedures through which the Governance framework assesses compliance.

- The three dimensions for trust in AI (**technology**, **data governance** and **business model**) should be the first bullet in the summary box KEY GUIDANCE FOR REALISING TRUSTWORTHY AI.

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

- **Monopolization of data is a critical,** immediate threat to the development and implementation of AI and could be an ethical question in and of itself. Since access to data, and more relevantly behavioral data, is a requirement for the development of AI, and mostly all these data are going

to be in the hands of a few companies, AI will be controlled by such companies. Indeed, the ability to track the behavior of billions of users worldwide, through the provision of multiple conglomerate services, is possible just and only for an extremely short list of global digital players; only they will be able to gather such an extensive and diverse amount of data, indispensable for the training of AI algorithms. No one will be able to compete as they will dominate the end-to-end ecosystem decide how AI evolves.

This is an ethical concern that should be raised within the assessment list in Requirements of Trustworthy AI: the requirement of Respect for Human Autonomy refers to protecting citizens in all their diversity from private abuses made possible by AI technology, ensuring a fair distribution of the benefits created by AI technologies. Certainly, in a monopolized data market where a few companies control the development of AI, and no alternatives are allowed, citizens could be subject to abuses from such companies.

Unless abuses in the data gathering by the biggest digital players are avoided, allowing the emergence of alternative AI providers, benefits would not be fairly distributed among users and providers, neither geographies. At the end of this section the following paragraph could be added "Access to **data and behavioral data is critical for the development and implementation of AI**, and thus its monopolization in the hands of a very few companies could limit the emergence of AI solutions from other alternative players, thus enabling potential abuses on citizens and uneven distribution of AI benefits; situations critical data for the development of AI being monopolized should be avoided".

- Architectures for Trustworthy AI do not seem to include the entire lifecycle of AI—including how data was collected and the business model of the whole system.

- We recommend the addition of a reference to technical tools with the following capabilities:

  - **Detection of correlations** between sensitive variables and normal (apparently harmless) variables
  - **Detection of bias in** data sets (IBM, Accenture, Pymetrics, ...)
  - **Correction of bias**: in the data set, and through the algorithm (using GANs, https://blog.godatadriven.com/fairness-in-ml)
  - **Checking of the risk of re-identification of anonymized** data
  - **Visualization of the impact of false positives and false negatives** on a certain domain

Chapter III: Assessing Trustworthy AI

Recommend the addition of the following question within the assessment list:

- **Respect for privacy**:
  - Are end users informed of which data sets are being treated, for which purposes and which is the expected output of those treatments?