QUBIT PRIVACY BRIEFINGS

Qubit Privacy

BRIEFING 2 - APRIL 2021

ETHICS IN ARTIFICIAL INTELLIGENCE

On 8 April 2019, the High-Level Expert Group on AI (AI HLEG) presented Ethics Guidelines for Trustworthy Artificial Intelligence. The guidelines put forward a human-centric approach on AI and list 7 key requirements that AI systems should meet in order to be trustworthy.
This document summarises the main points of the guidelines

INTRODUCTION

AI has the potential to significantly transform society and it is a means to increase human flourishing. AI systems (AIS) need to be human-centric, and developers should seek to maximise the benefits of AI solutions while preventing and minimising their risks.

TRUSTWORTHY AI

Trustworthiness is a prerequisite for people and societies to develop, deploy and use AIS. Trustworthy AI has three components: lawful, ethical and robust AI

- *Lawful AI:* AI must comply with all applicable laws and regulations. In the EU it means complying with EU primary law (TEU, TFEU, CFR) secondary law (GDPR, Product Liability Directive, etc), human rights conventions (ECHR), and EU MS laws.
- *Ethical AI*: AI must ensure adherence to ethical principles and values.
- *Robust AI*: AIS must contain safeguards to prevent any unintended adverse effects or harms. This applies both from a technical perspective (ensuring the system's technical robustness as appropriate in a given context, such as the application domain or life cycle phase) and from a social perspective (considering the context and environment in which the system operates)

QUBIT PRIVACY BRIEFINGS

Qubit Privacy

BRIEFING 2 - APRIL 2021

FOUNDATIONS OF TRUSTWORTHY AI

AI ethics focuses on the ethical issues raised by the development and deployment of AI. It tries to identify how AI can advance or raise concerns to the good life of individuals, whether in terms of quality of life, or human autonomy and freedom necessary for a democratic society.

Fundamental rights as moral and legal entitlements

AI ethics based on the fundamental rights enshrined in the EU treaties, CFR and international human rights law. Respect for fundamental rights, within a framework of democracy and the rule of law, provides the most promising foundations for identifying abstract ethical principles and values, which can be operationalised in the context of AI.

Fundamental rights as a basis of trustworthy AI

Many fundamental rights are enforceable under EU law:

- *Respect for human dignity*: every human being possesses an 'intrinsic worth', which should never be compromised by others. AIS should not treat human beings as objects to be sifted, scored, conditioned or manipulated
- *Freedom of the individual*: human beings should remain free to make life decisions for themselves;
- *Respect for democracy, justice and the rule of law*: AIS should serve to maintain and foster democratic processes and respect the plurality of values and life choices;
- *Equality, non-discrimination and solidarity*: AIS cannot generate unfairly biased outcomes
- *Citizens' rights*: AIS have the potential to improve the scale and efficiency of government in the provision of public goods, but also to negatively impact them

QUBIT PRIVACY BRIEFINGS

BRIEFING 2 - APRIL 2021

Qubit Privacy

ETHICAL PRINCIPLES IN THE CONTEXT OF AI SYSTEMS

4 ethical principles must be respect to ensure that the AI systems are developed and deployed in a trustworthy manner.

- *Respect for human autonomy*: AIS should not unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans, and they should leave meaningful opportunity for human choice
- *Prevention of harm*: AIS should neither cause not exacerbate harm or otherwise affect human beings. AIS must be safe and secure, technically robust and ensure that they are not open to malicious use.
- *Fairness*: the principle of fairness has two dimensions
  - *Substantive dimension*: ensuring equal and just distribution of benefits and costs and ensuring that individuals and groups are free from unfair bias. AIS should not create or reproduce unfair biases
  - *Procedural dimension*: ability to contest and seek effective redress against decisions made by AIS
- *Explicability*: processes need to be transparent, the capabilities and purposes of AIS openly communicated and decisions explainable to those directly and indirectly affected by them. Without such information a decision cannot be duly contested.

Tensions may arise between the principles: e.g. principle of prevention of harm and the principle of human autonomy in predicting policing. But certain fundamental rights and principles are absolute and cannot be balanced (e.g. human dignity)

QUBIT PRIVACY BRIEFINGS

BRIEFING 2 - APRIL 2021

Qubit Privacy

REALISATION OF TRUSTWORTHY AI

For the realisation of AI, all 7 key requirements must be implemented: *i) human agency and oversight; ii) technical robustness and safety; iii) privacy and data governance; iv) transparency; v) diversity, non-discrimination, and fairness; vi) societal and environmental wellbeing; and vii) accountability*

Human agency and oversight

AIS should support human autonomy and decision-making, as prescribed in the principle of respect for human autonomy. AIS should support:

- *Fundamental rights*: AIS can enable and hamper fundamental rights. In situations where such a risk exists, a fundamental rights impact assessment is advised.
- *Human agency*: users should be able to make informed autonomous decisions regarding AIS. They should be given the knowledge and tools to understand and interact with AIS. Key to this is the right not to be subject to a decision based solely on automated processing when this produces legal effects on users or similarly significantly affects them.
- *Human oversight*: it helps ensuring that an AI system does not undermine human autonomy or causes adverse effects. Oversight may be achieved by governance mechanisms like human-in-the-loop, human-on-the-loop, or human-in-command.

Technical robustness and safety

AIS should be developed with a preventive approach to risks and in a manner such that they reliably behave as intended while minimising unintentional and unexpected harm. This is linked to the principle of prevention of harm. Main features:

- *Resilience to attack and security*: AIS should be protected against vulnerabilities

- *Fallback plan and general safety*: AIS should have safeguards that enable a fallback plan in case of problems (e.g. from a statistical to a rule based procedure). The level of safety depends on the magnitude of the risks
- *Accuracy*: it concerns the ability to make correct judgments, predictions, recommendations. The AIS should indicate how likely errors are
- *Reliability and reproducibility*: AIS should work properly with a range of inputs an in a range of situations (reliability) and show the same behaviour when repeated under the same conditions (reproducibility).

Privacy and data governance

Linked to the principle of prevention of harm.

- *Privacy and data protection*: AIS must guarantee privacy and data protection throughout a system's entire lifecycle
- *Quality and integrity of data*: AIS must ensure a high quality of data, since data can contain socially constructed biases, inaccuracies and errors. Also, the integrity of data must be ensured since feeding malicious data may change the system behaviour
- *Access to data*: procedures to ensure the access to the data must be put in place

Transparency

Linked to the principle of explicability

- *Traceability*: the datasets and the processes that yield the AIS's decision, and decision itself, should be documented
- *Explainability*: the ability to explain the technical processes of the AIS and the related human decisions, and any trade-offs between explainability and accuracy made by the system, in a clear and easily understandable manner
- *Communication*: individuals have the right to know that they are interacting with an AIS

QUBIT PRIVACY BRIEFINGS

BRIEFING 2 - APRIL 2021

Qubit Privacy

Diversity, non-discrimination and fairness

Linked to the principle of fairness

- *Avoidance of unfair bias*: Datasets used by AIS may suffer from the inclusion of inadvertent historic bias, incompleteness and bad governance models. The continuation of such biases could lead to unintended direct or indirect prejudice and discrimination
- *Accessibility and universal design*: AIS should be user-centric and designed in a way that allows all people to use AI products regardless of their age, gender, abilities, or features
- *Stakeholder participation*: in the development it is advisable to consult stakeholders who may directly or indirectly be affected.

Societal and environmental wellbeing

Linked to the principles of fairness and prevention of harm

- *Sustainable and environmentally friendly AI*: AIS promise to help tackling some of the most pressing societal concerns, yet it must be ensured that this occurs in the most environmentally friendly possible way
- *Social impact*: ubiquitous exposure to social AIS may alter our conceptions of social agency, or impact our social relationships and attachment
- *Society and democracy*: the impact of AIS should be assessed from a societal perspective, considering its effects on institutions, democracy and society at large

Accountability

Linked to the principle of fairness and it ensures responsibility and accountability for AIS and their outcomes

- Auditability: assessment of algorithms, data and design processes, which does not necessarily imply revealing information about business models or IPR

- *Minimisation and reporting of negative impacts*: the ability to report on actions or decisions that contribute to a certain system outcome and to respond to the consequences of such an outcome must be ensured
- *Trade-offs*: implementing these requirements will lead to tensions and many times a trade-off will be needed. Where no ethical acceptable trade-offs can be identified, the development and deployment of the AIS should not proceed in that form
- *Redress*: Where unjust adverse impact occurs, accessible mechanisms should foresee that ensure adequate redress

This briefing was prepared by Federico Marengo for QUBIT PRIVACY

QUBIT PRIVACY is a consultancy firm established in Italy that provides tailor-made services for individuals and companies to comply with the requirements established in the General Data Protection Regulation.
Federico Marengo is a lawyer, master in public administration (University of Buenos Aires), LLM (University of Manchester), and PhD candidate (Università Bocconi, Milano).
He currently provides data protection consultancy services for Qubit Privacy and also works as of counsel for consultancy firms.
He is the author of "Data Protection Law in Charts. A Visual Guide to the General Data Protection Regulation", e-book released in 2021, and authored several publications on international data transfers and international trade law.
As a PhD researcher, his research deals with the potential and challenges of the General Data Protection Regulation to protect data subjects against the adverse effects of Artificial Intelligence.
He is also teaching assistant at Università Bocconi.