THE
NATIONAL LAW REVIEW

# European Commission’s Ethics Guidelines on Artificial Intelligence

DrinkerBiddle
Article By
Kenneth K. Dort
Alice D. Czyzycki
Drinker Biddle & Reath LLP
Insights

- Global
- Communications, Media & Internet

- European Union

Tuesday, April 16, 2019

“Artificial intelligence” can be defined as the theory and development of computer systems able to perform tasks that normally require human intervention. Artificial intelligence (AI) is being used in new products and services across numerous industries and for a variety of policy-related purposes, raising questions about the resulting legal implications, including its effect on individual privacy. Aspects of AI related to privacy concerns are the ability of systems to make decisions and to learn by adjusting their code in response to inputs received over time, using large volumes of data. (See Artificial intelligence and Privacy Report, January 2018, p. 6)

Following the European Commission’s declaration on AI in April 2018, its High-Level Expert Group on Artificial Intelligence (AI HLEG) published Draft Ethics Guidelines for Trustworthy AI in December 2018. A consultation process regarding this working document concluded on February 1, 2019, and a revised draft of the document based on the comments that were received is expected to be delivered to the European Commission in April 2019.

The Guidelines call for the development of “Trustworthy AI,” which features a “human-centric” approach that emphasizes the goal of increasing human well-being, rather than the development and use of artificial intelligence as a means in itself. The standard for Trustworthy AI set forth by the Guidelines has two components: (1) respect for fundamental rights, applicable regulations, and core principles and values, ensuring an “ethical purpose” and (ii) technical robustness and reliability, to avoid unintentional harm caused by a lack of technological mastery.

The goal of the Guidelines is not only to provide a list of core values and principles but also to offer concrete guidance on the implementation and operationalization of the values and principles applicable to AI systems. However, the AI HLEG makes it clear that the Guidelines are not an official document from the European Commission with legally binding effect, and are not intended to be an end point but rather the beginning of an open-ended debate.

The Guidelines set forth the following three-chapter framework for achieving Trustworthy AI.

## Chapter I: Key Guidance for Ensuring Ethical Purpose

The principles and values identified in the Guidelines are based on the fundamental rights commitment of the EU Treaties and Charter of Fundamental Rights. These rights (dignity, freedoms, equality and solidarity, citizens' rights and justice) represent a human-centric approach, where humans have primacy in the civil, political, economic and social fields. The AI HLEG believes that a rights-based approach will bring an additional benefit of limiting regulatory uncertainty, thereby building on decades of application/development of fundamental rights in the European Union.

The Guidelines set forth the following principles and values in the context of artificial intelligence:

- Beneficence: "Do Good" – the systems should be designed and developed to improve individual and collective well-being by generating prosperity, value creation and wealth maximization and sustainability, seeking achievement of a fair, inclusive and peaceful society, and helping to increase citizens' mental autonomy with equal distribution of economic, social and political opportunity.

- Non-Maleficence: "Do No Harm" – the systems should not harm human beings, whether through physical, psychological, financial or social harm, or harm to the environment. As artificial intelligence—specific harms may stem from the treatment of individuals' data, the collection and use of data for training the systems should be done in a way that avoids discrimination, manipulation or negative profiling, with greater attention to the data of vulnerable demographics. Further, the systems should be developed and implemented in a way that protects societies from ideological polarization and algorithmic determination.

- Autonomy: "Preserve Human Agency" – human beings should have freedom from subordination to, or coercion by, the systems. This freedom includes a right to decide whether or not to be subject the artificial intelligence decision-making, a right to knowledge of interaction with artificial intelligence systems, a right to opt out and a right of withdrawal.

- Justice: "Be Fair" – the systems should maintain freedom from bias, stigmatization and discrimination for individuals and minority groups. Positives and negatives resulting from artificial intelligence should be evenly distributed, and artificial intelligence systems must provide effective redress if harm occurs or if practices are no longer aligned with the human beings' preferences. Those developing or implementing artificial intelligence systems should be held to high standards of accountability.

- Explicability: “Operate Transparently” – to build and maintain the citizens’ trust, the systems should be auditable, comprehensible and intelligible by human beings at varying levels of expertise, and human beings should be knowingly informed of the intentions of the developers and implementers of artificial intelligence systems. This principle requires informed consent from individuals interacting with artificial intelligence systems, and that accountability measures be put in place.

This chapter goes on to discuss critical concerns raised by artificial intelligence, although the AI HLEG acknowledged that it did not reach consensus on such concerns and asked for specific input in this area during the consultation process. The current Guidelines include the following critical concerns:

- Identification without consent
- Covert artificial intelligence systems
- Normative and mass citizen-scoring without consent in deviation of Fundamental Rights (noted above)
- Lethal autonomous weapon systems
- Potential longer-term concerns (such as artificial consciousness, artificial moral agents or unsupervised recursively self-improving artificial general intelligence).

## Chapter II: Key Guidance for Realizing Trustworthy AI

The Guidelines set forth the following non-exhaustive list of concrete requirements in the context of artificial intelligence:

- Accountability – the systems should include accountability mechanisms, such as monetary compensation, faultfinding or reconciliation without monetary compensation.
- Data Governance – the datasets gathered to train the systems must include high-quality data, with inevitable biases trimmed away prior to engaging in training or by conducting training that requires symmetric behavior over known issues in the training dataset. The data must be properly divided between training sets and validation sets to achieve a realistic picture of the performance of the artificial intelligence system, and the integrity of the data gathering must be ensured such as by keeping a record of the data that is used. The data that is gathered must not be used against the individuals who provided the data.
- Design for All – systems should allow all citizens, particularly those with disabilities, to use the products or services; systems should be user-centric and consider the full range of human abilities, skills and requirements.

- Governance of AI Autonomy (Human Oversight) – the systems must continue to behave as expected in areas such as safety, accuracy, adaptability, privacy, explicability, compliance with the rule of law and ethical conformity. The greater the degree of autonomy given to an artificial intelligence system, the more extensive testing and stricter governance is required. A user of an artificial intelligence system should have the ability to deviate from a path or decision chosen or recommended by the system.

- Non-Discrimination – the systems should not exploit the differences in characteristics between individuals or groups to vary results to negatively impact such individuals or groups, whether intentionally or unintentionally. Harm also may result from exploitation of consumer biases or unfair competition, such as homogenization of prices by means of collusion or non-transparent markets.

- Respect for (and Enhancement of) Human Autonomy – the systems should be designed to uphold rights, values and principles; protect citizens from governmental and private abuses made possible by artificial intelligence technology; ensure a fair distribution of the benefits created by artificial intelligence technologies; protect and enhance a plurality of human values; and enhance self-determination and autonomy of individuals and communities. Systems that are designed to help the user must provide explicit support to promote the user's preferences but set limits for system intervention, ensuring that the well-being of the user as explicitly defined by the user is central to system functionality.

- Respect for Privacy – privacy and data protection must be guaranteed at all stages of the life cycle of a system. This includes all data provided by the user and all information generated about the user through interactions with the system. Organizations must ensure full compliance with the GDPR as well as other applicable privacy and data protection regulations.

- Robustness – the systems' algorithms should be secure, reliable and robust enough to deal with errors and inconsistencies during the design, development, execution, deployment and use phases of the artificial intelligence system, and to adequately cope with erroneous outcomes. This requirement includes the sub-requirements of reliability and reproducibility; accuracy, resilience to attack and redundancy (having a fallback plan).

- Safety − the system should do what it is actually intended to do without harming human physical integrity, resources or the environment. This requirement includes minimizing unintended consequences and errors in operation, putting in place processes to clarify and assess potential risks associated with use of artificial intelligence systems, and incorporating formal mechanisms to measure and guide the adaptability of such systems.

- Transparency – the systems should have the capability to describe, inspect and reproduce the mechanisms through which such systems make decisions and learn to adapt to their environments, as well as the provenance and dynamics of the data used and created by the systems.

This chapter goes on to discuss the following technical and non-technical methods to achieve Trustworthy AI:

- Ethics and rule of law by design
- Architectures for Trustworthy AI
- Testing and validating
- Traceability and auditability
- Explanation
- Regulation
- Standardization
- Accountability governance
- Codes of conduct
- Education and awareness to foster an ethical mindset
- Stakeholder and social dialogue
- Diverse and inclusive design teams

## Chapter III: Key Guidance for Assessing Trustworthy AI

The AI HLEG proposes the use of an Assessment List to assess how artificial intelligence systems can meet the requirements of Trustworthy AI set forth above, for use by developers, deployers and innovators. The AI HLEG again asked for specific input in this area during the consultation process, and plans to include use cases in the next iteration of the document. The current Guidelines include relevant questions related to each of the requirements listed in Chapter II, above, as a preliminary, non-exhaustive Assessment List.

Although the Guidelines are not legally binding at this time, like other documents promulgated by the European Commission we expect the framework proposed by the AI HLEG to either be or become the foundation for an important and widely accepted standard in the development, use and governance of artificial intelligence going forward.

**Source URL:** https://www.natlawreview.com/article/european-commission-s-ethics-guidelines-artificial-intelligence