Universal Guidelines for Artificial Intelligence

Explanatory Memorandum and References

October 2018

Context

The Universal Guidelines on Artificial Intelligence (UGAI) call attention to the growing challenges of intelligent computational systems and proposes concrete recommendations that can improve and inform their design. At its core, the purpose of the UGAI is to promote transparency and accountability for these systems and to ensure that people retain control over the systems they create. Not all systems fall within the scope of these Guidelines. Our concern is with those systems that impact the rights of people. Above all else, these systems should do no harm.

The declaration is timely. Governments around the word are developing policy proposals and institutions, both public and private, are supporting research and development of "AI." Invariably, there will be an enormous impact on the public, regardless of their participation in the design and development of these systems. And so, the UGAI reflects a public perspective on these challenges.

The UGAI were announced at the 2018 International Data Protection and Privacy Commissioners Conference, among the most significant meetings of technology leaders and data protection experts in history.

The UGAI builds on prior work by scientific societies, think tanks, NGOs, and international organizations. The UGAI incorporates elements of human rights doctrine, data protection law, and ethical guidelines. The Guidelines include several well-established principles for AI governance, and put forward new principles not previously found in similar policy frameworks.

Terminology

The term "Artificial Intelligence" is both broad and imprecise. It includes aspects of machine learning, rule-based decision-making, and other computational techniques. There are also disputes regarding whether Artificial Intelligence is possible. The UGAI simply acknowledges that this term, in common use, covers a wide range of related issues and adopts the term to engage the current debate. There is no attempt here to define its boundaries, other than to assume that AI requires some degree of automated decision-making. The term "Guidelines" follows the practice of policy frameworks that speak primarily to governments and private companies.

The UGAI speaks to the obligations of “institutions” and the rights of “individuals.” This follows from the articulation of fair information practices in the data protection field. The UGAI takes the protection of the individual as a fundamental goal. Institutions, public and private, are understood to be those entities that develop and deploy AI systems. The term “institution” was chosen rather than the more familiar “organization” to underscore the permanent, ongoing nature of the obligations set out in the Guidelines. There is one principle that is addressed to “national governments.” The reason for this is discussed below.

Application

These Guidelines should be incorporated into ethical standards, adopted in national law and international agreements, and built into the design of systems.

The Principles

The elements of the Transparency Principle can be found in several modern privacy laws, including the US Privacy Act, the EU Data Protection Directive, the GDPR, and the Council of Europe Convention 108. The aim of this principle is to enable independent accountability for automated decisions, with a primary emphasis on the right of the individual to know the basis of an adverse determination. In practical terms, it may not be possible for an individual to interpret the basis of a particular decision, but this does not obviate the need to ensure that such an explanation is possible.

The Right to a Human Determination reaffirms that individuals and not machines are responsible for automated decision-making. In many instances, such as the operation of an autonomous vehicle, it would not be possible or practical to insert a human decision prior to an automated decision. But the aim remains to ensure accountability. Thus where an automated system fails, this principle should be understood as a requirement that a human assessment of the outcome be made.

Identification Obligation. This principle seeks to address the identification asymmetry that arises in the interaction between individuals and AI systems. An AI system typically knows a great deal about an individual; the individual may not even know the operator of the AI system. The Identification Obligation establishes the foundation of AI accountability which is to make clear the identity of an AI system and the institution responsible.

The Fairness Obligation recognizes that all automated systems make decisions that reflect bias and discrimination, but such decisions should not be normatively unfair. There is no simple answer to the question as to what is unfair or impermissible. The evaluation often depends on context. But the Fairness Obligation makes clear that an assessment of objective outcomes alone is not sufficient to evaluate an AI system. Normative consequences must be assessed, including those that preexist or may be amplified by an AI system.

The Assessment and Accountability Obligation speaks to the obligation to assess an AI system prior to and during deployment. Regarding assessment, it should be understood that a central purpose of this obligation is to determine whether an AI system should be established. If an assessment reveals substantial risks, such as those suggested by principles concerning Public Safety and Cybersecurity, then the project should not move forward.

The Accuracy, Reliability, and Validity Obligations set out key responsibilities associated with the outcome of automated decisions. The terms are intended to be interpreted both independently and jointly.

The Data Quality Principle follows from the preceding obligation.

The Public Safety Obligation recognizes that AI systems control devices in the physical world. For this reason, institutions must both assess risks and take precautionary measures as appropriate.

The Cybersecurity Obligation follows from the Public Safety Obligation and underscores the risk that even well-designed systems may be the target of hostile actors. Those who develop and deploy AI systems must take these risks into account.

The Prohibition on Secret Profiling follows from the earlier Identification Obligation. The aim is to avoid the information asymmetry that arises increasingly with AI systems and to ensure the possibility of independent accountability.

The Prohibition on Unitary Scoring speaks directly to the risk of a single, multi-purpose number assigned by a government to an individual. In data protection law, universal identifiers that enable the profiling of individuals across are disfavored. These identifiers are often regulated and in some instances prohibited. The concern with universal scoring, described here as “unitary scoring,” is even greater. A unitary score reflects not only a unitary profile but also a predetermined outcome across multiple domains of human activity. There is some risk that unitary scores will also emerge in the private sector. Conceivably, such systems could be subject to market competition and government regulations. But there is not even the possibility of counterbalance with unitary scores assigned by government, and therefore they should be prohibited.

The Termination Obligation is the ultimate statement of accountability for an AI system. The obligation presumes that systems must remain within human control. If that is no longer possible, the system should be terminated.

REFERENCES

Asilomar AI Principles (2017)

Aspen Institute Roundtable on Artificial Intelligence (2016)

Association for Computing Machinery, U.S. Public Policy Counsel, Statement on Algorithmic Transparency and Accountability (Jan. 2017)

Council of Europe, Convention 108 (1981)

Council of Europe and Artificial Intelligence (2018)

Data and Society, Governing Artificial Intelligence (2018)

European Commission, High Level Expert Group on Artificial Intelligence (2018)

EU General Data Protection Regulation (2018)

IEEE, Ethically Aligned Design (2016)

Japan, Ministry of Internal Affairs and Communications, AI R&D Guidelines (2016)

Garry Kasparov, Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins (2017)

Madrid Privacy Declaration (2009)

OECD, Artificial Intelligence (2018)

OECD, Privacy Guidelines (1980)

Cathy O’Neil, Weapons of Math Destruction (2016)

Frank Pasquale, The Black Box Society: The Secret Algorithms That Control Money and Information (2015)

Privacy International, Artificial Intelligence (2018)

US Privacy Act (1974)

Toronto Declaration (2018)

Joseph Weizenbaum, Computer Power and Human Reason (1976)

Universal Declaration of Human Rights (1948)