THE DIGITAL
CONSTITUTIONALIST

#4 TRUSTWORTHY AI THROUGH REGULATION? SKETCHING THE EUROPEAN APPROACH

Post Views: 3376

Reading Time: 7 minutes

In this #4 post of the Symposium 'Hitchhikers Guide to Law & Tech', ***Nathalie Smuha*** and ***Anna Morandini*** continue asking fundamental questions on the interaction between regulation and technology. Can the European AI Act mitigate the ethical and legal concerns raised by this hyped technology? Which trail is the EU blazing to secure 'Trustworthy Artificial Intelligence' in Europe, as distinct from the laissez-faire approach in the US and the state-centric approach in China? In this post, both authors unpack the proposed AI regulation and evaluate its merits and pitfalls. After explaining the build-up towards the proposal, they set out the scope of the Act and its four categories of risks as part of a 'risk-based approach' to regulate AI. While the transformation from ethics guidelines to ambitious legal framework is an important milestone – with a potential for global reach – both authors conclude that the Act contains several Achilles' heels that must be addressed during the ongoing legislative negotiations to ensure it *truly* protects EU values.

Artificial Intelligence (AI) has been raising high hopes, not only for techno-solutionists. By identifying patterns in large volumes of data, AI techniques can help analyse, predict, recommend and even take decisions in all domains of our lives – a feature welcomed by many. This typically boils down to the analysis and influence of social phenomena through their quantification. Such processes aren't perfect however: often only indirect data-points (proxies) can be collected with some aspects lost in translation. Moreover, it is feared that '[i]f you digitize a broken process, you have a digitized broken process'. Lawyers and ethicists in particular, always the spectre of new societal rushes, are mainly intrigued by one question: If gathering unprecedented amounts of data only shows us a glimpse of how the world *is* rather than how it *should be*, who is then to make the normative choices inherent to AI's design and use – and how?

Building the EU approach

States have raced to harvest the opportunities of AI and – not always with as much enthusiasm – to mitigate its risks by adopting AI Strategies (e.g., China 2017, EU 2018, US 2019). The European Strategy has aimed to boost AI's uptake and research, tackle related socio-economic challenges, and ensure an adequate ethical and legal framework. Such regulatory endeavors already faced fundamental difficulties when it came to merely defining AI. This was the first agenda item of the European Commission's High Level Expert Group on AI (HLEG) that was tasked with preparing AI ethics guidelines and policy recommendations. The HLEG extensively discussed the normative framework that should guide AI regulation, and convincingly argued it should be based on the protection of fundamental rights, democracy and the rule of law. Its debate topics also spanned the dependence of AI's risks on the context and mode of application (think of its use for medical treatments versus its use for song recommendations) and, technically, the relevance of the difference between traditional rule-based AI with prior codified rules and data-driven AI, which itself derives rules from data.

While stressing that all use of technology should be trustworthy, the HLEG argued why AI may deserve a specific regulatory approach – *inter alia* due to a lack of transparency and accountability that isn't adequately tackled by existing rules. In its guidelines, the HLEG therefore defined 'trustworthy AI' as AI that is lawful, ethical, and robust. To interlink law and ethics, the guidelines' ethics principles were derived from fundamental rights and concretized into 7 key requirements that developers and deployers of AI systems should meet. Rather than a simple tick-the-box exercise, these requirements must already be considered in the design phase, with the help of an assessment list that operationalizes each requirement through specific questions to address. In terms of policy suggestions, the HLEG recommended a risk-based approached to AI regulation, and the introduction of new legislation to ensure accountability for AI's adverse impact.

Scope of the Act

The European Commission was willing to take up this suggestion. In April 2021, it proposed the AI Act, aimed at covering legal gaps in existing legislation such as the GDPR and consumer protection laws rather than filling a legal vacuum. Following the 2020 White Paper on AI, the proposed regulation strives to create an 'ecosystem of excellence and trust' in Europe, protecting fundamental rights, and eliminating obstacles to trade. With a broad geographic scope, it shall apply to all providers of AI systems that have an impact within the Union. While various AI application domains are targeted, systems exclusively developed or used for military purposes are excluded, which – given the significant risks associated therewith – has faced critique.

In terms of defining AI, the Commission took inspiration from the definition proposed by the HLEG and the OECD. The Act defines AI broadly, covering not only AI systems based on machine learning approaches, but also those driven by logic- and knowledge-based approaches as well as statistical approaches, producing 'outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with' for human-set goals. As a wide AI definition also entails a wide scope of AI applications subjected to the Act, a definitional battleground is emerging during the ongoing negotiations on the Act.

Regulating by risk

Adopting a risk-based approach, the Act introduces four categories of risks. First, AI systems with *minimal or no risks* – which according to the Commission would constitute 'the vast majority' of all AI systems – are permitted without restrictions (though voluntary codes of conduct are encouraged). Second, users of emotion recognition and biometric categorization systems, as well as providers of AI systems with a risk of deception (such as deep fakes and chatbots) are required to implement transparency measures. Importantly, these transparency measures are formulated as *obligations* rather than as a *right* to receive information – an example of the notable absence of rights for individuals throughout the Act.

The third and arguably most important category of *high risk* AI systems spans (1) AI systems intended to be used as safety component of products that are subject to third party ex-ante conformity assessment and (2) a selected list of 'other stand-alone AI systems with mainly fundamental rights implications' incorporated in an Annex. As it currently stands, this category includes the use of AI applications in, for instance, the operation of critical infrastructure, judicial decision-making, the evaluation of asylum applications, the assessment of individuals' creditworthiness, risk assessments by law enforcement, the evaluation of exams to determine access to education and for recruitment purposes. These *high risk* applications must meet certain requirements inspired by the 7 key requirements for Trustworthy AI formulated by the HLEG (such as transparency, human oversight and data governance) and undergo an ex ante conformity assessment.

While this obligation can enhance protection against their adverse impact, one can certainly question this list-based approach, which risks being over- and under-inclusive. Moreover, under its current shape, this list is blatantly too short (for instance, applications used for biometric categorization and emotion recognition are missing). Furthermore, the requirements that these high-risk systems must meet are not going far enough. Too narrow appear the obligations for the provider to judge the overall residual risk of the system as acceptable (a clear lack of an independent forum of accountability) and for essential information about the system to only be provided to users rather than individuals subjected thereto. Practically unachievable seems the obligation to only use datasets that are free of errors and complete. In addition, providers are presumed to be in conformity with the requirements if they are in conformity with harmonised standards, which essentially comes down to the outsourcing of normative requirements – which are political in nature and pertain to fundamental rights – to standardisation bodies which are anything but democratic in nature.

AI systems of the fourth and last category, *unacceptable risk*, are prohibited. This category is – understandably – very narrow, yet – less understandably – not subject to periodic revision, unlike the *high risk* category. As regards the prohibition on 'manipulative' AI systems, the qualification that they must cause physical or psychological harm may be hard to prove. More importantly, the prohibition on 'real-time' biometric identification by law enforcement in public places (such as facial recognition technology), is open to several exceptions, hence almost rendering it deceptive to be included in a list of prohibitions. While the exceptions seem reasonable at first sight (e.g. finding missing children or 'suspects of a criminal offence'), they open pandora's box by essentially allowing a mass-surveillance infrastructure to be built and implemented in public places, and – in light of path dependency – irreversibly so. Hence a significant number of civil society organisations are calling for a complete ban on facial recognition technologies.

Debating (beyond) the novel regulation

Further weaknesses of the proposal include its entirely nationalised enforcement, which judging by the GDPR's example risks leading to uneven protection of EU citizens, strongly based on individual Member States' resources and interests. The proposal also lacks a consumer complaint mechanism and worker protection. Furthermore, discussions are to be anticipated on the Commission's delegated powers (especially to update the Annexes) as well as on the scope of the regulation (e.g., narrowing down the definition of AI, or introducing exemptions for national security and law enforcement), which can further weaken the protection afforded by the Act. Yet despite these points of critique, it does show significant progress in the EU's approach of acknowledging AI's impact on fundamental rights and mitigating these risks by harmonising legislation.

The initial wish of AI developers for a unified list of existing legal requirements applicable to AI systems (as many different existing regulations may apply to the application) shows the complexity of the regulatory landscape that all stakeholders face. The AI Act does not constitute a codification of all legislation applicable to AI. Rather, it seeks to provide answers on how to apply a set of additional horizontal rules to AI systems so as to protect the health, safety and fundamental rights of individuals. However, this aspired fundamental rights-based approach does not deliver on its promise. The deficiencies discussed above – such as the inadequate risk categorization of certain AI systems, the lack of essential accountability mechanisms, the absence of a complaint mechanisms for individuals affected, and reliance on national enforcement even for very large transnational corporations – jointly threaten the Act's effectiveness in mitigating AI's risks to fundamental rights.

Moreover, the Act also falls short of considering societal harms that go beyond an impact on the fundamental rights of individuals. Importantly, AI systems can also adversely affect societal interests such as the rule of law and the democratic process. The legislator should therefore ensure that all EU values listed in Article 2 TEU are safeguarded against AI's adverse impact, for instance by explicitly referring thereto. Furthermore, rather than only imposing obligations on providers, the AI Act should acknowledge the role of citizens and ensure they have a right to redress to strengthen the Act's accountability mechanism. By enabling both public and private enforcement mechanisms, EU values stand a better chance of being protected against AI's risks.

With its proposed AI Act, the EU thrust itself forward as a pioneer in regulating AI. By relying on its first mover advantage, it may de facto set global standards (the Brussels effect), as achieved with the GDPR. However, getting in early doesn't necessarily mean getting it right. It remains to be seen whether the Act can still be improved during the ongoing negotiations in the Council and Parliament and whether an appropriate regulatory balance can be struck between the different rights and interests at stake. By strengthening the AI Act and sharpen its legal safeguards, hopefully ethicists and lawyers can soon be joined by others to ensure its practical effectiveness in protecting EU values in the age of AI.

Nathalie Smuha

Legal Scholar and Philosopher at KU Leuven

Nathalie Smuha is a legal scholar and philosopher at KU Leuven, focusing on the impact of AI and digital technologies on human rights, democracy and the rule of law.

Anna Morandini

PhD Researcher at European University Institute

Anna Morandini is a PhD researcher at the European University Institute in Florence. She studies European Digital Law with a focus on platform accountability.