Spotlight on risk

# Nine recommendations for the governance of AI systems

Marie-Valentine Florin

28 February 2022

Some governance functions traditionally performed by humans are increasingly informed and sometimes automatically executed by machine learning algorithms (governance by machine learning) to benefit society. Therefore, it is necessary to think also about the governance, or regulation, of digital technologies. This is the path that the EU has taken in a sequence of policy initiatives, with important milestones including the General Data Protection Regulation (GDPR) and the proposal for a legal framework on AI. The context is marked by a desire to stimulate innovation while recognising that the digitalisation of society comes with risks that must be attended to.

Over the course of the Horizon 2020 Trigger project, its partners have analysed various aspects of EU governance of and by digital technologies. This issue is increasingly important, as digital technology becomes an ever more central feature of the global governance landscape. Their research on digital technology governance culminated in a

report[1] produced in 2021 by IRGC that sets out nine recommendations for how the EU could proceed in this area. The recommendations draw on earlier work, including research on two digital technology domains (artificial intelligence (AI)/machine learning (ML) and blockchain technologies) as well as wider cross-cutting themes[2] and explorative scenarios[3] of evolving EU digital policy. We summarise the nine recommendations below.

## 1. Prioritise regulatory attention to algorithmic decision-making

At the core of the EU's values and traditions is the principle that technology should be at the service of humans and not the other way around. This is especially relevant in the case of artificial intelligence. Concerns arise when a machine learning system is used in an automated decision-making system that makes and implements decisions without human intervention or control, such as with autonomous driving or, in some cases, facial recognition. More benign-seeming cases exist with, for example, appointment scheduling systems in hospitals, although failure to recognise the urgency of a case may lead to harmful consequences for patients. Therefore, governments should ensure that any algorithmic decision-making system concerning critical matters for consumers and citizens, which occurs without appropriate human oversight, is treated for regulatory purposes as a high-risk application and prioritised in regulatory attention. This is the goal of the proposed EU legal framework on AI[4], and we believe it is correct. However, case-specific decisions should be made when this approach would lead to banning the use of AI systems in applications that could cause unacceptable risks but also high benefits.

## 2. Be clearer about how principle-based and risk-based regulations are used

There are important differences between risk-based and principle-based regulation. Both have a role in the EU's governance of digital technologies, and both must be developed in a clear, nuanced, consistent and implementable manner. The EU has emerged as a global regulatory leader with its principles-based ethos, emphasising fundamental individual rights as a keystone of technology governance. For example, the EU High-Level Expert Group on AI established ethics principles for trustworthy AI[5]: respect for human autonomy, prevention of harm, fairness and explicability. But it is also important to see the elaboration and operationalisation of such governance principles as an ongoing task relevant to business and the public sector. First, there must be clarity and consensus on what each principle means. Second, on a technical level, it must be possible to operationalise these principles: developers must be able to understand and implement them in code, and the industry must be able to translate them into clear assessment and auditing systems. For example, this is what a group of industries convened in the Veritas Consortium[6] have done for, first, defining operational components of the fairness, ethics, accountability and transparency principles that align with corporate objectives in finance and insurance; and then developing guidelines for assessing fairness in credit risk scoring, customer marketing, and insurance predictive underwriting and fraud detection.

The EU has identified specific application domains where high risks to individuals may manifest. The proposed legal framework on AI adopts a risk-based approach, implying that possible infringements on the principles have been translated into risks, with causes and consequences. This also implies that evidence of damage or harm must be made to justify restrictions or prohibitions on certain applications of AI. We now have to work to define those as clearly as possible, yet remain open to flexibility as the regulation must adapt to technological developments and societal choices.

## 3. Consider a precautionary approach to some applications of AI/ML

The precautionary principle[7] has been mainly applied in the environmental and human health domain to avoid potentially very severe yet uncertain risks. It may be worth learning from it and possibly explicitly expanding its scope to cover potential severe risks posed by AI in some domains. The speed with which new digital technologies propagate across societies can be faster than the pace at which a robust evidence base about the impacts of these technologies can be developed. Application of the precautionary

principle can be contentious and seen as hindering innovation. However, if it is interpreted as requiring harm-avoidance measures until there is sufficient evidence that a technology's impacts will not cause severe harm, or that techniques exist to prevent harm, then adopting a precautionary approach would encourage the development of, for example, better privacy-by-design or non-discriminating techniques. Thus, precaution can be seen as not only protecting citizens but also stimulating innovation towards fairness, ethics and other fundamental values. It is worth noting that the proposed expansion of the precautionary principle to cover digital technologies is in line with the "Ethics Guidelines for Trustworthy AI"[5] mentioned above and recommended under certain conditions by the group in its subsequent "Policy and Investment Recommendations"[8].

## 4. Focus on domain-specific regulation

This is one of the key recommendations that emerged from the Trigger project in 2020, and was already one of the main suggestions of IRGC's work on decision-making algorithms[9], produced in 2018. Today, it is a major aspect of the EU proposal for a regulation on AI, which addresses the risks of AI in specific application domains. The primary focus of policymakers should not be on authorising or prohibiting a technology per se, but rather on its specific applications and uses, because this is where benefits and risks arise. This is similar to the approach taken by the OECD.AI Network of Experts that developed the OECD Framework for Classifying AI Systems[10] as a tool for policymakers, regulators, legislators and others so that they can assess the opportunities and risks that different types of AI systems present in different industries and application domains, and to inform their national AI strategies. In addition, policy should not focus only on those aspects of a technology that may need to be restricted or regulated in some way. Where appropriate, policymakers should also advocate for the increasing use of digital technologies that can mitigate domain-specific risks. A balanced risk assessment is needed, encompassing not just potential undesirable outcomes that require regulation, but also the desirable outcomes that technologies may be able to deliver.

## 5. Invest in the development and implementation of technology for privacy and trustworthiness

The EU should invest in and incentivise the development and use of technologies that help to protect fundamental rights 'by design'. This could become a leading advantage of Europe versus the US and China. Doing this implies paying greater attention to 'governance by digital technology' alongside the more familiar 'governance of digital technology', recognising that digital technology can help solve some important governance challenges. Enabling technologies, such as various confidential computing techniques, can offer a potential "risk-superior" solution on the trade-offs between security, privacy (as a fundamental right) and innovation (which implies broader data sharing than currently generally allowed in Europe). The EU should thus seek to incentivise those technological solutions that contribute to achieving one or more of the requirements for trustworthiness. Trust-building / privacy-preserving technologies[11] are particularly relevant. This potentially includes using solutions that would embed legal rules in technical specifications that could be mandated across the EU, although there could be concerns regarding the legitimacy of doing so.

## 6. Define ethical red lines

A transparent and legitimate process is needed to assess whether any applications of digital technologies should be ruled out, regardless of their potential benefits, because the risks they pose are too great or incompatible with the EU's fundamental values. An example of such an application might be a lethal autonomous weapons system ("killer robot") in which killings are decided on by an algorithm. The proposed EU legal framework on AI[12] adopts this approach for banning AI systems that present 'unacceptable risk' considered a "clear threat to the safety, livelihoods and rights of people... This includes AI systems or applications that manipulate human behaviour to circumvent users' free will (e.g. toys using voice assistance encouraging dangerous behaviour of minors) and systems that allow 'social scoring' by governments."

## 7. Clarify the scope, rationale and goals of digital sovereignty

Greater clarity is needed regarding the intention and concrete implications of a country's (or the EU's) goal of digital sovereignty. There are fundamental questions about how a government that wants to promote digital sovereignty can engage with the rest of the world. Digital sovereignty can be understood in this context as the objective of ensuring that the country retains the value of its digital resources and has the capacity to make and enforce decisions about the use of digital technologies across its territories. However, what does this mean in practice? Before engaging in this, governments should spell out what they see as the costs as well as the benefits of prioritising sovereignty, and they should also explain in greater detail how sovereignty in the technological domain, in general, might interact with developments in other major domains of global interdependency, including industry, trade and competition policy.

## 8. Balance public and private forms of governance

The EU and individual governments should weigh the relative pros and cons of public and private forms of governance with regard to maximising their effectiveness at shaping the type of governance landscape they would like to see develop. With the GDPR, the EU has shown that it has the heft required to project rules globally, and member states support it. However, it would be unwise to generalise the case of data protection to digital technologies more generally and conclude that flagship regulations are the most effective way of proceeding. There may be instances where European governments and the EU would enjoy more leverage by seeking to influence sectoral standards, guidelines and codes of conduct, ex-ante conformity assessments or self-regulation more generally. A good example is what NIST[13] is doing in the US with the proposed risk management framework for AI[14]. Articulating the pros and cons of hard and soft laws will be particularly needed for implementing the regulations that will result from the proposed EU legal framework on AI, which will perhaps be even more difficult than implementing the GDPR. However, compliance and enforcement are a particular challenge with private forms of governance. Platform governance, perhaps with the EU Digital Services Act Package[15], is likely to be a key test-bed for mixed public-private approaches to digital technology governance.

## 9. Develop a strategy for working with other key global governance actors

The EU and member states should clarify how they intend to work with other key actors, the most important of whom are the US and China, given these countries' clear leadership role in developing and deploying digital technologies. Acknowledging the complexity of relationships with these countries is a crucial starting point for Europe to find a consistent and durable way of acting on its goal of increased sovereignty and autonomy. It will be necessary to clarify whether or how a joint agenda might constrain the EU's digital sovereignty. Decisions regarding the hosting of public servers and cloud computing, where data is moved across countries, are not neutral in this respect and should be scrutinised, especially if servers are located outside of regulatory jurisdictions. There may be differing levels of consensus that could be achieved with different groupings: only a very thin agreement might be possible between all three of the US, China and the EU, whereas a much greater level of overlap is likely to be possible between the EU and US.

**Acknowledgements**
The author would like to thank Jim Larus (EPFL), Elettra Ronchi (WHO) and Michael Veale (UCL) for reviewing this article. However, the views presented are not a consensus judgement by IRGC or its reviewers.

**Disclaimer**
A first version of this article was published in a report of the Trigger project, authored by Aengus Collins, and as a blog post: trigger-project.eu/2021/07/30/nine-recommendations-for-eu-governance-of-and-by-emerging-technologies/

1 Florin, M.-V. & Collins, A. *D4.6 WP4 final report*. (2021).
2 Renda, A. *D4.4 Cross-cutting themes from tasks 1, 2 and 3: principles and guidelines for an overarching governance framework*. (2020).
3 Renda, A. *D4.5 Explorative scenarios of governance by and of emerging technologies with far- reaching consequences on society and the economy*. (2020).
4 EC. *Proposal for a regulation laying down harmonised rules on artificial intelligence*. (2021).
5 EC, Directorate-General for Communications Networks, Content and Technology. *Ethics guidelines for trustworthy AI*. (2019).
6 Veritas consortium. *Veritas document 1. FEAT fairness principles assessment methodology*. (2020).
7 EC. *Communication from the Commission on the precautionary principle (COM/2000/0001 final)*. (2000).
8 AI HLEG. Policy and investment recommendations for trustworthy artificial intelligence. (2019).
9 IRGC. *The governance of decision-making algorithms*. (2018).
10 OECD.AI. *The OECD framework for the classification of AI systems*. (2022).
11 Center for digital trust (C4DT). C4DT.
12 EC. Europe fit for the Digital Age: Commission proposes new rules and actions for excellence and trust in artificial intelligence. *European Commission* (2021).
13 National Institute of Standards and Technology (NIST). AI risk management framework. *NIST* (2021).
14 National Institute of Standards and Technology (NIST). *AI risk management framework. Concept paper*. (2021).
15 EC. The Digital Services Act package. (2022).