December 2019

Considerations for EU Proposals to Regulate AI

Artificial intelligence (AI) systems will help us more effectively address a wide range of social challenges in fields such as healthcare, transportation, and environmental sustainability. They will also help European companies large and small better compete locally and globally, and European governments better anticipate and meet the needs of their citizens.

AI systems will also raise new questions and potential risks. This will require both the private and public sectors to adopt new approaches to manage these risks and to protect individuals from harm.

Microsoft believes that companies that create technology must accept greater responsibility for helping secure the promise of its future. To that end, we have identified six principles to guide our development and use of AI: fairness, reliability and safety, privacy and security, inclusivity, transparency, and accountability.

While we believe that responsible development and deployment of AI is primarily the responsibility of industry, over time, we see regulation of certain AI systems and deployments as inevitable to ensure and maintain trustworthiness. We are, however, acutely conscious of the need to move forward with care. Given the nascent nature of AI, and the almost infinite scenarios and domains in which it can be deployed, it will be important for any regulation of AI to carefully target undesirable behaviors and outcomes, but without restricting beneficial uses of AI or undermining incentives for innovation.

Drafting rules for a future that we can barely predict is challenging—something we know from our own experience. Microsoft is working hard to develop and test standards, frameworks, and processes for the development and use of AI systems that adhere to our six principles. This is not a simple exercise. Many of the questions raised by AI are technically difficult. AI can also raise exceedingly difficult ethical and sociotechnical questions, such as how to balance an AI system's clear benefits to one group against potential risks to another, or how to assess what is and isn't "fair"—questions that may yield different answers in different cultures and contexts.

As the European Commission considers how best to regulate AI, it will need to grapple with these and many other difficult questions. We list some of these considerations below, along with our initial thoughts:

- ***Which technologies are in scope?*** As a threshold matter, it will be critical to determine which AI systems or use cases need regulation, and then to clearly define those that are in scope. AI is not one technology, but rather an umbrella term that includes multiple technologies, including systems that perform computer-based

perception, learning, and reasoning. These technologies can be used separately or combined to yield systems that perceive, classify, recommend, predict, guide, or otherwise reason or act in an automated manner. Many AI systems rely on basic statistical analyses of data or richer machine learning, including supervised learning, unsupervised learning, and reinforcement learning. Even a single AI technology may raise no risk of harm in one setting (AI-powered steering in toy vehicles), but significant risks in another (the same in real vehicles). The Commission should work closely with industry to identify the specific AI systems or deployments that require regulation, and then craft legislative language that clearly defines what is and isn't regulated.

- ***Which actors are in scope?*** Discussions on AI regulation do not always clearly distinguish between AI developers and those deploying AI. Any legislation should be clear on this point, and any obligations should be differentiated as between developers and deployers. For instance, certain transparency obligations may make sense for developers (*e.g.*, to describe the limits of an AI system), and other for deployers (*e.g.*, to disclose the fact of AI processing to end-users). Developers may be uniquely situated to tackle the challenge of bias when training AI models, while deployers are typically in the best position to ensure that systems are not used in ways that unfairly discriminate. Similarly, while developers may hold important information about the datasets used to develop AI systems, only the deployer may have the ability to monitor the system once it has been released "in the wild".

  Further, any regulation should take account of the fact that there may be many links in the chain from development to ultimate deployment (*e.g.,* company A develops an AI solution that company B incorporates into its medical device, which it then sells to a hospital, is used by a doctor, and impacts a patient—which entities in this chain have which obligations?). Given the innumerable scenarios in which AI systems may be deployed, the Commission should ensure that legislation is both appropriately targeted and sufficiently flexible to adapt to these myriad scenarios, and that all stakeholders can easily determine which obligations apply to them.

  Distinctions should also be drawn between public- and private-sector deployments. Because users often don't have alternatives to interacting with governments (*e.g.,* to obtain certain social services or benefits), and because governments have unique surveillance and sanctioning powers, public-sector deployments of AI may pose significantly higher risks to individual rights than most private-sector deployments. Legislation should reflect this, imposing more significant obligations on public-sector developers and deployers of AI.

- ***Risk-based or one-size fits all?*** Many AI systems pose extremely low (or even no) risks to individuals or society (*e.g.,* AI systems that optimize storage of items in a warehouse, suggest music playlists, or fix typing errors). These systems do not require new regulation and should be expressly excluded from scope. Any legislation should instead focus on AI systems or use cases that pose a material risk of individual or societal harm—i.e., that have "consequential impact," such as systems that may

adversely affect people's legal rights, their physical or emotional well-being, or their ability to access healthcare, financial services, or employment opportunities.

The fact that an AI system might have a consequential impact should not necessarily preclude its development or use. But the scope of any regulatory obligations should be a function of the degree of risk and potential scope and severity of harm. In high-risk deployments—which are likely to be primarily (but not always) in the public sector—it may be appropriate to impose mandatory obligations.

- ***Prescriptive or governance-based?*** Given the nascent nature of the technology and sociotechnical quality of many of its most significant challenges, prescriptive regulation of AI (*e.g.,* "an AI system must meet this level of accuracy or accountability") will likely be unworkable in practice. Similarly, requiring developers or deployers to ensure that every possible use of an AI system is "fair" or "unbiased" will be impractical. Fairness is a complex concept that is deeply contextual; given the many human and technical sources and often incongruous definitions of unfairness, it simply will not be possible to fully guarantee a system's fairness, or lack of bias. AI systems also may necessitate trade-offs between competing priorities, meaning a system may be fairer from the perspective of one group than another.

  For these reasons, we think that at this stage, a governance-based approach to legislation—i.e., one that identifies broad objectives, and then sets out the processes that developers and deployers should follow to best achieve those objectives—is likely to be more effective than a prescriptive one. The goal of these governance processes should be to help developers and deployers of covered AI systems identify and quantify any relevant risks of harm to individuals or society and, where those risks are determined to be significant, to implement measures to mitigate against them.

- ***What core goals should the regulation advance?*** As explained above, AI legislation should articulate the core goals or outcomes that it seeks to achieve, and then identify the processes and procedures that covered entities should undertake to advance those goals. Especially given the nascent state of AI development and the potentially broad reach of the legislation, we encourage the Commission to focus on goals around which there is broad consensus, and on processes that will be impactful without being unduly burdensome or impractical. In that regard, certain key principles in the High Level Expert Group on AI ("HLEG") Ethics Guidelines (several of which have analogues in Microsoft's own principles) can provide a useful starting point. These include:

  - **Fairness.** Those who develop or deploy AI systems should take steps to help ensure that the system treats people fairly. This means working to identify and mitigate potential harms, such as uses that may perpetuate undesirable social biases or unlawfully discriminate.

  - **Reliability and safety.** AI developers should strive to provide systems that perform reliably and safely, including with robust protections from cyberattacks.

- **Accountability.** People who design and deploy AI systems should be accountable for how their systems operate. In particular, any AI system having a consequential impact on any person or group should require human oversight / human-in-the-loop, as well as the possibility for human review of adverse decisions.

- ***What type of obligations should apply?*** As noted, the goal of any horizontal AI legislation should be to help developers and deployers of AI systems identify potential risks so that they can provide transparency about those risks and take appropriate steps to mitigate them. With that objective in mind, the Commission might wish to consider the appropriateness of the following types of obligations:

  - **Governance and oversight.** For organizations engaged in the development or deployment of AI systems with consequential impacts, it might be appropriate for them to adopt an AI governance framework. This framework could entail having the organization adopt principles for trustworthy AI (based on objectives set out in legislation), and assigning specific individuals or groups within the organization to promote compliance with the principles. The framework could also entail having the organization take steps to raise internal awareness of the need for such compliance, including through company-wide guidance and trainings, and to implement an escalation process through which employees could raise compliance concerns and have those concerns resolved.

  - **System envisioning (impact assessments).** As discussed above, developers and deployers of AI systems with consequential impacts should understand these impacts and identify appropriate mitigations. An appropriately tailored impact assessment could help them do so. This would consider the system's purpose; the key intended use cases, as well as foreseeable misuses; the domains in which the AI system is most likely to be used (*e.g.,* school? commercial organization? hospital? retail sector?); the relevant stakeholders (i.e. those who are responsible for, will use, or will be affected by the system); and the nature and scope of the potential risks and harms, including upon vulnerable groups. The assessment should illuminate tensions between the interests of various stakeholders, and consider how the system is likely to evolve—for example, how might feedback loops and changing deployment conditions affect the system's potential impacts? Outcomes and risk mitigation strategies should align with the organization's ethical principles for AI development.

    In structuring any such impact assessment, we encourage the Commission to take account of the feedback provided to the HLEG in the context of their pilot of their Trustworthy AI assessment. Microsoft was one of several entities that piloted that assessment. Our engineers and researchers felt the assessment did not fit well with the lifecycle in which AI systems are typically developed and deployed. Also, many questions required a "yes/no" response, when in fact the answer was more nuanced. An impact assessment that helps users spot and evaluate relevant issues is more appropriate in the context AI, so that developers and deployers can identify

and weigh concerns and make informed decisions about whether and how to utilize the relevant AI system. Microsoft's team also concluded that it would be hard for smaller organizations to implement the HLEG assessment list; almost all sections required multiple individuals to answer the questions, which might be beyond the means of smaller companies and start-ups.

Finally, to the extent that any regulation requires that impact assessments consider impacts on fundamental rights, it will be essential to define precisely the rights to be evaluated. Reference to specific EU (or international) instruments, and to the articles and specific rights themselves, will be important. Those rights might include—depending on the AI system and its anticipated use cases—some or all of the rights to a fair trial, effective remedy, privacy and protection of personal data, equality and non-discrimination, and freedom of expression, information and association. Organizations will also require guidance on how to balance "competing" fundamental rights (*e.g.,* the right to freedom of expression as against a prohibition on discrimination), and the extent to which any objective justifications or permitted derogations to these rights are relevant in the context of an AI system. Learnings from human rights impact assessments, which some organization currently use, for instance, to evaluate their supply chains, could be useful models here.

- **Transparency**. Transparency is key to developing trust in AI systems. Users won't trust AI if they don't understand it. Transparency is also important to mitigate risk. Developers and deployers should disclose, for example, what they know about how and why bias might be introduced into the system; any limitations in a system must also be known and understood. This transparency entails several related but distinct elements:
    - *Communication*. Developers and deployers of AI systems should communicate their systems' capabilities, limitations, and risks, including any high-risk applications that should be avoided. Developers should also disclose any known system requirements, while those deploying AI systems should inform people when consequential decisions about them may have been influenced by the system, and who is accountable for the system.
    - *Explainability*. People should be able to understand, monitor, and respond to the technical behavior of an AI system.
    - *Traceability*. Developers and deployers of AI systems should document their goals, definitions, and design choices, and any assumptions they have made in designing their systems. Such documentation should conform to emerging best practices for documenting key characteristics of the datasets used to train and test AI systems, as well as resulting models.
- **Documentation.** Documentation is closely linked to, and an essential element of, the transparency / traceability obligation described above. It's value extends

beyond traceability, however, and should also include documentation of key aspects of the data used to train or test the AI system, as well as system requirements and architecture (including major system components, models, assumptions, and mitigation strategies) and responsible release criteria.

- **Appropriate safeguards.** Where a risk assessment identifies significant consequential impacts, safeguards (in addition to transparency) may be appropriate. Regulation should not dictate what those safeguards should be, however; instead, a requirement to impose appropriate measures would give developers and deployers the flexibility to determine and implement those mitigations that best fit the context. Conceptually, this standard could borrow from the security obligations set out in the GDPR. GDPR Article 32, for example, requires that organizations adopt measures that ensure a level of security "appropriate to risk," taking into account "the state of the art, the costs of implementation and the nature, scope, context and purposes of the processing as well as the risk of varying likelihood and severity for the rights and freedoms of natural persons." Measures that might be appropriate (and which any legislation could reference in a recital for consideration) could include, for instance, prototyping / "ring testing" the system in scenarios that reflect likely conditions of use; a mechanism to monitor and receive feedback about issues that arise after the system has been deployed; "humans in the loop," i.e., subjecting the system to ongoing human oversight and review; and a rollback plan in the event that the system does not perform as expected.

- ***What about sensitive use cases?*** As AI systems become more prevalent, the types of sensitive uses that merit regulation will become clearer. Accordingly, it will be important for the Commission to regularly revisit the question of whether specific legislation is required for sensitive uses. The HLEG Guidelines identify some potential applications that are particularly sensitive, such as lethal autonomous weapons. Courts and regulators are also currently grappling with the legality of public-sector uses of facial recognition technologies ("FRT"), including for surveillance purposes. While EU law already provides clear parameters for assessing the lawfulness of biometric technologies from a data protection perspective, the rules that govern the ethics and other risks of FRT deployments are less well defined. For that reason, the Commission might wish to consider specific rules governing the use of FRT by the public sector in particular, given the heightened risks inherent in governmental use of this technology.

- ***How frequently should any regulation be revisited?*** AI systems are inherently disruptive, which means that our ability to identify today the ways in which such systems might have negative impacts tomorrow is limited. At the same time, many AI systems will be deployed in ways that render them subject to existing legal protections (*e.g.,* rules prohibiting discrimination). To account for these factors, AI regulation should be flexible and not be static. Any horizontal AI regulation should anticipate that further, sector-specific regulation might be necessary for certain scenarios (*e.g.,* autonomous driving), or for new sensitive use scenarios that may emerge. In this same

spirit, we encourage the Commission to ensure that any new horizontal rules do not overlap or conflict with any existing regulatory obligations.

Microsoft looks forward to collaborating closely with the Commission on this important endeavor. For further information, please contact [redacted] [redacted] [redacted]@microsoft.com).