CENTER FOR DATA INNOVATION

RECOMMENDATIONS TO THE EU HIGH LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE ON ITS DRAFT AI ETHICS GUIDELINES FOR TRUSTWORTHY AI

The Center for Data Innovation is pleased to submit feedback to the High-Level Expert Group (HLEG) on AI on its draft AI Ethics Guidelines for Trustworthy AI. The Center is a nonprofit research institute focused on the intersection of data, technology, and public policy. With staff in Washington, DC and Brussels, the Center formulates and promotes pragmatic public policies designed to maximize the benefits of data-driven innovation in the public and private sectors. It educates policymakers and the public about the opportunities and challenges associated with data, as well as technology trends such as artificial intelligence, open data, and the Internet of Things. The Center is affiliated with the Information Technology and Innovation Foundation (ITIF), the top-ranked science and technology policy think tank in the world.

GENERAL COMMENTS

The Center acknowledges that this initiative is timely and supports it for having involved a broad diversity of stakeholders within the HLEG, and for its non-legally binding nature. The guidelines are an opportunity to further the conversation on AI, which given the stakes, is much needed to provide a sense of urgency to the European policymakers, business community, academics, and the general public about the potential opportunities to use AI to improve the economy and society. The emphasis on addressing the needs of vulnerable groups, ensuring diversity and inclusion, and addressing skills are key contributions in the HLEG's document.

The guidelines aim to provide concrete guidance on how to implement and operationalize "trustworthy AI" systems that "maximize the benefits of AI while minimizing its risks." While this goal is worthwhile, the guidelines have five main problems: 1) they present an overall negative tone towards AI; 2) they overlook the importance of EU leadership on AI adoption as a means of influencing global AI ethics; 3) they incorrectly suggest that developing a European AI ethics governance system will allow the EU to significantly differentiate its AI solutions, thereby gaining global market share; 4) they inaccurately frame AI as a technology that requires ethical tradeoffs, instead of one that can be used to improve ethical behavior; and 5) they propose principles such as transparency and explainability that would limit AI development.

First, when the HLEG recognizes that "on the whole, AI's benefits outweigh its risks," it is damning with faint praise. In fact, the overall narrative it presents about AI is negative and unbalanced, especially given the vast number of tangible examples of AI's benefits already in existence and the relatively few instances of substantial and unmitigated AI harms from systems that have actually been deployed (as opposed to being tested). Indeed, there are several examples in the document suggesting that AI has greater potential to cause harm rather than to produce benefits. For example, in Chapter 1, Section 3.3 ("Respect for democracy, justice and the rule of law"), the HLEG suggests that AI systems "interfere with democratic processes" and "undermine the plurality of values and life choices." Such allegations are not supported by

evidence and stand to diminish public acceptance of AI, which would slow down adoption. In contrast, Chapter 1, Section 3.5 states that AI systems only "hold potential" in terms of how they can "improve scale and efficiency of government in the provision of public goods and services to society." This statement mischaracterizes the numerous examples of AI systems already in production in governments around the world, while significantly overstating actual real-world AI harms that have occurred. There are two reasons why harms are likely to be vastly less than portrayed. The first is that in existing EU laws and regulations would apply to most applications of AI, giving governments the right to bring action against potentially harmful cases. The second is that those laws and regulations, along with oversight by civil society and pressures from market forces (e.g., the desire of companies to sell AI applications and maintain healthy public reputations) will lead the vast majority of companies to work diligently to ensure that the AI systems they deploy are accountable and beneficial.

To address this shortcoming, the HLEG should provide more representative descriptions of AI's capabilities, and a clearer acknowledgement of where it is already delivering benefits and where concerns are merely speculative or have occurred but have been easily remedied. In particular, the HLEG should focus on informing the public about many of the positive use cases of AI, including industry-specific examples, as this will help create an environment that is more conducive to adoption of AI to the benefit of EU businesses, consumers, and others. For example, a negative tone could prove harmful to the development of a workforce with the technical skills that will be necessary for AI in Europe. European students will be unlikely to pursue a career in AI or related fields if those who contribute to its development are demonized. The HLEG's guidelines should not discourage policymakers from responding to legitimate concerns and discussing challenges, but they should also not encourage alarmists to delay progress.

Second, for all of its concern about the future of AI, the HLEG ignores the fact that the EU is unlikely to be able to influence global AI ethics if Europe is not a leader in AI development and adoption. Ensuring "technological mastery" to foster "trustworthy AI"—an objective the draft guidelines set forth—requires the EU to be a global leader in AI. Europe is facing intense global competition in AI, but the HLEG ignores the need for the EU to focus on boosting public and private sector investment, raising technical skills of its workforce, and designing a regulatory environment conducive to AI so that it can compete with countries like China and the United States. For example, leading AI research is coming from North America and China where large tech companies have set up their own AI research labs because they have better access to talent, funding, and data. In addition, EU regulators have not been sufficiently supportive of AI. For example, regulators should foster voluntary data sharing to increase access to valuable data sets that may enable advances in machine learning. Often, the public and private sectors hold valuable data but lack mechanisms to securely and efficiently share it. Moreover, some provisions of the GDPR limit data collection and sharing and include other measures that will limit AI adoption. Amending the GDPR to ensure it does not impede innovation should be seen as a priority. Yet the draft guidelines rarely refer to the importance of increasing R&D, improving

workforce training, or reforming regulations to make the EU more competitive in AI. The HLEG should draw attention to the fact that Europe is lagging in all three areas and should identify these priorities as a necessary precursor to influencing the global debate on AI ethics. In short, it is much easier for leaders to influence the overall direction of AI ethics, not only through market leadership but also through technological capability.

Third, the HLEG's guidelines naively suggest that "user trust" will enable Europe to be globally competitive in AI. This, to be blunt, is wishful thinking that is not supported by evidence or real logic. Past studies that have quantified user trust in digital technologies have found that the levels of consumer trust in the EU are similar to those in the United States, even though the U.S. privacy regulatory system is not as stringent as Europe's. It is not that well-established that user trust—beyond a baseline level—deters digital adoption, and there is little evidence that user trust will be a major driver of AI adoption. What will be the major drivers of AI adoption will be the innovativeness, quality, cost, effectiveness and breadth of AI applications.

Fourth, the HLEG incorrectly presents AI and ethics as a trade-off. For example, throughout the text, the draft guidelines suggest that an increased use of algorithms would lead to a host of harms, including exacerbating existing biases, discrimination, and inequalities. If the HELG is going to present such claims, it needs to thoroughly document them with more than assertions from civil society groups with an interest in limiting AI adoption. Moreover, it needs to examine all claims of harm not just from a first-order perspective (e.g., did a particular version of AI lead to troubling or problematic results), but from a second-order perspective as well (e.g., did the next version of the AI application fix that problem? did the application lose out in the marketplace to other applications that did not have that problem? etc.). A major problem with making these accusations, and implying that AI is inherently problematic, is that it will engender support for policies to regulate algorithms in ways that would harm consumers, businesses, and democratic values alike. Combating bias and protecting against harmful outcomes is important, but it should be made clear that if an algorithmic system produces unintended and potentially discriminatory outcomes, it is not because the technology or the developer is malicious. Rather, unforeseen limitations in the design of the system or reflections of real-world biases from training data may cause these types of errors, something one would expect with any new technology or system where developers are still learning and improving. But even where bias in AI systems may occur, in many cases, these systems are still likely to generate less bias than similar human processes. In addition, these biases can be identified and quickly improved, which is exactly what occurs in virtually all identified cases in the marketplace. Indeed, rather than treating AI as a technology that presents inherent ethical risks, the HLEG's draft guidelines should focus more on how AI could be used to address existing ethical problems by automating activities where humans have a propensity to act unethically, often unconsciously.

Finally, the HLEG should eliminate some of the principles and requirements it proposes in the draft guidelines, such as transparency and explainability. By proposing these concepts as

requirements for AI systems, they would hold algorithmic decisions to a standard that simply does not exist for human decisions and limit the use of some advanced algorithms that cannot easily be explained but offer greater accuracy. In addition, transparency requirements could entail code disclosure. The economic impact of asking for companies to reveal their source code would be significant as it would prevent them from capitalizing on their intellectual property and future investment, and AI R&D would slow because businesses could simply copy the work of others. A better alternative to transparency and explainability is algorithmic accountability—the principle that an algorithmic system should employ a variety of controls to ensure the operator can verify algorithms work in accordance with its intentions and identify and rectify harmful outcomes.

The draft guidelines, while well-intentioned, miss the mark in terms of outlining a path forward for how the EU can be a global leader in AI, and through this leadership, answer important ethical questions about the future uses of AI. Rather than attempting to proceed on its own at setting global norms on AI ethics, the EU should work to establish itself as a leader in AI development and use, and work with other countries to develop common baseline approaches to AI ethics.

INTRODUCTION: RATIONALE AND FORESIGHT OF THE GUIDELINES

The draft guidelines begin with an introduction that includes the definitions of key terms, including artificial intelligence (AI), ethical purpose, bias, trustworthy AI, and human centricity. It also recalls the process and the purpose of the HLEG, the intent of the consultation, the role of ethics in AI, and the scope of the guidelines. The HLEG should update and revise some of those definitions which fall short and clarify the implications for any stakeholders who may choose not to endorse the voluntary guidelines.

The definition of AI and bias in the glossary merits further elaboration. In particular, AI is not entirely, as stated by the draft guidelines, "designed by humans." Some forms of AI, particularly those using machine learning and deep learning, build models from data that require little to no manually engineered intervention. Indeed, a goal of many companies is to construct machine learning systems that can build other machine learning systems, such as Google's AutoML. In addition, the guidelines' definition of AI specifically does not make the distinction between two very different types of AI: narrow and strong. Narrow AI, also known as weak AI, refers to machine intelligence able to perform a specific narrow task for which they have been programmed, such Apple's Siri virtual assistant, which interprets voice commands. Strong AI, also referred to as artificial general intelligence (AGI), is a hypothetical type of AI that can meet or exceed human-level intelligence and apply this problem-solving ability to any type of problem.

The draft guidelines note that a "mechanism will be put in place that enables all stakeholders to formally endorse and sign up to the Guidelines on a voluntary basis." However, the draft guidelines contain no further information about the nature of this mechanism and what would be the consequences for stakeholders who do not wish to "formally endorse" the guidelines. There

is a risk that these voluntary guidelines may become an attempt at backdoor regulation, such as penalizing companies who do not adhere to it. To avoid that, the HLEG should not endorse any particular mechanism for stakeholders to adopt the guidance, but instead put it forth and let it stand on its own merits.

CHAPTER I: RESPECTING FUNDAMENTAL RIGHTS, PRINCIPLES AND VALUES - ETHICAL PURPOSE

The first chapter lists selected fundamental rights, principles, and values which, according to the HLEG, AI should comply with to ensure its "ethical purpose" and trustworthiness. For instance, the fundamental right "respect for human dignity" leads to the "principle of autonomy," which reflects the freedom of individuals to make their own choices and is operationalized by the value of "informed consent." The chapter concludes with a section on "critical concerns" raised by certain uses, applications or contexts of AI, such as citizen scoring and Lethal Autonomous Weapons Systems (LAWS).

This chapter contains multiple examples of a negative tone, flawed references, vague statements, and unrealistic requirements. First, accusing AI systems and industry of, for example, working against democratic processes and values, limits this document's legitimacy and credibility. The HLEG's statements about AI should be fair and balanced, and clearly distinguish when it is referencing speculative concerns versus proven ones. In addition, to understand the potential tradeoffs of limiting or slowing the advancement of AI, the HLEG should include examples of how AI solves many economic and societal challenges. Second, several instances in the guidelines suggest AI systems should be held responsible for achieving complete equality—an unreasonable standard that does not exist for non-AI systems and processes. The HLEG should also revise and clarify other unrealistic constraints and impracticalities, such as references to "high standards of accountability" which, left undefined, could lead to confusion and stifle innovation in Europe. Finally, concerns raised in the final section of this chapter with respect to explainability do not sufficiently credit the vast amount of research taking place to improve AI explainability. Other references even seem to discourage the integration and use of new technologies such as facial recognition, and fail to acknowledge the strategic importance of developing autonomous systems. This will limit Europe's competitiveness and its ability to protect its infrastructure while China and the United States, for instance, will be catching up and gain a competitive edge.

Chapter 1, Section 3.1 ("Respect for human dignity") suggests that businesses developing AI systems would treat people "merely as data subjects" and not with dignity or respect. This accusation does not accurately reflect how businesses using AI treat their customers and look to AI to improve product and service quality. Indeed, many businesses are investing in AI to deliver better quality or value to their customers. Accusing industry of such attitudes further feeds into the false narrative throughout the document that AI is negative.

Chapter 1, Section 3.2 (“Freedom of the individual”) states that protecting this freedom in the context of AI “requires intervention from government and non-governmental organizations to ensure that individuals or minorities benefit from equal opportunities.” However, the report does not discuss how AI systems can be used to support this goal, such as by reducing gender biases in recruitment processes. Moreover, it implies that companies employing AI systems are more likely to discriminate against certain groups.

Chapter 1, Section 3.3 (“Respect for democracy, justice and the rule of law”) asserts that AI systems destabilize democratic processes and societies, and “undermine the plurality of values and life choices.” But again, these claims are not made on the basis of careful research and review of evidence. Moreover, such unfounded claims will not help nourish trust in AI from users, negatively impacting social acceptance of AI and, in turn, slow down the adoption of AI technologies.

This section also would require AI systems to take on responsibilities that would be impractical. For example, the HLEG says AI system could “abide by mandatory laws and regulation, and provide for due process by design.” Without mentioning which laws and regulations—and whether they are local, national, regional, or global ones—they refer to the “right to a human-centric appeal, review and/or scrutiny of decisions made by AI systems.” However, the guidelines do not specify what this “right” would entail or how it could be operationalized, setting up a vague standard that most businesses will be unable to commit to.

Chapter 1, Section 3.5 (“Citizens’ rights”) rejects all types of “systematic scoring by government,” to which “citizens should never be subject.” Many scoring systems have long been widely used throughout the EU, such as for credit ratings, and should not be dismissed out of hand. For example, most educational systems, including in EU member states, use scoring systems, and these scores may be biased by the judgment of a teacher. Yet AI systems could reduce the level of subjectivity in grading assessments and other types of scoring systems. To be sure, governments can abuse such systems, as the Chinese government is doing with its social credit scoring system. But that should not be used as an attack on the technology any more than steel technology should be criticized because totalitarian regimes use steel to build prisons holding political prisoners.

This section also suggests that AI systems only “hold potential” in terms of how they can “improve scale and efficiency of government in the provision of public goods and services to society.” Yet there are many examples of how government are using AI systems effectively, and there is widespread agreement among AI experts that these systems will be even more impactful going forward.

The introduction of section 4 includes vague language such as “in particular situations” or “Given the potential of unknown and unintended consequences of AI.” This should be clarified.

Chapter 1, Section 4 ("Ethical Principles in the Context of AI and Correlating Values") provides a number of potential ethical principles but does not elaborate on how organizations are already using AI for these goals. For example, the HLEG writes that "AI systems can be a force for collective good" but gives few details on this under its description of "The Principle of Beneficence: 'Do good.'"

Other principles, such as "The Principle of Non Maleficence: 'Do no Harm'" which states that "AI systems should not harm human beings," are aspirational, but unrealistic. For example, if an organization truly abided by this principle to never cause harm, it could never use AI to eliminate a particular worker's job, even if on net workers came out ahead through higher living standards, or use AI in for autonomous vehicles that might result in human injury, even if on net there were many fewer accidents and injuries.

Similarly, "The Principle of Autonomy: 'Preserve Human Agency,'" provides no explanation of how a "right to opt out and a right of withdrawal" can work in practice for certain uses of AI, such as facial recognition, where individuals may not have an interface to the technology. The draft guidelines are also vague about what it means to have "a right to decide to be subject to direct or indirect AI decision making," or what qualifies as an "indirect" decision. This also sets up a false comparison as there are a vast array of situations in Europe where individuals are subject to decisions where they do not know the reasons behind a decision (e.g., being accepted to a college, obtaining a job, getting a loan, etc.).

Similarly, in "The Principle of Justice: 'Be Fair,'" the directive that data practices be aligned with "individual or collective preferences" is quite possibly unachievable, as there are as many preferences as there are individuals, and the collective preferences may not reflect individual ones. Likewise, this principle says that "the positives and negatives resulting from AI should be evenly distributed" which again may be aspirational, but not a standard that can be perfectly achieved and one that is not expected for human-led processes.

Finally, "The Principle of Explicability: 'Operate Transparently,'" overemphasizes the importance of auditability and explainability, even those these requirements can limit the use of more accurate algorithms and undermine attempts to protect intellectual property by forcing companies to disclose source code. Having to explain the logic behind algorithmic decisions to as broad an audience of users as possible is an impractical requirement that could compel companies to make trade-offs between accuracy and interpretability of their computer models. This section also fails to acknowledge the important research advances that might allow future AI systems to provide explanations. For example, the U.S. Defense Advanced Research Projects Agency (DARPA) is investing heavily in its Explainable AI program to spur breakthroughs in machine learning techniques that could explain themselves or be more interpretable by humans without sacrificing performance. Explainable AI would be enormously beneficial for applications

ranging from judicial decision-making to medical diagnostic software, and would alleviate pervasive concerns about the potential for AI to be biased and unfairly discriminate. Rather than call for companies to use explainable AI before it has been fully developed, the HLEG should call for more research in this area and limit requirements for explainable AI to instances where accuracy is not more important.

Chapter 1, Section 5 ("Critical concerns raised by AI") acknowledges that "our understanding of rules and principles evolves over time and may change in the future." This point is important, and since rules and principles are not timeless, the EU should be cautious about imposing static regulations on such an early and dynamic technology. Rules, principles, and concerns will likely change in the future, but regulations tend to lag behind technological developments. Therefore these guidelines should not mandate strict government standards. The HLEG should also clarify whether there may be any consequences for those organizations that do not choose to endorse these guidelines. It should also refrain from using language such as "requirements" given that this is intended to be a voluntary set of guidelines.

Chapter 1, section 5.1 ("Identification without consent") refers to facial recognition as an example of "involuntary methods of identification using biometric data" and recommends overhauling the mechanisms through which consumers give consent, arguing that they are ineffective because "consumers give consent without consideration." First, facial recognition is not always involuntary, and so the guidelines should be updated to clarify this point. Second, it is not practical for consumers to give consent to many uses of facial recognition, such as when it is being used in a public place for public purposes, so that should not be the standard.

Chapter 1, Section 5.2 ("Covert AI Systems") suggests that AI systems are necessarily risky and therefore people should have a right to know when they are interacting with them. As this requirement presupposes that AI systems pose some kind of inherent risk, it should be eliminated, and the guidelines should explicitly avoid rules that discriminate against the use of AI systems. Moreover, such a requirement will seem anachronistic in a decade or two when AI is used to improve significant parts of people's daily lives.

Chapter 1, Section 5.4 ("Lethal Autonomous Weapon Systems (LAWS)") raises concerns over the "unknown number of countries and industries" which are actively "researching and developing" LAWS. Europe should begin to understand the potential military applications of AI. Rather than sitting back while other countries explore these uses of AI, Europe should work to understand its potential use by and against adversaries, especially to protect its infrastructure and strategic interests. But decisions about whether to pursue LAWS should not be part of the HLEG's mandate as it encompasses many broader questions about regional and national security that are outside the area of focus of the HLEG members.

To that end, the guidelines should avoid conflating the broader debate about AI ethics with calls for banning "killer robots." That may be an important debate, but it is almost completely separate and distinct from the one about how AI will impact Europe's economy and society as a whole. Should policymakers succumb to baseless fears that military AI research will lead to a dystopian world full of rogue systems taking over the world, it will set back important AI research poised to deliver many benefits to Europeans. Debating how nations should govern and use autonomous weapons has its place in policymaking, but the HLEG should be careful to recognize that this technology is not just about "killer robots." By comparison, policymakers in the early 20th century did not conflate debates about the internal combustion engine with questions about using that technology to power military tanks. Sabotaging important AI research that can serve the public good as a means of avoiding confronting these issues head on is counterproductive and will harm innovation.

The HLEG states that it will add a final section (5.5) to explore "Potential longer-term concerns." The draft guidelines note that this section is "highly controversial" within the HLEG itself. Given that these concerns are so speculative as to be closer to science fiction than science, such as positing risks from AGI, they should be excluded from this report. As noted by AI expert Max Versace, CEO of robotics and computing company Neurala and founding director of the Boston University Neuromorphics Lab, "The likelihood of an AI scientist building Skynet is the same as someone accidentally building the space station from Legos." If the HLEG decides to include these purely speculative long-term risks, then it should also include a similar section outlining the potential long-term benefits of unforeseeable advances in AI.

CHAPTER II: REALISING TRUSTWORTHY AI

The second chapter of the draft guidelines attempts to map the general principles of the first chapter into concrete requirements for the development and use of AI systems, and suggest a number of technical and non-technical methods to this purpose. But there are several problems with this section.

First, some of the requirements to embed ethics within the design and development of AI systems would be unnecessary and counterproductive. Contrary to what the guidelines suggest, the developers and designers of AI applications cannot always be held responsible for ensuring equality and equity in the use of their technologies. AI is a multipurpose tool, and the ones who should be responsible for ensuring its appropriate use are the operators who deploy the technology. Should there be any oversight, it should be built around algorithmic accountability—the principle that an algorithmic system should employ a variety of controls to ensure the operator can verify algorithms work in accordance with its intentions and identify and rectify harmful outcomes.

Second, requirements to have humans review certain algorithmic decisions raise the labor costs of using sophisticated AI systems which offer better accuracy. As a result, a right to human review

of algorithmic decisions will force companies to use less accurate AI systems that may actually increase bias. Due process and scrutiny should always be appropriate to the nature and seriousness of the decision at hand, and not be based on whether the decision was made by a human or an algorithm.

Third, the guidelines' methods recommend relying on human decisions to solve the "limitations" and "biases" of AI. This incorrectly portrays AI as inherently biased and human ones as unbiased. Yet human decisions are often less accurate, more arbitrary, and more susceptible to bias than algorithmic decisions—which is the reason why many organizations choose to adopt AI systems in the first place. Humans are also far more like "black boxes" than are algorithms, which heightens the folly of subjecting human decisions to lesser scrutiny than algorithmic decisions. In most cases these systems are less biased than human decision making, where subconscious or overt biases permeate every aspect of society. It is certainly true that AI systems, like any technology, can be used unethically or irresponsibly. And combating bias and protecting against harmful outcomes is of course important. But those who resist AI based on this concern fail to recognize a key point: AI systems are not independent from their developers or the organizations using them. If an organization wants to systematically discriminate against certain groups, it does not need AI to do so. A more constructive approach would be to recognize that human decision-making is subjected to less scrutiny than AI yet operates within "black boxes" of its own and greater use of AI could mitigate some human biases.

Fourth, with respect to privacy, the HLEG fails to identify opportunities to use AI to increase individual privacy, such as by automating certain processes that would otherwise require an individual to reveal personal information to another individual. AI offers an important opportunity to increase privacy, and the HLEG should identify some of these opportunities where AI has a net positive impact on consumer privacy and encourage those uses.

Fifth, the use of broad language and unclear terms is concerning. For example, the guidelines (see Chapter 2, Section 1.7, "Respect for Privacy") mention the importance of companies fully complying with the GDPR "as well as other applicable regulation dealing with privacy." The HLEG should specify which other applicable regulations the guidelines are referring to so as not to leave this open-ended to possibly include future regulations or ones in other countries. The HLEG states that adoption of these guidelines should be voluntary, but the guidelines recommend "formal" mechanisms, frameworks, constraints, procedures, and regulation. Moreover, the guidelines include references to "requirements" which could suggest there would be consequences to non-endorsement and non-adherence.

Finally, the guidelines call for transparency and explainability, but make no distinction between the two. The two terms are commonly conflated in discussions about governing algorithms, and the guidelines reflect this particular misunderstanding as well, as they define "explainability—as a form of transparency." Transparency refers to disclosing an algorithm's code or data (or both),

while explainability refers to the concept of making algorithms interpretable to end users, such as by having operators describe how algorithms work or by using algorithms capable of articulating the rationales for their decisions. The guidelines should clarify this distinction. Moreover, while transparency and explainability are fundamentally different concepts, they share many of the same flaws as a solution for regulating algorithms. In particular, they hold algorithmic decisions to a standard that simply does not exist for human decisions. If an evaluation of their decision-making process happens at all, humans are rarely asked to explain is prior to the decision. In addition, mandating that companies make their propriety AI software publicly available would prevent companies from capitalizing on their intellectual property and future investment because other companies would simply copy their algorithms. Similarly, requiring explainability will limit the use of AI in Europe, and thus related investment, which will likely slow down research dedicated to this purpose. As a result, these guidelines could paradoxically act against their own advice by slowing research into AI.

CHAPTER III: ASSESSING TRUSTWORTHY AI

The third chapter provides a list of questions to guide developers when designing AI systems, and to help them assess whether these comply with the requirements and ethical principles of "trustworthy AI." The use cases that will illustrate how this would work in practice will be provided in the next iteration of the guidelines and will be helpful to evaluate whether these questions make sense.

Based on the comments and observations offered for the previous chapters of the guidelines, several questions could be refined or deleted.

For the requirement "Accountability," the first question "Who is accountable if things go wrong?" is too broad and points to AI as holding intrinsic risks. Why not ask "Who is accountable if things go right?" Moreover, "wrong" is Manichean language that is not adapted to the way businesses make decisions, measure risk, and assess results.

The guidelines encourage organizations to consider "diversity and inclusiveness" policies when recruiting staff working on AI. This is an important element. In many EU countries, such policies are compulsory, but not always efficiently implemented. Yet given the skills available in Europe may not match the demand and the needs for the development of AI and diversity, this may not always be practical for a business. Therefore, it cannot be yet a reliable measure of accountability, and this question should be positioned under another requirement, such as "Non-discrimination" or "Design for all."

The guidelines ask "Has an Ethical AI review board been established?" While some companies may choose to use review boards, there is no evidence that this should be a standard. Moreover, this framing suggests that organizations can and should put in separate accountability mechanisms for uses of AI as opposed to other technologies or processes. AI is likely to be deeply

integrated into organizations, and it will likely not be possible to always treat AI accountability and ethics questions separate from other organizational accountability and ethics questions.

The draft oddly categorizes "ethical oath" as a skill and knowledge, according to another question listed under "Accountability."

For the requirement "Data governance," the question "Who is ultimately responsible?" implies that organizations can easily and clearly determine who may be liable, which may not always be the case. It would be relevant to add some elaboration to this question, such as "Who is ultimately responsible for X part of process Z?"

For the requirement "Respect for (& Enhancement of) Human Autonomy," the requirement for businesses to offer users the possibility to "interrogate algorithmic decisions in order to fully understand their purpose, provenance, the data relied on, etc." may be impractical for many organizations. In addition, the HLEG wrongly associates "risks to mental integrity" with "nudging." "Nudging" remains undefined, and could broadly include any recommendation, therefore including it in the list is not appropriate or practical for an assessment.

CONTACT

For questions about these recommendations, please contact Daniel Castro (dcastro@datainnovation.org) or Eline Chivot (echivot@datainnovation.org).