EU WHITE PAPER ON ARTIFICIAL INTELLIGENCE

SUBMISSION TO THE EUROPEAN COMMISSION

BY IBM

JUNE 2020

IBM is the largest technology and consulting employer in the world, with over 350,000 employees serving clients in 175 countries. Today, 47 of the Fortune 50 companies rely on the IBM Cloud to run their business and IBM Watson enterprise AI is deployed in more than 20,000 engagements. IBM is one of the world's most vital corporate research organizations, with 26 consecutive years of patent leadership.

With more than 100 years of commitment in Europe, IBM is one of the largest technology employers in the EU and has many cloud data centres, research labs, innovation spaces, centres of excellence, etc. spread across Europe. IBM scientists from 50+ nationalities work in Europe on cutting-edge research and IBM will build Europe's first quantum computer in Germany.

IBM's expertise is in the intersection of technology and business, providing artificial intelligence (AI) and cloud-based solutions that are changing the way the world works. Above all, guided by principles for trust and transparency and support for a more inclusive society, IBM is committed to being a responsible technology innovator and a force for good in the world. For more information, visit www.ibm.com.

INTRODUCTION

IBM welcomes the opportunity to contribute to the European Commission's consultation on their February 2020 White Paper on Artificial Intelligence, and to offer our views on the measures we believe can help ensure responsible development and deployment of AI systems.

It is rare that a technology attracts the level of attention that AI has from policy-makers, business, academics, media and the public, especially at a relatively early stage in its adoption. That it has done so reflects both the enormous positive potential that all stakeholders recognize in AI, as well as the many concerns – some well-founded, others perhaps less so.

The European Commission has been to the forefront of global efforts to understand and assess the risks and the benefits that AI offers, and to set out a public policy framework that balances the management of those risks with promoting the innovation and uptake necessary for the benefits to be realized. We welcomed the ethical principles developed by the Commission's High-Level Expert Group on AI last year, as well as the OECD's AI Principles, and we continue to be active participants in multistakeholder dialogues seeking to address the issues surrounding trustworthy AI.

Principles are important to help communicate a government's or a company's commitments to citizens and consumers, and to set a direction in a complex and evolving area. However, it is now time to move from principles to clear policies that all stakeholders can rely on. For that reason, we welcome the Commission's AI White Paper and this consultation process as an important step towards operationalizing the high-level principles already established.

RESPONSE TO THE EUROPEAN COMMISSION'S WHITE PAPER ON AI

The White Paper correctly emphasizes the paramount importance of trust, both for companies building and deploying AI, and anyone making use of this technology. IBM has been a vocal supporter of responsible data stewardship and privacy protection for many years, and this strongly informs our approach to AI. Our Trust and Transparency Principles[1] describe our commitment to the belief that

1. The purpose of AI is to augment – not replace – human intelligence.
2. Data and insights belong to their creator.
3. New technology, including AI systems, must be transparent and explainable.

On the basis of these principles, IBM supports targeted policies that would increase the responsibilities for companies to develop and operate trustworthy AI. As the White Paper recognizes, an effective governance framework must first of all be risk-based, and it must seek to regulate not the technology itself but rather its uses. We strongly agree with this overall approach, and believe it is consistent with what we set out in our January 2020 policy paper on AI regulation[2], where we proposed that a risk-based governance framework for AI should be based on the pillars of accountability, transparency, fairness and security.

In the sections below we provide comments on the contents of the White Paper, following its section numbering. Our main points in relation to the proposed regulatory framework for AI are:

- For regulatory purposes the definition of AI systems needs to be narrow and clear, so that the focus is on those systems that cause serious concern and avoiding non-AI systems being inadvertently included.
- There should be a single risk assessment framework to identify high-risk applications regardless of sector, and without lists of exceptions.
- In sectors where established conformity assessment mechanisms already exist, these should also cover high-risk AI applications in those sectors. In other sectors, an appropriate level of compliance for high-risk applications can be achieved with a combination of ex-ante self-assessment and ex-post auditing and enforcement.
- While voluntary labelling systems can be helpful to consumers or end-users in some markets, we do not believe they would be effective across such a broad field as AI applications.

[1] "IBM's Principles for Trust and Transparency," *THINKPolicy*, 30 May 2018, available at https://www.ibm.com/blogs/policy/trust-principles/.

[2] *"Precision Regulation for Artificial Intelligence," IBM Policy Lab, 21 Jan. 2020, Ryan Hagemann and Jean-Marc Leclerc: https://www.ibm.com/blogs/policy/ai-precision-regulation/.*

4. AN ECOSYSTEM OF EXCELLENCE

In contrast with some of the negative commentary around Europe's global standing in AI, we see a Europe that is well-placed both to take advantage of widespread AI adoption, and to contribute strongly to the development of the underlying science and technologies. The approach set out in the "Ecosystem of Excellence" section of the White Paper will help to further strengthen that position and help to build the capacities needed to drive successful adoption and further development of AI in Europe, including through promoting research and innovation, developing skills, coordinating national policies and initiatives, encouraging adoption by the public sector, and building strong public-private partnerships.

In particular, we support the emphasis on skills, where AI needs to be reflected appropriately throughout all levels of the formal education system, including in vocational training and apprenticeships, as well as in targeted re-skilling and lifelong learning initiatives. In addition to the long-standing emphasis on Science, Technology, Engineering and Mathematics (STEM) skills and competences, it is important to recognize that arts and creative disciplines are also necessary for a vibrant AI ecosystem. There is also significant potential for AI to drive improvements in the delivery of education across all domains.

Public/private partnerships can be valuable tools for encouraging strong private sector involvement in developing an ecosystem of excellence and ensuring a broad-based and coherent approach. However, to be effective, partnership models should remain open to participation by any company that can provide relevant capabilities and that complies with European regulations and values, even if they are headquartered outside the EU.

The potential for AI to significantly improve the delivery of public services for the EU's citizens is significant, despite the particular challenges that many public services face in adopting AI. The Commission is right to focus on actions to address this, including developing technical capabilities within public bodies and agencies, ensuring public servants are familiar with AI's implications, and that anyone using AI systems in the public service is appropriately trained.

As the European Data Strategy identifies, there is enormous potential in public sector data to improve the lives of EU citizens. Unlocking that value should be a key focus, including through making non-sensitive public sector data available for research, and by establishing rules around permitted usage of public sector data by private enterprise.

5. AN ECOSYSTEM OF TRUST: REGULATORY FRAMEWORK FOR AI

We agree with the need for a consistent EU-wide regulatory framework for trustworthy AI, and that this would help give businesses and consumers alike the confidence to develop and adopt AI-based solutions.

Any successful governance framework for AI will need to account for the many and varied applications of AI both in use today and likely to be used in the future. The best means of striking an appropriate balance between effective rules that protect the public interest and the need to promote ongoing innovation and experimentation is with a precision regulation approach that creates expectations of accountability, transparency, fairness and security according to the *role* of the organization (whether a provider, owner, or some mix of both) and the *risk* associated with each use-case of AI.

5A. PROBLEM DEFINITION

The White Paper focuses on the challenges that AI poses for the application of rules protecting fundamental rights, and addressing safety and liability, noting that *"opacity ('black box-effect'), complexity, unpredictability and partially autonomous behaviour"* create challenges for verification of compliance and effective enforcement of existing EU law meant to protect fundamental rights. However, it is also true that human beings using AI (or not) and governed by existing bodies of law pose similar challenges.

In that light, we suggest that the degree to which autonomy and the judgment of a human actor are ceded to an AI system should be a key factor in determining the degree to which AI poses new problems to be addressed by new regulation. Where a person receives the output of an AI system as one factor in making a decision for which that person retains responsibility, while being given sufficient information about the training and functioning of that system to make reasonable judgements about its utility, current laws should suffice. Where, however, human autonomy or judgment are substantially ceded to an AI system, whether by choice (for example, automatic reviews of loan applications) or by the nature of the activity (for example, autonomous driving), then the problem at hand might be fairly characterized as new and unique.

5B. POSSIBLE ADJUSTMENTS TO EXISTING EU LEGISLATIVE FRAMEWORK RELATING TO AI

Even though much of it was written without having digital technologies explicitly in mind, the EU's existing legal framework already applies to AI applications, including for example, fundamental rights protections, the General Data Protection Regulation (GDPR), product safety, and consumer protection regulations. The General Product Safety Directive (GPSD) has a proven track-record as a 'safety net' complementing sector-specific regulations. In particular, the technology-neutral approach to formulating requirements for safe products has proven to be

effective with new challenges like interconnectivity, cyber threats, human-machine-interaction, AI, etc.[3]

Effective application and enforcement of existing legislation

It is true that in some cases there is a lack of clarity about exactly how these existing frameworks would apply or be enforced in the context of AI solutions, so there is a need for assessment and for clarification or guidelines, particularly for high-risk use cases. However, new legislation should only be considered for specific high-risk cases where it is determined that existing frameworks cannot be adequately applied, clarified or adapted. Any new legislative proposals on AI, or in related areas (such as the draft e-Privacy Regulation), should be consistent with existing legal frameworks to avoid diverging rules and legal uncertainty, which could have a negative effect on innovation and the uptake of AI within the EU. They should also clearly specify what additional risks any new legislation is seeking to address.

Changing functionality of AI systems

The White Paper notes that software updates during the operational lifetime of an AI system could change its functionality, and potentially its risk profile. (Note that the use of machine learning does not necessarily imply that a system continues to learn, or evolve its operation, during its operational lifetime. In many cases an AI model remains static once deployed, unless changed by a subsequent software update.)

However, the existing legal framework already addresses this issue, and we do not believe changes are required to address AI. Under the existing framework, if a product is changed significantly once placed on the market, the manufacturer must undertake a new risk assessment. This could be triggered by a modification of the intended use or reasonably foreseeable use of the product, a change in the nature of a hazard or an increase of the level of risk. This is a well-established and accepted practice, with for example the Blue Guide stating in chapter 2.1: *"A product which has been subject to important changes or overhauls aiming to modify its original performance, purpose or type may be considered as a new product. The person who carries out the changes becomes then the manufacturer with the corresponding obligations."*

---

[3] *For example, this recent RAPEX notification published on the EU Safety Gate's website shows that market surveillance authorities were able to take legal action under the GPSD in relation to cybersecurity shortcomings in a smartwatch for children:*
https://ec.europa.eu/consumers/consumers_safety/safety_products/rapex/alerts/?event=viewProduct&reference=A12/0157/19&lng=en

This concept fits appropriately in a world of software-enabled products where a software update can apply important changes to the product. Note that software updates often consist of purely bug fixes, in which case there should be no obligation to undertake a re-assessment.

Allocation of responsibilities between different economic operators

The use of AI systems, and therefore any resulting liability, is context-specific: the focus of risk should lie on a specific application and the context of its use. There is often a complex chain of various producers and intermediaries involved - for example, the various technology producers (of data, software, hardware components, or physical products with embedded software), the systems integrators (who train the AI system) and the owner/operator who is using the system. This is why a future-proof regulatory framework should ensure allocation of liability to the actor who is closest to the risk, instead of introducing joint liability, as it would not make sense for anyone involved in making an AI system to be liable for problems they had no awareness of or influence over.

In a business-to-business context, parties can negotiate for an efficient allocation of risk which takes account of each specific context and use. Contractual liability is working well and should therefore be maintained. Any changes to the liability framework should be consistent with the scope of the Product Liability Directive.

We disagree with the idea of expanding the definition of "product" in the Product Liability Directive to include software. It is difficult to envisage how standalone software could result in property damage, bodily injury or death. Generally speaking, services will still require a physical infrastructure in their execution, therefore physical products remain the basis for the guidance and the application of the Directive. In most cases the relationship between provider and end-user is already covered by a contractual relationship, while services that are inherently dangerous or pose specific risks to the users are usually already regulated and subject to insurance (e.g. healthcare or legal services).

5C. SCOPE OF A FUTURE EU REGULATORY FRAMEWORK

We strongly agree with the Commission's view expressed in the White Paper that *"the definition of AI will need to be sufficiently flexible to accommodate technical progress while being precise enough to provide the necessary legal certainty"*. Much work has been done, for example, by the EU High Level Expert Group and by ISO/IEC SC42 on defining AI and related terms for ethical and technical purposes, but for the regulatory context we suggest that an extension of the approach taken by the OECD might be suitable, for example:

> *An AI system can be broadly defined as one that makes predictions, recommendations or decisions, influencing physical or virtual environments, and whose outputs or behaviours are not necessarily pre-determined by its developer. AI systems are typically trained with large quantities of structured or unstructured data, and may be designed to operate with varying levels of autonomy or none, to achieve human-defined objectives. ("Autonomy" means acting, physically or virtually, without human intervention or oversight.)*

This definition is intended to focus on the features most likely to distinguish AI from non-AI systems. However, given the dynamic nature of the field, it is not possible to come up with a perfectly-bounded definition – emphasizing the importance, for regulatory purposes, of focusing on specific use-cases and the risks associated with them.

The Commission's White Paper identifies the features of AI systems that are of concern because they could make the application and enforcement of existing regulations difficult – transparency/opacity, traceability, and human oversight. These features are particularly associated with AI systems developed using certain machine learning techniques and not, for example, those using rules-based or symbolic approaches to AI. For that reason, we believe the Commission should consider focusing any new regulation specifically on applications that depend on these machine learning approaches.

The Definition of High-Risk

We agree with the Commission's view expressed in the White Paper that *"A risk-based approach is important to help ensure that the regulatory intervention is proportionate"*, and in particular that *"it requires clear criteria to differentiate between the different AI applications, in particular in relation to the question whether or not they are 'high-risk'."*

We believe that the approach proposed in the White Paper, of explicitly listing sectors where high-risk applications might occur, would be problematic in practice.

- Since the technology is evolving rapidly and the application of AI becoming more and more pervasive, it is difficult to definitively rule out any sector from potentially having high-risk applications, and incorrect to assume that every use of AI in particular sector is high risk.
- It is likely that any explicit list of sectors will have to be frequently updated, leading to uncertainty and issues over the retrospective application of laws.

- While it is true that in certain sectors the use of sensitive data is more common and therefore a higher risk might exist, this higher risk applies to any application used in these sectors, whether AI-driven or not.

Instead, we believe that focusing solely on the second criterion proposed in the White Paper would be a more effective approach, provided that there is clarity about how risk is to be assessed. We suggest that the key factors to consider should be the degree to which human judgment and agency is replaced (autonomy) and the risk of negative impact of the application on human lives (severity and likelihood).

- Autonomy:
  It is IBM's position that AI should generally augment not replace human decision-making. In practice, there will be a spectrum of levels of autonomy for different applications, and we believe the degree of human involvement in decision-making should play a large role in determining whether an application is considered high-risk. For example, if a doctor uses a clinical decision-support tool which lays out potential treatment options in a context which gives the doctor reasonable information about the tool, discretion to ignore the recommendation and time to make his/her own decision, that should not necessarily be considered a high-risk application, even though it's in the medical domain. In other words, in certain use cases the risk to the eventual subject of a decision remains largely in the hands of the human user, rather than with the AI element of a decision-support system.

- Severity and Likelihood:
  In cases where an AI system has significant autonomy from human intervention or oversight, we agree with the overall approach in the White Paper that an application could be considered high risk when "*significant risks are likely to arise*" and there is potential to cause significant impact on the affected parties.
  In considering when significant risks are likely to arise, however, both the severity of the harm and the likelihood of it occurring need to be taken into account. For example, there may be situations when even a very low likelihood of a risk occurring constitutes "high-risk" because the severity of harm is so high.
  In considering what constitutes "significant impact", we believe the emphasis should be on the most severe potential impacts: risk of injury, death or material damage over a reasonable threshold. "Immaterial damages" such as loss of privacy, limitations to the right of freedom of expression, human dignity, or discrimination are highly important issues, and AI systems should certainly be required to respect the extensive legal protections (such as privacy legislation, consumer protection, etc.) already in place in these areas.

Exceptional Instances

In the White Paper, the two-criteria approach is complicated by the proposal that the use of AI for certain purposes would be always considered as high-risk. Examples of such exceptional purposes are given: the use of AI applications for recruitment processes; use in situations impacting workers' rights; and the use of AI for the purposes of remote biometric identification. We believe this is both unnecessary and problematic, as these examples are defined in an open-ended way, making the scope unjustifiably broad and unpredictable. As set out above, a clear approach to risk assessment should be sufficient to identify all high-risk cases without the need for any lists or exceptions.

In particular, we would argue that the use of AI technologies in an area such as workforce management does not per se qualify an application as high-risk. To do so would amount to regulating the technology rather than the use of the technology. Naturally, the use of AI technology in the employment environment could raise concerns about bias, control or monitoring. However, AI solutions also offer significant benefits to workers, including reducing the effect of human biases, providing customized insights about potential jobs or careers, or personalized training paths. This reinforces why it is paramount to identify the specific risk foreseen from the use of AI in a particular context rather than risking the exclusion of the employment sector from the potential benefits of AI. For example, it is already the case in some European countries, that the implementation of standard (non-AI) software solutions is subject to consultation with employee representatives. There is little evidence of the need for additional regulation in an area where there are significant existing controls (e.g. under GDPR).

We would also clarify that, to avoid over-regulation, the definition of high-risk and any regulatory requirements flowing from that should not apply to AI systems during their research and development, but only to those deployed or placed on the market, given that any risks will only occur during operational use and not in the research and development phase.

Finally, it is important that as any risk assessment framework evolves there is strong and practical guidance provided for companies so that there can be clarity, consistency and transparency about whether applications will be deemed to be high-risk and why.

5D. TYPES OF REQUIREMENTS

The White Paper outlines the kinds of legal requirements that could be imposed on AI actors in relation to high-risk applications, under each of the subheadings below.

a) Training Data

We fully support the aim to ensure that any high-risk AI solutions developed or deployed in Europe should reflect European values, rules, and citizens' rights. However, we do not

believe that this aim can be achieved by placing prescriptive requirements on training data. As the White Paper acknowledges, the focus should be on the outcomes of the system.

It is reasonable to place a requirement on the relevant actor responsible for the training of a high-risk application to ensure a specific outcome, for example, the absence of discrimination. However, input-specific requirements, such as those that need to be considered in selection of the training data (appropriate quality, diversity, lack of bias etc.), the processing of that data, the training of the model, should not be prescribed in regulation, since the relevant technologies and state-of-art are evolving rapidly, and existing laws may already provide sufficient coverage.[4]

b) Keeping of records and data

For high-risk applications we agree that developers and operators should be required to maintain relevant information to enable possible subsequent investigation of problematic outcomes, by competent authorities. That could include information about the development process, characteristics of the training data, algorithms, testing methodologies and results. In some cases, it may be justified to retain the actual training data. However, adequate protection must be provided for confidential and commercially sensitive information.

c) Information provision

In addition to information retained for possible use by competent authorities, as outlined above, it is appropriate to disclose certain information about high-risk applications to end users or members of the public. It is also important to provide for appropriate information-sharing between other actors in the AI supply chain since, for example, the developers of AI systems often draw upon AI services from other organizations, typically through an Application Programming Interface (API).

IBM has proposed the use of FactSheets[5] as a general approach to AI transparency. A FactSheet is a collection of relevant information about an AI model or service that is created during the machine learning lifecycle. Given the diversity of AI application domains and model types, a single FactSheet template or schema is not realistic, but could include:

---

[4] *IBM has created tools to help address bias, such as AI Fairness 360, an open source software toolkit that can help detect and remove bias in machine learning models: https://developer.ibm.com/technologies/artificial-intelligence/projects/ai-fairness-360/*

[5] *"Factheets for AI Services", IBM Research Blog, August 2018, Alexandra Mojsilovic: https://www.ibm.com/blogs/research/2018/08/factsheets-ai/*

- Information from the business owner (e.g. intended use and business justification);
- Information from the data gathering/feature selection/data cleaning phase (e.g. data quality, features used or created, cleaning operations);
- Information from the model training phase (e.g., bias, robustness, and explainability information);
- Information from the model validation and deployment phase (e.g., key performance indicators).

d) Robustness and accuracy

We agree that trust in AI systems will depend critically on them demonstrating technical robustness and accuracy in terms of their outputs. Robustness must also take into account cybersecurity issues and the possibility of adversarial threats.[6] As with non-AI systems, it is appropriate to hold relevant actors in the supply chain for high-risk systems accountable for the robustness and accuracy of such systems. Any regulation should steer clear of mandating specific technical approaches to achieving this, given that technology-neutral regulation has proven to be more effective and adaptable and therefore able to foster innovation.

e) Human oversight

As outlined earlier, we submit that the lack of informed and empowered human oversight should be a key factor in determining whether an application should be deemed a high-risk application. In some cases, the lack of human oversight creates *incremental risk* that AI applications will produce significant effects for the rights of an individual or a company, e.g. risk of injury, death or significant damage, *over and above* the risks raised by unaided human activity that are already addressed through existing law. There could be situations where an AI system may make better decisions without a human involved, though with less transparency and accountability.

On that basis, we agree with the principle set out in the White Paper that it would be appropriate to require the operator of an AI system to ensure a human can intervene, override or revise its output in a high-risk context. The details, however, will depend on the specific use case and the risk assessment.[7]

---

[6] *For example, IBM has developed Adversarial Robustness 360 Toolbox, an open source software library to help both researchers and developers defend deep neural networks against adversarial attacks: https://developer.ibm.com/technologies/analytics/projects/adversarial-robustness-toolbox/*

[7] *IBM has created tools to help open up the "AI blackbox" and improve explainability. For example, Watson OpenScale tracks and measures outcomes from an AI system across its lifecycle, explaining how recommendations are being made and detecting and mitigating bias: https://www.ibm.com/cloud/watson-openscale*

f) Specific requirements for remote biometric identification

We welcome the distinction made in the White Paper between biometric identification and biometric authentication/verification. This is the kind of precise definition of use cases that is required to ensure regulation is appropriately targeted. We also welcome the intent to launch a broad European debate on the specific circumstances, if any, which might justify the use of remote biometric identification in public places, and on common safeguards.

In a June 8 letter[8] to members of the United States Congress, IBM CEO Arvind Krishna stated:

> *"IBM no longer offers general purpose IBM facial recognition or analysis software. IBM firmly opposes and will not condone uses of any technology, including facial recognition technology offered by other vendors, for mass surveillance, racial profiling, violations of basic human rights and freedoms, or any purpose which is not consistent with our values and Principles of Trust and Transparency. We believe now is the time to begin a national dialogue on whether and how facial recognition technology should be employed by domestic law enforcement agencies."*

5E. ADDRESSEES

We strongly agree with the Commission's view that in any future regulatory framework, legal obligations should be imposed on the actor(s) in the AI supply chain best placed to address any potential risks.

Over the course of an AI system's lifecycle, many organizations will play varying roles in the system's development and operation. They may contribute research, the creation of tooling, or APIs; in later stages, organizations will train, manage, and control, operate, or own the AI models that are put to real-world use. These different functions may allow for a distinction between "providers" and "owners/operators," with expectations of responsibilities based on how an organization's role falls into one or both categories.

Differentiating between providers and owners/operators can help to better direct resources and oversight to specific applications of AI based on the severity and likelihood of potential harms

---

IBM AI Explainability 360 is an open source toolkit that helps developers provide explainable AI solutions: https://developer.ibm.com/technologies/artificial-intelligence/projects/ai-explainability/

[8] *IBM CEO's Letter to Congress on Racial Justice Reform, June 8 2020: https://www.ibm.com/blogs/policy/facial-recognition-susset-racial-justice-reforms/*

arising from the end-use and user of such systems. Risk-based, precision regulation approaches like this — which also allow for more manageable and incremental changes to existing rules — are ideal means to protect consumers, build public trust in AI, and provide innovators with needed flexibility and adaptability. They will also be more resilient and adaptable as the technology evolves.

In connection with these concepts, we would point back to our comments in Section 5B above, focused on the role of disclosure and informed human oversight in addressing novel risks posed by the adoption of AI.

We agree with the Commission's view on geographic scope – that EU requirements should be applicable to all relevant economic operators providing AI-enabled products or services in the EU, regardless of whether or not they are established in the EU.

5F. COMPLIANCE AND ENFORCEMENT

In sectors where established conformity assessment mechanisms already exist, we agree that these should also cover high-risk AI applications in those sectors. In other sectors, we believe an appropriate level of compliance for high-risk applications can be achieved with a combination of ex-ante self-assessment and ex-post auditing and enforcement, following the successful approach taken for technical regulation in Europe. It is reasonable to require deployers of AI systems to make an initial assessment and disclosure about whether an AI system they propose to implement is high-risk or not, focusing on the impact of the decisions being supported by the AI and the degree to which informed human oversight is being provided.

If conformity assessment by a third party is required, there should be an emphasis on outcome-based testing rather than examination of (potentially proprietary) inputs such as algorithms, code and data. Moreover, any assessment should be done against open standards.

As set out in our comments above, where human judgment is governed by existing standards, we suggest that auditing and enforcement mechanisms focus on evidence that informed human oversight is appropriately established and maintained. Another critical factor for the safe operation of a high-risk system is the appropriate training of the human operator or user.

We do not agree with the suggestion that shortcomings identified might be remedied by re-training a system in the EU or on EU data. The aim of ensuring that AI is trustworthy, secure and in respect of European values and rules cannot be achieved by simply requiring the use of European training data or that the training be done within the EU. The focus should be on the outcomes of the system, and the quality of the training data (including appropriate diversity, lack of bias etc.) rather than its geographical source.

5G. VOLUNTARY LABELLING FOR NO-HIGH RISK AI APPLICATIONS

While voluntary labelling systems can be helpful to consumers or end-users in some markets, we do not believe it would be effective across such a broad field as AI applications. Given the hugely diverse range of AI products and services that will be developed and deployed across all sectors in the coming years, a one-size-fits-all labeling scheme would be too complex to be workable in practice. For these reasons we do not support the proposal to standardize a voluntary labeling system for all non-high-risk AI systems while remaining supportive of innovation and experimentation in this space.

5H: GOVERNANCE

In sectors where established governance structures already exist (medical devices, aviation etc.) we believe these existing bodies are best placed to also cover high-risk AI applications in their sectors. They have the necessary deep sectoral expertise, the operational relationships and track-record with relevant stakeholders. There may be some value in a new European mechanism that provides best practice sharing and guidance across sectors, but its scope must be limited and its relationship to existing regulatory bodies clearly defined, so as to avoid fragmentation, inconsistency and the risk of stifling innovation.

IBM is supportive of the need for multistakeholder forums that allow industry, government, academia, and others to contribute to an evolving set of requirements for standards and best practices that can keep pace with changes in the technological reality of AI. We believe the optimal approach to ensuring adherence to and ongoing improvement of such standards and best practices is one that prioritizes co-regulatory mechanisms. The open and transparent processes of global and European standardization are best suited to support this work.

We would point to a number of existing national frameworks that effectively prioritize the need for voluntary standards and co-regulatory governance of AI, such as Singapore's *Model AI Governance Framework*. In determining how best practices and standards can fit into a broader governance framework that includes regulatory authorities, we would also point to ongoing efforts in the United States – in particular, the Office of Management and Budget's *Guidance for Regulation of Artificial Intelligence Applications.*

Independent auditing of high-risk AI systems may be appropriate in certain circumstances, such as law enforcement applications, provided it focuses only on the operational use of the system and its outcomes, and not on other stages of the system lifecycle. Otherwise, requiring third-party audit of, for example, iterative improvements in AI models and systems at all stages of the development lifecycle, will create an environment in which standards competition is replaced with adherence to static minimum mandates.

CONCLUSION

In conclusion, we welcome the emphasis in the Commission's White Paper on the need for a consistent EU-wide regulatory framework for trustworthy AI, and agree that this will be essential to give all stakeholders the confidence to develop and adopt AI-based solutions, and to realize the enormous benefits that they offer.

Building trust requires acknowledging valid concerns that exist in relation to accountability, transparency, fairness and security, and putting in place appropriate regulatory mechanisms to manage those risks, while continuing to promote ongoing innovation and experimentation – getting that balance right requires a precision regulation approach that is clear and targeted.

We set out our views on this in a recent paper[9] on "Precision Regulation for Artificial Intelligence", including policy proposals for both governments and companies, and believe the approach in the Commission's White Paper is largely consistent with our own. We hope that our comments are useful and look forward to contributing to the debate in the months ahead.

[9] *"Precision Regulation for Artificial Intelligence," IBM Policy Lab, 21 Jan. 2020, Ryan Hagemann and Jean-Marc Leclerc:* https://www.ibm.com/blogs/policy/ai-precision-regulation/.