ETHICAL PURPOSE

SOCIETAL BENEFIT

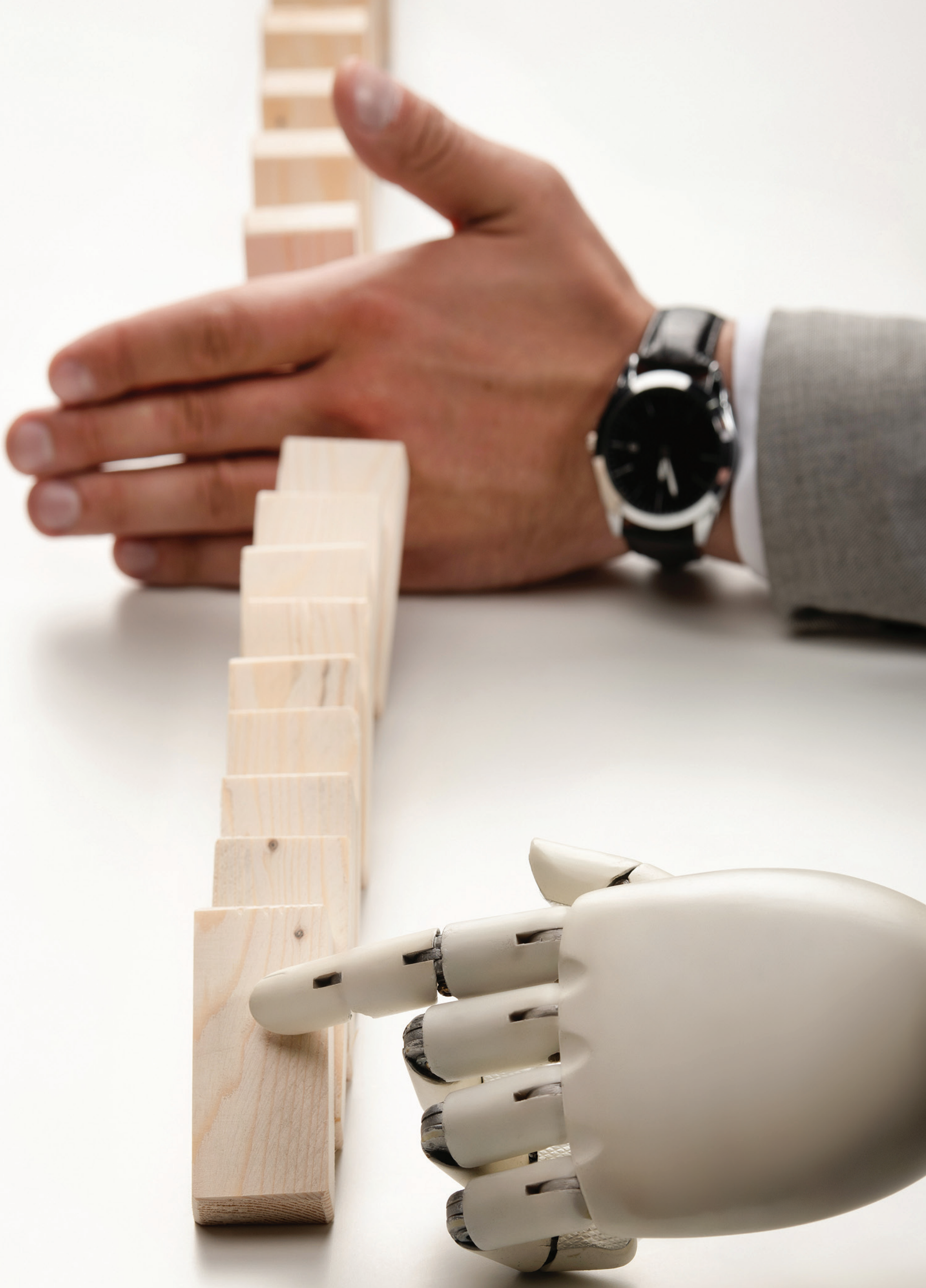
ACCOUNTABILITY

TRANSPARENCY

EXPLAINABILITY

FAIRNESS

NON-DISCRIMINATION



Responsible AI

A GLOBAL POLICY FRAMEWORK

SAFETY

RELIABILITY

OPEN DATA

FAIR COMPETITION

PRIVACY

INTELLECTUAL PROPERTY

The EU AIA

A Green Paper Policy Analysis

ITECHLAW®
INTERNATIONAL TECHNOLOGY LAW ASSOCIATION

The EU **AIA**

A Green Paper Policy Analysis

EDITED BY

John Buyers
Patricia Shaw
Susan Barty

ITECHLAW®
INTERNATIONAL TECHNOLOGY LAW ASSOCIATION

McLean, Virginia, USA

This book does not provide legal advice. It is provided for informational purposes only.

In the context of this book, significant efforts have been made to provide a range of views and opinions regarding the various topics discussed herein. The views and opinions in this book do not necessarily reflect the views and opinions of the individual authors. Moreover, each of the contributors to this book has participated in its drafting on a personal basis. Accordingly the views expressed in this book do not reflect the views of any of the law firms or other entities with which they may be affiliated. Firm names and logos, while used with permission, do not necessarily imply endorsement of any of the specific views and opinions set out herein.

The authors have worked diligently to ensure that all information in this book is accurate as of the time of publication. The publisher will gladly receive information that will help, in subsequent editions, to rectify any inadvertent errors or omissions.

International Technology Law Association
7918 Jones Branch Drive, Suite 300
McLean, Virginia 22102, United States
Phone: (+1) 703-506-2895
Fax: (+1) 703-506-3266
Email: memberservices@itechlaw.org
itechlaw.org

Cover and text design by Troy Scott Parker, Cimarron Design

This book is available at www.itechlaw.org

ISBN-13: 978-1-7339931-2-8

Printed in the United States of America.

Contents

Introduction

It was back in 2018 that the International Technology Law Association started what would be a forensic journey down the path of Artificial Intelligence technology analysis—not only in terms of examining the ethical aspects of what has proven to be a very different and transformative technology, but also considering what the ideal legal position should be when evaluating laws attempting to legislate for such technologies. Our Responsible AI Global Policy Framework and its eight associated principles stand as a formidable achievement in this regard, the culmination of the effort of 27 law firms across the world, many from the EU, now recently updated with a 2021 edition.

The European Union's proposed new law on the regulation of Artificial Intelligence—the so-called Artificial Intelligence Act or "AIA," which is the first major trans-national legislative foray into this space—attempts to regulate the use of AI across the 27 EU member states. It is a boldly innovative measure which has the potential to significantly impact trade with that bloc and may force us to reconsider the manner in which such technologies are created and deployed, much as the GDPR did back in April 2016.

Given the transformative impact of the AIA, we have prepared this Green Paper which evaluates and benchmarks the embryonic legislative measure against our Responsible AI Global Policy Framework. The Green Paper itself is the work of 34 lawyers in 32 law firms across 14 international jurisdictions. As such, it is able to provide a unique global and contextual perspective of the draft regulation.

We recognise at the time of publication that the AIA is going through a legislative process and is liable to be amended and updated. As such, we intend to keep pace with those changes and ultimately update this Green Paper once the measure becomes law. Until then, we hope that it will serve as an informative, thought-provoking and in-depth consultative resource to aid both the potential enterprises impacted by the new law as well as the legislators implementing it.

This inaugural version of the Green Paper is based on the Slovenian Presidency Compromise text of the Artificial Intelligence Act released on 29th November 2021, No. 8115/20. Whilst we are aware that the French Presidency released subsequent partial Compromise texts during the editing of this Green Paper, we have not commented on these as they do not yet cover the entire proposed Artificial Intelligence Act.

How to Use This Green Paper

The Green Paper is subdivided into sections which relate to each of the core Principles in the ITechLaw Responsible AI Global Policy Framework, 2021 Update Edition (referenced in this Green Paper as the "Responsible AI Framework"). As such, it is intended as a companion guide to ITechLaw's Responsible AI Framework, which is available separately from the International Technology Association's website. For convenience, the Principles from the Responsible AI Framework are listed beginning on page 91.

In each case, we have attempted to identify in detail where we believe the AIA is managing to meet the standards set out in our framework, as well as those areas which require further attention or which may be demonstrably missing, and provide recommendations for an improved approach where relevant.

Recommendations in the text are indicated as follows:

→ **Recommendation**

Where we are of the view that there is a significant omission or oversight in the AIA which requires immediate attention, we indicate it as follows:

⚠ **Caution**

We should stress that this publication is prepared in the constructive spirit of attempting to improve what is, by any evaluation, a hugely innovative, ambitious, and complex legislative measure.

Our core assumptions, findings, and recommendations on approach are contained in the Executive Summary.

We welcome feedback on the contents of this Green Paper. If you have any comments or opinions, please do pass your feedback to ITechLaw on the following email address: greenpaper@itechlaw.org.

– John Buyers, Osborne Clarke LLP
Patricia Shaw, Beyond Reach Consulting Limited
Susan Barty, CMS LLP
March 2022

Executive Summary

This Executive Summary is intended to provide a "snapshot" view of the new Artificial Intelligence Act (as reflected in the draft Compromise Text issued by the Slovenian Council Presidency at the end of 2021) and as such picks up on some wide-ranging themes that we find run through our analysis.

A summary of issues based on our eight Responsible AI Framework Principles can be found in Part A of this Green Paper. Part B offers a more detailed review of the AIA. Part C provides the full text of our Responsible AI Framework Principles, 2021 Edition.

We should stress that this publication is prepared in the constructive spirit of attempting to improve what is—by any evaluation—a hugely innovative, ambitious and complex legislative measure. Given its breadth, it is inevitable that there will be "crossover" areas which will apply in relation to several of our principles.

The Need for Further Legislation

First and foremost, we should emphasise that, frustratingly, the EU AIA is not a holistic piece of legislation. What we mean by this is that it relies on a corpus of pre-existing law (such as the GDPR) and will need yet-to-be-determined-laws (such as a measure to enable effective AI liability mechanisms) in order for it to function within the EU *acquis*. We consider it essential that the EU, at the very least, provide a clear legislative road map (and appropriate legislative decision making) to ensure that those likely to be fundamentally affected by this legislative measure are able to understand its far-reaching consequences.

Product Safety Focus

A focus on European product safety and liability and an attempt to piggyback the nascent AI regime on that legislative mechanism is undoubtedly an elegant response to the problem of creating a new regulatory framework without disrupting existing laws. On one level, we can see how this approach fits neatly in relation to the undoubtedly difficult problem of regulating AI systems. We are concerned, however, that in reality it creates fundamental problems in relation to focussing on what we would deem to be the most important issue: that of harm to the individual and harm to groups of individuals (group harm). The EU product-safety regime is inimical to a subjective harm or outcomes-based approach which have been adopted by equivalent measures, such as Article 22 of the General Data Protection Regulation (GDPR). By focussing on product compliance, the EU has made it commensurately difficult to allow for the outcome approach described above to be applied. Necessarily, the EU will be forced down a track which is limited to specific use cases and classes of AI system, whilst avoiding what we would deem to be the single most important element of AI regulation.

Equally speaking, we do not understand why the EU does not mandate a set of core ethical principles to underpin the operation of all EU-wide AI systems. Confining regulation to systems dependent on use case (i.e., whether they are prohibited, High Risk within Annex III, or anything else) seems to us to be an approach that removes the core elements of any effective AI measure—which, at the very least, should be to ensure consistency and minimum baselines in performance. We would like to see the AIA align with the European Parliament Resolution of 20 October 2020 on a framework of ethical aspects of artificial intelligence, robotics, and related technologies.

A product-safety emphasis also means that, inevitably, the focus is on compliance and conformity assessments. We see this in particular in relation to the transparency and explainability measures in the AIA which again are focussed on providing definitive (and essentially binary) explanations to market surveillance authorities and national regulators, rather than on delivering outcomes. Our nervousness stems from the fact that AI technologies, especially those such as deep neural networks, are inherent "black boxes" that cannot be unpacked in the manner envisaged by the proposed law. Again, a different emphasis on acceptable parameters and outcomes (related to what is desirable and what is not) could provide a more technologically sustainable and future-proofed legislative framework.

Citizen Redress and Human Accountability

Following on from this, we find that the Users (by which we mean the EU citizens that may be the subject of AI decision making) are not afforded the rights we would expect to see in such a landmark piece of legislation. In terms of End Users' ability to understand how decisions have been made, there is little transparency and no right to receive Technical Documentation under Annex IV or in relation to their ability to seek individual redress in the event they have been harmed (e.g., through discrimination or biased decision making). In short, the AIA does not provide an adequate route for recourse.

There is a clear and urgent need to enable private individual remedies under the AIA. We suspect (although this has not yet been definitively confirmed by the EU) that either this is being planned under a forthcoming (and unspecified) AI liability measure, or undue reliance is being placed on recourse mechanisms in the GDPR in so far as personal data is concerned.

Related to the above point, we find that the AIA does not take a clear stand in relation to AI and legal personality. It should therefore expand on the need for human oversight by expressly stating that those who can be held responsible for the acts and omissions of an AI system are—and always will be—human.

Surveillance and Military Usage

It is disappointing but perhaps inevitable to see that military-use cases are entirely outside of the scope of the AIA. Whilst we understand that defence is inherently the domain of the EU 27, we would like to see some level of commitment to the regulation of defence systems that are to be exported from the EU.

So far as surveillance is concerned, we would like to see an extension of the measures prohibiting real-time biometric surveillance to all of the agencies of the state and not just law enforcement, as well as the use of overly intrusive surveillance in the private sector. We see this as vital to reinforcing human autonomy and underpinning western democratic ideals.

Framing the AI Ecosystem

Given the complexity of the AI ecosystem, we are disappointed that this is not adequately reflected in the draft AIA. The AIA focuses on the concepts of manufacturer, distributor, importer, provider and users which contemplate an end-to-end supply chain for completely evolved AI systems. This overlooks the growing market in MLops, which is the component- or function-level supply of AI systems, which could (for example) include federated learning. Given the potential for certain actors in the industry to have differing roles in a non–end-to-end environment, we have concerns that this will lead to confusion as to who should "own" the responsibility and risk apportionment and the potential for multiple compliance assessments, as well as increased costs.

Compliance Burden

We find that the draft AIA has an overly complex and overlapping regulatory structure, including national supervisory authorities, notifying authorities and market surveillance authorities. Having multiple authorities means that there will be significant potential to create bureaucracy in relation to approvals, and jurisdictional conflict in relation to differing approaches taken by regulators in different EU nations. This is likely to create real difficulties and cost hurdles for businesses, especially SMEs, that are seeking to enter the European market. In the wider sense, this could also act to limit the overall effectiveness of the AIA and could act as a drag on innovation by (a) delaying or preventing systems beneficial to EU citizens from being launched in the EU and (b) limiting the market to those that can meet the cost of compliance.

Notwithstanding the burden we have identified above, we are also concerned that within the compliance mechanisms, the conformity assessment process depends on an organisation's compliance and governance maturity that itself risks the compliance process lacking rigour.

The disclosure requirements under the AIA are sweeping, wide-ranging, and lack specific custody and protection requirements which are commensurate with the investment that parties have made in the creation of their AI systems. A failure on the part of the EU to ensure adequate safeguards in relation to disclosed information relating to AI systems may ultimately act as a disincentive to EU-wide development and undermine effective trade-secret protection in such technology.

Compliance Emphasis

In noting that the AIA has an unduly onerous complex compliance burden, astute readers will notice calls for increased compliance in certain areas of our detailed commentary—in particular, calls to extend regulation to systems which are not "high risk" as defined by the AIA. We do not view these requests as inconsistent with our commentary on compliance burdens. We are merely acknowledging here that the focus on compliance should be on different areas which are complementary to our concerns around outcomes-based regulation, citizen redress, and human accountability.

Sustainability and the Environment

Finally, we are of the view that the AIA does not yet meet the increasingly high standards that are required in terms of sustainability and environment. It is true that the draft compromise text introduces AI systems that have the ability to control pollution and the environment as "High Risk," but that falls some way short of mandating that in any AI-use case, an environmental decision has to be made. AI, as we all know, is a heavy consumer of power and hence carbon intensive. Any decision to employ an AI system must therefore be done as sustainably and responsibly as possible on the basis that its particular use case outweighs the carbon cost. The EU has an ideal opportunity in the AIA to become a class leader in this respect.

Concluding Remarks

It is our hope that you find our recommendations and cautionary points constructive as to the design and potential operability of the currently proposed law, and conducive to its further improvement. Should any EU official or elected representative wish to discuss any matter which arises out of this Green Paper further, please contact:

By email: greenpaper@itechlaw.com

Address: International Technology Law Association
7918 Jones Branch Drive, Suite 300
McLean, VA 22102, United States

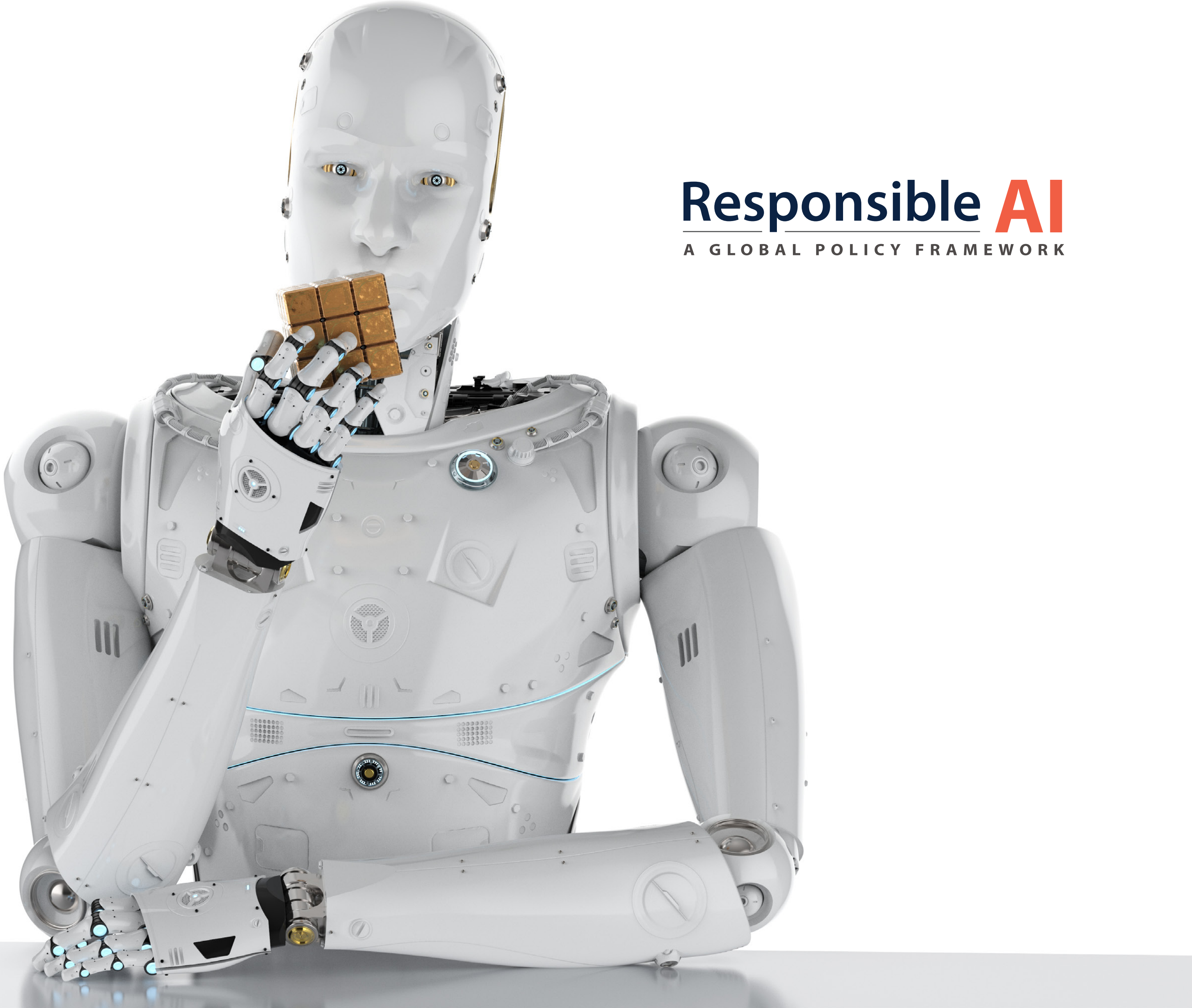
Editors: **John Buyers** | Osborne Clarke LLP
John.buyers@osborneclarke.com

Patricia Shaw | Beyond Reach Consulting Limited
trish@beyondreach.uk.com

Susan Barty | CMS Cameron McKenna Nabarro Olswang LLP
Susan.Barty@cms-cmno.com

PART A

Summary of Recommendations, Cautions, and Oversights



Responsible AI

A GLOBAL POLICY FRAMEWORK

THE EU AIA: A GREEN PAPER POLICY ANALYSIS

Summary of Recommendations

1 Ethical Purpose and Societal Benefit

- **Recommendation 1.1:** The EU should make overt the interdependencies required in relation to other legislative measures as these have been underpublicized to date. In addition, we would recommend that if there are further legislative measures to bring the AIA fully into effect, then an overall legislative plan is also communicated effectively by the EU.

- **Recommendation 1.2:** The AIA appears to overlook the potential for the use of AI systems in the workplace to have a detrimental impact on the psychological health of the human workers themselves (irrespective of the particular use case concerned), both by denying them dignity or fulfilment of purpose or by increasing the potential for workers to be involved in repetitive and menial tasks.

- **Recommendation 1.3:** We are disappointed that the environmental “cost-benefit” analysis is only reflected to a limited degree under Article 69 of the AIA as a potential candidate for a code of conduct to be adopted on a voluntary basis. Given the power consumption of such technologies, environmental impact should be a key consideration in evaluating the benefit of their introduction and should be reflected in the technical documentation requirements of Annex IV.

- **Recommendation 1.4:** Article 2.3 of the AIA and its associated Recitals should be aligned with one another, and they should be amended to emphasize that certain military AI systems are excluded solely on grounds of Member State competency, and where such systems are commissioned and/or deployed by the Member States.

- **Recommendation 1.5:** We encourage the EU institutions and bodies, including the European Defence Fund, to pass and act upon rules for the participation, codes of conduct, or other “soft law” that foster the principles of responsible AI in the military context. In doing so, the EU should apply higher and stricter standards than those adopted as part of the NATO Artificial Intelligence Strategy which contains significant ambiguities, for instance as regards bias minimization as opposed to avoidance of bias.

- **Recommendation 1.6:** In the context of addressing misinformation and online lies, we recommend that—at the very least—AI systems should not be used to actively propagate such content (through algorithmic reinforcement).

2 Accountability

Recommendation 2.1: We consider there are areas where the AIA is not attuned to the commercial realities of the AI ecosystem, the plight of the start-up or the SME, and the overall cost burdens of compliance. The current draft AIA does not encourage a culture of responsibility taking within organisations nor does it appropriately hold entities to account. The AIA should be amended to reflect the importance of embedding a responsible AI culture at an organizational level.

Recommendation 2.2: The AIA falls short of identifying those persons/roles who can be held responsible for the acts and omissions of an AI system There should be no room for ambiguity or confusion that might lead private actors to believe that they could blame an AI system for any non-compliant conduct, particularly for any violations of fundamental rights. The AIA should have a clear and unambiguous statement denying any independent legal status for AI systems. We also recommend that the AIA affirm our principle that there should always be a human behind the AI regardless of the criticality of AI.

Recommendation 2.3: We consider that if a multi-tiered regulatory compliance structure is to persist that there should be greater collaboration, coordination, cooperation, and communication, not just between the EU institutions and regulatory bodies across the EU but with governments and regulatory bodies outside the EU. Without such collaboration and coordination, the AIA may have the effect of potentially exporting EU values and the so-called “Brussels Effect” in an undesirable manner, raising some of the same extraterritoriality concerns for the AIA as has been seen as the EU GDPR has evolved.

Recommendation 2.4: The AIA should provide a clearer picture in relation to various ecosystem roles and responsibilities and clearly demonstrate how these may change when the AIA adds or shifts provider-type duties to another actor or creates interdependent duties.

Recommendation 2.5: We consider that further guidance, and significant scrutiny of AI governance and oversight within providers and other actors, should form part of the AIA in order to ensure that organisations maintain effective accountability systems internally as well as comply with the requirements to permit external accountability.

Recommendation 2.6: The information that must be detailed in the technical documentation should be improved and tailored to a supervising regulator and/or notified body, given the commercial sensitivities of the disclosure. Original AI code and training datasets should be held in escrow to allow assessment of systems in their original, pre-deployment form and also to ward off phoenix syndrome–like behaviour from AI businesses.

Recommendation 2.7: The categorisation of types of AI systems that require a conformity assessment involving an independent notified body should be expanded. The list should not just include biometric identification and categorisation systems (draft Annex III para 1) but should also include draft Annex III paragraphs 5 through to 8 as well. Independent use of notified bodies creates further accountability. If independent notified bodies are not suitable for reviews to be performed by a Market

Surveillance Authority, then additional safeguards should be implemented to maximize the independence and reliability of the conformity assessment.

Recommendation 2.8: Measures and processes should be developed to reduce the burdens of regulatory compliance, particularly for smaller businesses, by identifying circumstances where reliance on existing systems, datasets, or approaches reduce the risks of new or modified AI systems. Regulators should be cognizant of the anti-competitive effects that the draft AIA and any revisions to the AIA might engender. The AIA should also consider if there are areas that do not qualify for inclusion as a “High-risk AI,” but that should not be left entirely to voluntary compliance by providers.

Recommendation 2.9: The draft AIA needs to expressly provide for the decommissioning and sunsetting of AI systems, building in de minimis legal requirements. Practical and ethical steps are also needed to mitigate risks and the impact of business and social dependency in the event that an AI system is sunsetted or decommissioned.

Recommendation 2.10: The draft AIA needs to expand on and develop the principle of proportionality in the application of administrative fines and provide a sufficient framework or process to ensure reasonably harmonised schemes of penalties. Penalty schemes should also take account of and/or provide guidance concerning situations of dependency and the downstream effects of imposing sizable fines on providers, importers, or distributors.

Recommendation 2.11: The draft AIA needs to provide (or make reference to via external legislation) (1) a mechanism for queries from potentially harmed individuals to be investigated across regulated and unregulated arenas, because deployment of an application-agnostic technology may arise in many varied fora which overlap sectors, and (2) a mechanism which makes justice, equity, restitution, and (if necessary and appropriate) damages for individual and/or group harm accessible to all.

3 Transparency and Explainability

Recommendation 3.1: Usage of the term “transparency” within the AIA varies depending on context and the relevant article. Articles 12 and 13, for example, refer to “transparency” but in fact appear to be more concerned with the intelligibility of machine learning decisions (or explainability as it is defined by Principle 3). Article 52, in contrast, also refers to “transparency” but in a manner which is more consistent with the Principle 3 definition. We recommend that this definitional uncertainty is corrected in future drafts of the AIA.

Recommendation 3.2: Heightened transparency requirements should be generally applicable to all regulated high-risk AI systems and at least equivalent to those in Article 52.

Recommendation 3.3: We would like to see some element of transparency by design as reflected in Principle 3 of the Responsible AI Framework extended to the development of all AI systems, regardless of classification. Standard, minimal transparency obligations could be designed for mid- and low-risk AI systems.

- **Recommendation 3.4:** A hierarchy should be established within the AIA to minimise the inconsistencies between transparency and information keeping concepts.

- **Recommendation 3.5:** Further thought should be given to public disclosure issues by the EU. There are evident inconsistencies in disclosure levels in the current draft of the AIA which could harm public accessibility to justice for individuals harmed by AI decision making. For example, the documentation that arguably must contain a bias assessment does not need to be provided to users, the public, or those potentially affected by discriminatory algorithms. It is available only to regulators upon request. In contrast, the legislation clearly mandates assessments and disclosure of system accuracy.

- **Recommendation 3.6:** The AIA does not outline any mechanisms by which those harmed by AI systems may seek recourse and redress from the user of AI systems (with the exception of Article 52) similar to those which can be found in equivalent legislative measures such as the GDPR. We call on the EU for urgent clarification as to how these essential requirements will be met.

4 Fairness and Non-Discrimination

- **Recommendation 4.1:** Whilst we recognize that an outcomes-based approach may necessarily involve a significant "rethink" of the manner in which AI systems are regulated by the AIA, a failure to adequately include a subjective mechanism of assessment in terms of bias and discrimination is a significant omission. We call upon the EU to reconsider the inclusion of a subjective user-based harm-impact mechanism, such as that used within Article 22 of the GDPR.

- **Recommendation 4.2:** We would like to see a clearer inclusion of bias and fairness in the documentary requirements, particularly some specificity on the part of the EU in terms of key metrics which would assist organisations in determining the characteristics of biased outcomes and potential unfairness for the AI system in question. These need to be assessed together with the identification of biases or vulnerabilities in the AI's intended or unintended use and diversity of its application domain.

- **Recommendation 4.3:** In so far as the documentary requirements of the AIA are concerned, the Article 11 (Annex IV) requirements that arguably must contain a bias assessment does not need to be provided to users, the public, or those potentially affected by discriminatory algorithms. They are available only to regulators upon request. In contrast, the legislation clearly mandates assessments and disclosure of system accuracy. We would strongly recommend that some form of individual access right is introduced for individuals and groups who may have been discriminated against by AI decisions.

- **Recommendation 4.4:** As with the technical documentation that needs to be produced in accordance with Article 11 and Annex IV, we would like to see explicit further mention in the Conformity Assessment Procedure set out in Article 43 and Annex VII to bias-related testing. The same considerations as to user access mentioned above under Documentary Requirements applies to wider Conformity Assessment documentation, and we would see the Conformity Assessment

Documentation as forming a fundamental part of any evidence that could be accessed and used by individuals to seek remedy in cases of discriminatory harm.

Recommendation 4.5: We would like to see more details from the EU in terms of the way in which the fairness and non-bias goals in Article 10 and Article 15 are likely to be achieved in practical terms. A failure to provide any such benchmarks makes the obligations essentially self-policing by the applicable AI provider and makes it harder to determine the extent to which standards have been met (or otherwise).

Recommendation 4.6: Further mechanism and accountability requirements should be added to the AIA to remove any potential ambiguity and uncertainty, as the harms and impacts vary considerably depending on the purpose, scope, and context of the use of biometrics systems. As illustrated above, these emotion recognition systems and biometric categorisation systems could be deployed as part of the seven other high-risk AI systems listed pursuant to Article 6(2) and Annex III.

5 Safety and Reliability

Recommendation 5.1: We would hope for (industry) codes of conduct to be developed (see Article 69) that voluntarily introduce a requirement to establish a risk management system for such AI systems, but optimally we would call for the scope of these obligations to be extended to other AI systems and also to users of such AI systems.

Recommendation 5.2: The AIA strongly focuses on products and related AI systems, and links requirements to the EU product liability regime. However, AI systems may differ and mainly consist of services requiring different set-ups from a security perspective. These varying demands need to be reflected in the AI regulation, which is not yet apparent from the current draft. Therefore, we would strongly recommend that differentiation of security requirements depending upon the extent to which they are based on hardware, software, services, or product types should be considered and factored into the next draft.

Recommendation 5.3: In relation to security and cybersecurity aspects, the EU Commission should consider including additional remarks specifying the applicable standards to be complied with. Such specification would increase the clarity on respective requirements. In order to ensure a lasting adequacy, either a reference to non-static standards or a regular review and update thereof seems preferable. Specific definitions, requirements, and consequences in case of a breach related specifically to high-risk AI systems seem to be missing. This concerns mainly Articles 6, 8, 9, 10, 15, 16, and 27 of the AIA.

Recommendation 5.4: The European Commission should adopt an ethically underpinned approach within the AIA which will align it with the European Parliament Resolution of 20 October 2020 on a Framework of ethical aspects of artificial intelligence, robotics, and related technologies.

→ **Recommendation 5.5:** We recommend that the conformity re-assessment procedure in Article 43 is reassessed within the AIA. Currently, this procedure is only necessary when the AI system itself is modified (as stated in Article 43.4). We feel that the process would benefit from being widened to also apply when relevant circumstantial facts (such as the underpinning human-defined objectives of the system) change. If so, this would provide an opportunity to raise subsequent questions as to whether such relevant changed circumstantial facts should also include an assessment of the extent to which the AI system conforms to (and is underpinned by) recognized ethical standards.

→ **Recommendation 5.6:** There is a need for users of high-risk AI systems to define the ethical principles underpinning the high-risk AI system they are using. We encourage the EU to re-assess this aspect and also to introduce an obligation for governments and organizations using high-risk AI systems to document the ethical principles underpinning the high-risk AI systems they use.

→ **Recommendation 5.7:** We call for the scope of the documentation obligation for high-risk AI systems under Article 11 to be extended to any AI system (high risk or not, including “general purpose” systems and those used for scientific research and development) that (i) interacts with humans, (ii) is used to detect emotions or determine association with (social) categories based on biometric data, (iii) generates or manipulates content, or (iv) is designed for autonomous decision making.

→ **Recommendation 5.8:** For any documentation obligations concerning underpinning ethical standards or principles, the AIA does not make a distinction between AI systems that are designed to autonomously make decisions affecting humans and AI systems that do not. Consequently, we further recommend that this obligation be extended to AI systems (high risk or not) that are designed for autonomous decision making, especially where such decision making may negatively impact fundamental rights and/or ethical standards or principles.

→ **Recommendation 5.9:** We would encourage the European Commission to broaden the scope of the testing and logging obligations under Articles 9 and 12 to include reasonably foreseeable misuse.

→ **Recommendation 5.10:** Further specific limitations must be added to safeguard the processing of special categories of personal data, which should only be permitted to the extent strictly necessary and on a proportionate basis for the intended purpose and outcome of a high-risk AI system.

→ **Recommendation 5.11:** Self-regulation should be extended further to cover codes of conduct related to environmental sustainability, accessibility for persons with a disability, stakeholder participation in design and development, and diversity (all topics that are not covered in the current draft of the AIA). We are of the view that it does not matter whether codes of conduct are drawn up by individual providers, by representative organizations, or by both—we would like to see this flexibility promoted across the industry.

→ **Recommendation 5.12:** In terms of other specific issues within the regulatory regime proposed by the AIA, we observe that high-risk AI systems are required to produce log files which are subject to specific rules. The fact that these log files can simply be requested to be handed over to regulatory authorities in the absence of specific justification does not seem to present to us sufficient legal grounds and does not appear to be transparent. Whilst the connection between a system failing in compliance terms and disclosure of a relevant log file is implicit, we would comment that there needs to be an appropriate legal mechanism to be put in place for authorities to request log files for a limited list of legal reasons. Furthermore, we are of the view that the duty to store log files needs to be limited in time and related to certain incidents.

Recommendation 5.13: In terms of the power by regulatory authorities to disconnect or shut down high-risk AI systems, we are of the view that this needs to be more specific and more precise to enable it to sit comfortably within the balanced systems of rights and protections envisaged by European law. As it stands, the right very simply expressed could create a wide range of potential further issues if exercised (including matters such as liability, data protection issues, violations of other AIA provisions, and potential removal from EU citizens of valuable technological infrastructure supporting critical services). In our view, consideration needs to be given (particularly in circumstances where the affected high-risk systems may be depended upon by large numbers of people due to their criticality) that alternative methods of enforcement are considered, such as enforcing alternative non-AI technological methods to achieve the same goals.

6 Open Data and Fair Competition

Recommendation 6.1: We would recommend that the market monitoring and surveillance provisions of Article 63 are broadened to include a consultation procedure. This would be preferable since it would make competition law assessments more integrated with AI regulatory sector specific intervention, similar to the regulation in the electronic communication services sector.

Recommendation 6.2: We would suggest that the sandbox environment be expanded to include open data access and thereby serve as an enabler to ensure fair competition and lift smaller players up by way of also offering access to data (e.g., openly available training sets).

Recommendation 6.3: We consider that it would be preferable to see some alignment in the AIA with the principles in the Open Data Directive and also the forthcoming DGA and EU Data Act, including data altruism. Consistent with this, there should be recognition and alignment with new business models proposed by the DGA, such as data intermediation services. For example, Principle 6 of the Responsible AI Framework states that organisations that develop, deploy, or use data-driven systems and any national laws that regulate such use shall promote open source and decentralised frameworks. This means that an AI regulation should assess how the use of AI tech solutions and their outputs can be used in other situations or by other organisations, and private organisations should be encouraged or fostered through open access and portability of datasets. Public sector bodies in particular should be required or at least encouraged to ensure that data held by them and used within their AI systems are portable, accessible, and open if reasonably practicable.

Recommendation 6.4: Notwithstanding the forthcoming DGA and EU Data Acts, we also are of the view that the topic of open data and fair competition should be expanded in the AIA (to again align it with a consistent EU approach). This principle is more than a policy consideration and needs to be embedded in the regulatory approach as well.

Recommendation 6.5: The principle of fairness is embedded in the AIA. That principle of fairness focuses on equality and non-biased discrimination, and does not seem to cover fair competition. There is little reference in the proposed Regulation to regulatory tools that can facilitate fair competition (i.e., access to data) and no mention of the negative effects AI systems may have on competition, such as algorithmic collusion. We would like to see these aspects covered in significantly more detail in updated drafts of the AIA (or anti-trust legislation that may accompany it).

7 Privacy

Recommendation 7.1: The AIA needs to better address the practical realities of how an AI system may be developed, so it is clearer as to which stakeholder the relevant obligations should apply to and how. Alternatively, it could take a more outcomes-based approach, so that any party that engages with an AI system has an obligation to ensure appropriate operational safeguards are put in place to mitigate against adverse outcomes.

Recommendation 7.2: We would recommend that a supportive measure within the AIA might be to prescribe certain "compliant" methods of anonymisation (for example, that it should be done by a third-party processor). Subsequent provisions could clearly prohibit de-anonymisation and provide for penalties in the event of an infringement.

Recommendation 7.3: We are concerned that there is an open question as to which processing activities can be subsumed under the term "scientific and statistical purposes," especially as AI could include statistical approaches (see Annex I of the AIA). Interpreted widely, this could conceivably permit any AI processing for further purposes which we argue would be an unintended consequence. We would therefore urgently recommend a clarification on the status of AI systems vis-à-vis Article 89 of the GDPR.

Recommendation 7.4: We would submit that the goal of any machine learning system is to focus on algorithms that are designed to find patterns in data and use these patterns to make predictions. Hence, in many AI systems (in particular in the field of data mining that deals with the prediction of future developments), profiling is an indispensable component. This is why we consider that it would enhance the AIA in this context to define a framework which allows profiling as part of machine learning and provides for a context-driven explanation of the consequences of this profiling, rather than just a blanket proscription.

Recommendation 7.5: For clarity, it would also be beneficial to provide further provisions to regulate private use of remote biometric identification in publicly accessible spaces in a way that is consistent with Article 9 of the GDPR. The AIA should explicitly clarify that existing EU data protection legislation applies to any processing of personal data falling under its scope regarding biometric data, including the GDPR. As the AIA would be a global standards-setting piece of legislation—much like the GDPR—it would assist if any potential ambiguities in this regard were removed.

8 AI and Intellectual Property

Recommendation 8.1: The introduction of specific requirements regarding processes for the secure management of information by the national competent authorities and notified bodies involved in the application of the AIA is considered highly appropriate, also by reference to existing international standards and ISMS frameworks (e.g., ISO 27001, NIST, etc.) and audit best practices, also including a specific requirement regarding time limitations in the storage of or access to data and documents, limited to what is strictly necessary for the performance of their duties under the AIA.

Recommendation 8.2: Standard access and disclosure agreements should be elaborated by the Commission—as has happened, for example, with Standard Contractual Clauses under the GDPR—in order to facilitate standardisation of the access procedures and ensure the effectiveness of confidentiality measures among Member States, and to provide an efficient allocation of responsibilities in case of intentional or non-intentional breach of confidentiality of data and information accessed by the national competent authorities and notified bodies.

Recommendation 8.3: We would like to see the respective bodies responsible for assessing AI innovation and systems consult with each other with respect to the principles they are to implement in performing their duties and to ensure that there is a consistency of principle and approach between them.

Recommendation 8.4: A more effective approach to determining the operation of an algorithm when implementing it in an AI system would be to test the operation of the AI system by utilising a reference input to determine a satisfactory output. In this way, the issue of what goes on in the "black box" is irrelevant providing the output fulfils the necessary criteria when tested against a particular input. This may be considered an outcome-based approach rather than a design-based approach where the design of the algorithm and its configuration is to be transparent and explainable. Such an outcomes-based approach may obviate the need for access to data and documents to assess an AI system provided it behaves within what are deemed to be acceptable parameters.

THE EU AIA: A GREEN PAPER POLICY ANALYSIS

Summary of Cautions and Oversights

1 Ethical Purpose and Societal Benefit

⚠ **Caution 1.1:** There is substantial definitional uncertainty around what constitutes a "User" under the AIA, which is in need of further clarification, particularly in respect of private individual remedies.

⚠ **Caution 1.2:** The public policy reason for prohibiting surveillance activity in Articles 5.1 (d), 5.2, and 5.3 should be to reinforce democratic human agency and provide a check and balance on state power as well as on overly intrusive private sector "surveillance capital." We consider that the addition of national security purposes as an exclusion—in addition to the failure to encompass private sector surveillance (other than in the limited context of third party agencies acting on behalf of law enforcement authorities) together with all of the potential mechanisms of state and government—to be significant and fundamental weaknesses in the proposed legislation.

⚠ **Caution 1.3:** The AIA falls some way short in mandating responsible, environmentally friendly use of AI systems. Whilst the legislative measure is not designed to be one that dictates environmental protection, it should at least recognise the environmental cost involved in the use of all such technologies. Environmental issues are currently dealt with by the AIA on a narrow use-case basis only, without an explicit acknowledgement to the environmental impact of the use of all AI systems. We would recommend that the EU have regard to the latest draft of the UNESCO guidance on Artificial Intelligence and Ethics, which places environmental use of AI systems at its heart.

⚠ **Caution 1.4:** Other than Article 52.3 on the use of "deep fakes" and the relatively narrow categories of Prohibited AI set out in Articles 5.1 (a) and 5.1 (b), the AIA really does not tackle in any meaningful sense information which has been "weaponized," distorted, or manipulated to serve particular agendas that could ultimately harm individuals, groups of individuals, or democratic institutions.

2 Accountability

⚠ **Caution 2.1:** The draft AIA does not effectively account for AI supply chain and/or AI components (whether procured through "AI as a Service" offerings or not) despite further revisions provided for in the Presidency Compromise text. The lines between manufacturer, distributor, importer, provider, and user can become blurred. When AI is put onto the market without either a specifically intended purpose or where the AI can perform generally applicable functions (so-called "General Purpose" AI), demarcation of responsibility could be difficult to unpick and become confused (Recital 70a and Article 52a). Even where the user does not modify the AI (by only taking AI or a pre-trained model "off the shelf"), the regulatory burden could potentially fall on that user as if they were a provider. This may have a disproportionate effect.

⚠ **Caution 2.2:** Who bears primary responsibility under the AIA needs further clarity, particularly in respect of "AI as a Service" or MLopS, with regard to how responsibility and duties (and ultimately liability) do and do not change over time and as circumstances change. If this remains unclear in the AIA, this will create uncertainty for businesses, and may have an undesirable and/or innovation-stifling effect.

⚠ **Caution 2.3:** The new Presidency Compromise text introduces two further exemptions from the scope of the AIA. These are AI developed and launched which is "general purpose" in nature (see the new proposed Article 52A) and AI which is either used for Scientific Research and Development purposes or any research on AI (to the extent such systems are not placed onto the market) (see Articles 2(6) and 2(7)). These new exemptions seem to us to create a significant gap in the ability of the AIA to ensure accountability across the EU AI ecosystem. The concerns stem from the fact that the exemption seems to (a) side-step the potential consequences of the use of such unregulated systems and (b) ignore the role that the actual developers of such systems (as opposed to legally identified "producers" under the AIA) in using best practice to ensure basic algorithmic hygiene, utilise clear system architectures, and manage training probity. This would appear to us to create a fundamental accountability problem in that these parties are presumed to escape liability or consequence for such systems despite the fundamental role they have had in their creation.

⚠ **Caution 2.4:** The right of a notified body to carry out periodic audits should be evaluated to enhance the role of regulators beyond the confines of mere management system quality assurance, and to identify and scrutinise key aspects of AI systems for potential violation of fundamental human rights.

⚠ **Caution 2.5:** Unfortunately, the rigour of expected compliance with the conformity assessment requirements is questionable. What the AIA says should be done and what is actually done in practice may differ. Without any scrutiny of an organisation's conduct with respect to AI through a third-party "notified body" assessment or through consideration of user feedback, any assessment conducted internally by an organisation may result in a flawed assessment that is based on a closed feedback loop or is biased. Depending on organisations to conduct their own conformity assessments risks the compliance process itself lacking rigour.

⚠ **Caution 2.6:** The bifurcation of the proposed AIA for High Risk and other AI systems overlooks a range of circumstances in which intermediate levels of oversight and regulation may be needed to ensure proper accountability—particularly when the costs of regulatory compliance can create barriers to entry for start-ups, scale-ups, and SMEs—and invite the use of anti-competitive practices by increasingly large market players.

⚠ **Caution 2.7:** The draft AIA lacks a suitable "one stop shop"-style redress mechanism that can provide an effective and available means for all those injured by an AI system to seek compensation for their injuries.

3 Transparency and Explainability

⚠ **Caution 3.1:** There are few substantive AI transparency obligations within the AIA within the definition of Principle 3. The AIA is light on information that must be disclosed to the people who are affected by AI systems. We would question the alternative focus in the draft measure on AI explainability which is not a complete substitute for transparency obligations and may in fact be technically very difficult to achieve in the context of some AI systems based in deep learning.

⚠ **Caution 3.2:** Article 52A on "General Purpose" system exemption seems to us to move the AIA further away from an approach which should extend at least some transparency obligations to all AI systems and consequently further away from the goal of enabling users of AI systems to understand the manner in which decisions have been made on a consistent basis. We view it as a retrograde step.

4 Fairness and Non-Discrimination

⚠ **Caution 4.1:** We consider taking a risk-based approach based on use cases to be a relatively unsustainable way to proceed as the EU is now faced with having to anticipate, on a case-by-case basis, which types of AI application could have an increased potential to cause bias or unfair outcomes. A far better way of achieving this is by way of an outcomes-based approach, which is necessarily use-case agnostic.

⚠ **Caution 4.2:** The new Presidency Compromise text allows for "general purpose" AI to be exempted from the provisions of the AIA, as is AI used for Scientific Research and Development. Again, we see these exclusions as highly problematic and indicative of the use-case methodology that the EU appears to be adopting. In our view, it overlooks the consequences that the use of such systems could have on their users in terms of bias and discrimination.

⚠ **Caution 4.3:** Principle 4 requires that parties who are harmed by the decisions of AI systems should have effective ways to seek remedies in discriminatory or unfair contexts. We find the AIA significantly lacking in this regard, and have raised this as a substantial concern in the context of different Responsible AI Principles elsewhere in this Green Paper.

5 Safety and Reliability

⚠ **Caution 5.1:** We are concerned that the risk management requirements of Article 9 of the AIA only apply on a compulsory basis to systems that are classified as "High Risk" and thus not to low-/minimal-risk AI systems, not even if the AI system (i) interacts with humans, (ii) is used to detect emotions or determine association with (social) categories based on biometric data, (iii) generates or manipulates content, or (iv) is designed for autonomous decision making (and such decision making may negatively impact fundamental rights or ethical standards or principles).

⚠ **Caution 5.2:** Security, and specifically cybersecurity, are very important topics in the AIA. However, although these topics are mentioned throughout the AIA, neither security nor cybersecurity are defined in terms of the actual steps which should be taken by Providers. We are concerned by this as embedding cybersecurity (and hence security more widely) into an AI system effectively needs to be done at inception or, to borrow the phraseology of the GDPR, by design—it is not just an ongoing, post-implementation process.

⚠ **Caution 5.3:** Although the AIA (according to Recital 5) supports the EU objective of being a global leader in the development of secure, trustworthy, and ethical artificial intelligence (as stated by the European Council), and ensures the protection of ethical principles (as specifically requested by the European Parliament), the Conformity Assessment procedure lacks a clear obligation for providers and users of (high-risk) AI systems in terms of adhering to ethical principles when designing, developing, putting on the market, putting into service, or using such systems.

⚠ **Caution 5.4:** As for data used in high-risk AI systems, the AIA currently requires all training, validation, and testing data to be relevant, representative, free of errors, and complete (Article 10.3). We would question whether it is possible to ensure a "letter-perfect" standard in this regard. Instead we recommend that AI systems should be free of errors "so far as is practicable under the given circumstances and taking into account the intended purpose of the AI system."

6 Open Data and Fair Competition

Not applicable.

7 Privacy

⚠ **Caution 7.1:** The AIA does not appear to provide for situations where there are multiple stakeholders involved in providing a high-risk AI system, for example, an IT company provides the engine on which it is built, a data analytics company overlays its expertise, and another company feeds in the data required to teach an AI system. It is not clear which party will be the "provider" under the AIA in this type of arrangement, so which party is required to comply with the requirements under Article 16?

⚠ **Caution 7.2:** Whilst the AIA sets rules for high-risk AI systems that use data for training (see Article 10), it fails to create a set of rules for the use of anonymous data, and even more so to create a clear distinction between pseudonymous and anonymous data. Our concern is that a significant proportion of AI models (not just high-risk AI systems) are trained with personal data often derived from other processing activities. Even if in this context the data are pseudonymised, the framework of Articles 5 and 6 of the GDPR will apply until such data are fully anonymous. This creates several legal questions because further processing of personal data in order to train, test, and evaluate AI systems will not be compliant in many situations. Therefore, anonymisation before change of purpose, or pseudonymisation thereafter with the consequence of exemptions from the GDPR, is a vital interest for providers of AI systems.

⚠ **Caution 7.3:** The AIA does not address the risk of certain AI systems (particularly General Adversarial Networks) to be used as tools to unpick privacy preserving techniques or to game a targeted AI system to create further vulnerability.

⚠ **Caution 7.4:** There must be additional safeguards in place beyond the mere application of data protection principles to prohibit AI system training when the balance between the fundamental rights of the data subject and the legitimate interests of the AI provider is not met, especially with high-risk AI systems. We are of the view that specific quality criterion have been omitted from the AIA in this regard—that the processed data must be, before being utilised for AI system training, in the overriding interest of the data subject and/or the AI user. In no circumstances should training be undertaken when the overriding interest is not with the AI user.

⚠ **Caution 7.5:** Profiling and automated decision making are activities typically reliant on AI and involve personal data. The extent to which and with what consequences the GDPR applies to AI systems/machine learning in this context is not specifically addressed by the AIA.

8 AI and Intellectual Property

⚠ **Caution 8.1:** Intuitively, allowing access to the information specified in Article 70 AIA to a public third party in the absence of specific protection requirements of such information has the potential to create a potential (and unnecessary) point of failure that, if exploited for example by a malicious third-party agent, could undermine the effective protection as trade secrets of the data and information disclosed.

⚠ **Caution 8.2:** In light of its disclosure requirements, there is a risk that the AIA may act to disincentivise and/or dissuade development and/or exploitation of AI systems in the territory of the European Union (EU), thereby depriving EU citizens and/or undertakings based in the EU of any benefit afforded by such AI systems, possibly to the detriment of EU society.

⚠ **Caution 8.3:** The explainability requirements of the AIA appear to be built upon the assumption that a technology creator is able to understand how the technology functions and to explain how an algorithm makes decisions based on the dataset parsed by it. These assumptions (particularly in cases where Deep Neural Networks are involved) may in certain circumstances be inaccurate, since such algorithms are not just protected as trade secrets under a legal perspective, but also constitute in most cases technological “black boxes” where the algorithm is secret by definition. The notion of black box AI refers to scenarios in which we can see only input data and output data for algorithm-based systems, without having insight into exactly what happens in between. In this respect, the complexity of the AI system/algorithm leads to a high level of difficulty in providing effective explanations.

PART B

Discussion Based on Responsible AI Principles

Responsible AI

A GLOBAL POLICY FRAMEWORK

Principle 1
Ethical Purpose and Societal Benefit

Organisations that develop, make available or use AI systems and any national laws or industry standards that govern such use should require the purposes of such implementation to be identified and ensure that such purposes are consistent with the overall ethical purposes of beneficence and non-maleficence, as well as the other principles of the Policy Framework for Responsible AI.

Principle 1 of the Responsible AI Framework is intended to provide an overall "Do no evil" wrapper around the development and use of AI Systems. As such, it encompasses a generally beneficent approach which emphasizes the importance of human agency and autonomy (aka "humans in the loop"), minimizes the inevitable impact that these technologies will have in the context of employment automation, requires consideration of the wider impact to the environment and requires military applications, particularly AI with the potential for lethal consequences, to be similarly regulated. For the full text of this principle, please see page 92.

Overall Approach and Legislative Basis of the AIA

The AIA adopts a general product-safety approach when legislating on the use, deployment, and adoption of artificially intelligent systems. Whilst this elegantly enables the regulatory measure to sit within an existing corpus of EU product safety legislation, which already has a highly evolved market monitoring and compliance regime, we note that this makes prioritisation of human rights considerations inherently harder to achieve as they are not the primary focus of the draft legislation. We further note that the legislative approach adopted is overtly linked to other EU legislative measures outside of product safety, such as the General Data Protection Regulation. These interdependencies in our view make the AIA inherently harder to evaluate (and potentially to implement) as it does not necessarily stand in isolation. Indeed we make other observations elsewhere in this paper in relation to omitted content which may be dependent on further legislative measures by the EU.

→ **Recommendation 1.1: The EU should make overt the interdependencies required in relation to other legislative measures as these have been underpublicized to date. In addition, we would recommend that if there are further legislative measures to bring the AIA fully into effect, then an overall legislative plan is also communicated effectively by the EU.**

Compliance Burden

We would tend to share the opinions of many who have commented on the draft AIA in relation to the potential for the compliance regime contemplated by it to be overly onerous. Whilst not specifically and directly linked to Principle 1 of the Responsible AI Framework (our commentaries under Principles 2, 3, and 5 provide for a more detailed analysis), we consider it sufficiently important to be raised here. Our concern is that this burden could impact the overall effectiveness of the measure and mean that either systems which could be beneficial to EU citizens are not actually launched in the EU and/or limit the market for the development of such systems to entities that are sufficiently resourced to meet the costs of such compliance. In either case, it is likely that this may affect the EU market for such technologies in a manner which unintentionally adversely affects EU Citizens and puts them in a disadvantageous position vis-à-vis citizens of other nations, particularly where such AI use is safe and advantageous to individuals.

Private Individual Remedies

⚠ **Caution 1.1: There is substantial definitional uncertainty around what constitutes a "User" under the AIA, which is in need of further clarification, particularly in respect of private individual remedies.**

We repeat this as a theme in some of our other commentaries in this Green Paper, principally in the sections dealing with Accountability (Principle 2) and Transparency and Explainability (Principle 3).

This definitional uncertainty manifests itself in two substantial ways in the AIA and has important implications for the rights of private individuals who may be harmed by the actions of "Providers." Firstly, although the term "User" as defined in the AIA encompasses either a natural or legal person, it is not clear whether that user is intended to be an entity which deploys an AI system in an authorised manner and not in general terms natural persons who are subjected to the outputs of the AI system (see in particular Article 29 in this regard). Secondly, if users are not intended to be general natural persons subjected to the outputs of AI systems, where are their potential remedies should they in fact be subjected to harm as the result of the use of such systems? There seems to be a tacit admission by the EU of this liability "gap" and the need to provide an appropriate liability framework for private citizens which is not adequately dealt with by the current draft of the AIA. In particular, the Commission acknowledges in an Inception Impact Assessment dated 30th June 2021[1] that the liability rules set out in national law and the EU Product Liability Directive[2] may need to be adapted to accommodate the unique nature of AI-based technologies.

Human Rights, Agency, and Autonomy

Article 5 of the AIA (Prohibited AI) is one of the specific areas in the draft measure where human rights considerations are particularly manifested, and we commend the attempts of the EU to proscribe certain use cases on the basis that they may be deceptive, cause physical or psychological harm or be overly

[1] Ref. Ares(2021)4266516 - 30/06/2021—Adapting liability rules to the digital age and circular economy.

[2] Directive 85/374/EEC.

inimical to accepted standards of the rule of law and human autonomy in Western democratic societies. We do not propose to list these use cases in this commentary but have some general observations to make in relation to them.

The Prohibited Use Cases set out in Articles 5.1 (a) and 5.1 (b) (systems deploying subliminal techniques and systems exploiting vulnerabilities) are all qualified and drafted in an exceptionally narrow manner which require a number of evidential hurdles to be met. They focus on the purpose of the proposed AI use rather than the outcome of that use on end-users. For example, very little attention is devoted in the AIA to risks associated with destructive psychological and behavioural manipulation, addiction, dependency, and attention deficit. In this regard, it is surprising that the EU has not adopted an approach which is equivalent to that used by Article 22 of the GDPR on Automated Decision Making (ADM), which refers to decisions having "legal effects" or "similarly significantly affecting" those individuals who are subject to ADM. As a legislative approach, Article 22 of the GDPR has the benefit and advantage of being "use case" agnostic and future-proofed in a manner which is extremely protective of individual human rights.

In a similar fashion, the limitations and/or prohibitions on surveillance by AI systems as set out in Articles 5.1 (d), 5.2, and 5.3 have extremely narrow application (i.e., the use of "real-time" biometric identification systems in publicly accessible spaces for the purpose of law enforcement), and even these prohibitions are subject to significant exceptions. It is commendable that the draft Presidency Compromise text has widened this prohibited-use case to potential instances of monitoring by third-party private agencies on behalf of law enforcement authorities (in addition to such activities by law enforcement authorities per se). However, we remain concerned that this definition does not successfully encompass all of the mechanisms of state and government, or indeed overly intrusive and/or unnecessary private sector surveillance.[3] In the Presidency Compromise text, we are concerned to see the addition of "national security purposes" to the exclusions from scope of the AIA.[4]

⚠ **Caution 1.2: The public policy reason for prohibiting surveillance activity in Articles 5.1 (d), 5.2, and 5.3 should be to reinforce democratic human agency and provide a check and balance on state power as well as on overly intrusive private sector "surveillance capital." We consider that the addition of national security purposes as an exclusion—in addition to the failure to encompass private sector surveillance (other than in the limited context of third party agencies acting on behalf of law enforcement authorities) together with all of the potential mechanisms of state and government—to be significant and fundamental weaknesses in the proposed legislation.**

Work and Automation

Principle 1 of the Responsible AI Framework specifies full enfranchisement of employees in the workplace in relation to the introduction of AI-related technologies. Whilst the AIA falls someway short of this ideal and does not allow for employee consultation on implementation, we are pleased to see that the concept of AI technologies being used in the workplace (and indeed in terms of Educational and Vocational

[3] See *The Age of Surveillance Capital: The Fight for a Human Future at the New Frontier of Power,* Zuboff, 2019.

[4] See Article 2.3 as amended.

Training) is acknowledged by the EU as sufficiently significant to merit categorization as "High Risk" AI in Annex III of the AIA.[5]

As currently drafted, the Presidency Compromise text refers to AI systems for specific "High Risk" use cases such as screening, recruitment, evaluation for promotion, or termination of work relationships.

→ **Recommendation 1.2: The AIA appears to overlook the potential for the use of AI systems in the workplace to have a detrimental impact on the psychological health of the human workers themselves (irrespective of the particular use case concerned), both by denying them dignity or fulfilment of purpose or by increasing the potential for workers to be involved in repetitive and menial tasks.**

Environmental Impact

The Environmental questions raised by the use of AI technologies are significant: machine-learning systems need massive computer-processing power, which in turn needs a huge amount of energy. Any decision to deploy AI will be a delicate balancing act between the demonstrable benefits of a particular use case as against the embodied carbon generated by its development and operation. These are concepts which are fully recognised in Principle 1 of the Responsible AI Policy Framework.

We acknowledge that the Presidency Compromise text enhances the position of AI systems which are designed to manage critical infrastructure and the protection of the Environment,[6] including AI systems designed to be used to control emissions and pollution,[7] and explicitly includes these as "High Risk" systems, now subject to the High Risk compliance regime in the AIA. Article 64 also requires the Providers of AI systems to report "Serious Incidents" (now expressly including environmental incidents) to national market surveillance authorities. In addition, Article 47 allows market surveillance authorities to derogate from conformity assessments under Article 43 in the interests (inter alia) of environmental protection[8], and Article 54 provides for environmental protection–based use cases in its proposed Regulatory Sandbox.[9]

⚠ **Caution 1.3: The AIA falls some way short in mandating responsible, environmentally friendly use of AI systems. Whilst the legislative measure is not designed to be one that dictates environmental protection, it should at least recognise the environmental cost involved in the use of all such technologies. Environmental issues are currently dealt with by the AIA on a narrow use-case basis only, without an explicit acknowledgement to the environmental impact of the use of all AI systems. We would recommend that the EU have regard to the latest draft of the**

[5] See sections 3 and 4 of Annex III.

[6] See Annex III, section 2.

[7] See Annex III, section 2(a).

[8] See Article 47.1.

[9] See Article 54.1 (a) (iii).

UNESCO guidance on Artificial Intelligence and Ethics, which places environmental use of AI systems at its heart.[10]

→ **Recommendation 1.3: We are disappointed that the environmental "cost-benefit" analysis is only reflected to a limited degree under Article 69 of the AIA as a potential candidate for a code of conduct to be adopted on a voluntary basis. Given the power consumption of such technologies, environmental impact should be a key consideration in evaluating the benefit of their introduction and should be reflected in the technical documentation requirements of Annex IV.**

Weaponised AI

Principle 1 of the Responsible AI Framework calls for the inclusion of lethal autonomous weapons systems (LAWS) within established conventions and human rights laws, and in particular for the banning of such systems to the extent they make autonomous kill decisions.

These systems (together with any use of AI in a military or national security context[11]) are expressly excluded from the scope of the AIA. It is clear that the reason military applications are excluded is due to the inability of the EU to legislate in that particular area as it lies within the competence of the individual EU Member States and is the exclusive remit of the Common Foreign and Security Policy regulated under Title V (Article 24) of the Treaty of the European Union. Notwithstanding this logical exclusion on the part of the AIA, we are of the view that this exclusion of military and national security applications allows for a much wider exclusion from scope than is necessary.

For example, it would be comparatively simple to add that the military (and now national security) exclusion only applies to the extent such systems are developed for and on behalf of the EU Member States. Military and national security AI-based systems developed for export outside the EU are also caught by the current exclusion from scope, which creates in our view an entirely avoidable and undesirable position where such technologies can be developed by private enterprise within the EU without regard to the principles in the AIA to suppress populations or sustain military conflict in third countries, provided they are exported away from the European Union. Recital 12 of the AIA as amended seems to contemplate that AI military systems developed for purposes outside Member State use (and hence the Member State competency) should be within the ambit of the legislation, but this is not reflected in Article 2.3.

→ **Recommendation 1.4: Article 2.3 of the AIA and its associated Recitals should be aligned with one another, and they should be amended to emphasize that certain military AI systems are excluded solely on grounds of Member State competency, and where such systems are commissioned and/or deployed by the Member States.**

→ **Recommendation 1.5: We encourage the EU institutions and bodies, including the European Defence Fund, to pass and act upon rules for the participation, codes of conduct, or other**

[10] See https://en.unesco.org/news/unesco-member-states-adopt-first-ever-global-agreement-ethics-artificial-intelligence.

[11] See Article 2.3.

"soft law" that foster the principles of responsible AI in the military context. In doing so, the EU should apply higher and stricter standards than those adopted as part of the NATO Artificial Intelligence Strategy which contains significant ambiguities, for instance as regards bias minimization as opposed to avoidance of bias.

False or Misleading Information

Finally in this analysis of the AIA in the context of Principle 1 of the Responsible AI Framework, we examine the issue of false and misleading information. Article 52.3 introduces a disclosure requirement in relation to so-called "deep fakes" which we welcome. We question in practical terms how such a provision could be enforced given the prevalence of such content on the internet and its origin, which in the vast majority of cases, we would determine to be from outside the EU.

The wider issue of weaponized information remains largely untouched in terms of regulation by the AIA. This concerns us given the detrimental impact "fake" news and misinformation can have on democratic societies (manifested recently and tragically in the context of the current war in Ukraine), and on public health (in the context of anti-vax disinformation). Quite how such fake content, propaganda, and lies can be addressed on a regulatory basis is likely to be a very difficult question.

→ **Recommendation 1.6: In the context of addressing misinformation and online lies, we recommend that—at the very least—AI systems should not be used to actively propagate such content (through algorithmic reinforcement).**

⚠ **Caution 1.4: Other than Article 52.3 on the use of "deep fakes" and the relatively narrow categories of Prohibited AI set out in Articles 5.1 (a) and 5.1 (b), the AIA really does not tackle in any meaningful sense information which has been "weaponized," distorted, or manipulated to serve particular agendas that could ultimately harm individuals, groups of individuals, or democratic institutions.**

Principle 2
Accountability

Organisations that develop, make available or use AI systems ought to be accountable for the consequences of their actions and shall designate an individual or individuals who are accountable for the organisation's compliance with the principles of the Policy Framework for Responsible AI or other adopted principles (including analogous principles that may be developed for a specific industry) with the objective of keeping humans behind the machines and AI Human centric.

Principle 2 of the Responsible AI Framework is intended to build and reinforce human accountability for the use of AI Systems. It does this by calling for robust governance and oversight with identifiable humans "behind the machines," as well as the carrying out of a Responsible AI Impact Assessment to fully determine and evaluate the consequences of using and deploying a particular AI system. Organizations using AI are also required to develop internal competencies and capacity in such systems by way of staff training and communications. Finally, Principle 2 rejects the concept of machine personality. For the full text of this Principle of the Responsible AI Framework, please see page 96.

Overall Approach to Accountability

We are pleased to see that at an organisational level, the AIA calls for risk management and quality management systems to be put in place with respect to High Risk AI, and it ultimately seeks to keep humans central in the governance and oversight of such AI. In general terms, greater transparency of organisational information is required, and it must be shared with authorities, notified bodies, and users, such as that seen in Annex IV in the technical documentation requirements. In addition legal (and natural) persons across the AI supply chain and ecosystem are sought to be held accountable under a new regime for the AI that they put into service or make available in the market.

→ **Recommendation 2.1:** **We consider there are areas where the AIA is not attuned to the commercial realities of the AI ecosystem, the plight of the start-up or the SME, and the overall cost burdens of compliance. The current draft AIA does not encourage a culture of responsibility taking within organisations nor does it appropriately hold entities to account. The AIA should be amended to reflect the importance of embedding a responsible AI culture at an organizational level.**

Affirmation of Human Accountability

Article 14 of the AIA emphasizes the need for high-risk AI systems to have effective oversight by natural persons. However, the Regulation does not take a clear stand concerning AI and legal personhood. The AIA should expand on the need for human oversight by expressly stating that those who can be held responsible for the acts and omissions of an AI system are, and always will be, human. There should be no room for ambiguity or confusion that might lead private actors to believe that they could blame an AI system for any non-compliant conduct, particularly for any violations of fundamental rights.

There is also no reason to limit the requirement for the need for human oversight to only high-risk AI systems. While the actual nature and extent of human oversight may differ, oversight should be extended to all forms of AI regardless of its criticality.

→ **Recommendation 2.2:** **The AIA falls short of identifying those persons/roles who can be held responsible for the acts and omissions of an AI system There should be no room for ambiguity or confusion that might lead private actors to believe that they could blame an AI system for any non-compliant conduct, particularly for any violations of fundamental rights. The AIA should have a clear and unambiguous statement denying any independent legal status for AI systems. We also recommend that the AIA affirm our principle that there should always be a human behind the AI regardless of the criticality of AI.**

Regulatory Oversight

For each Member State, the AIA provides for multiple or overlapping governmental authorities in the ongoing regulation of AI, including national supervisory authorities (Article 3.42), notifying authorities (Articles 3.19, 30), and market surveillance authorities (Articles 3.26, 26, 43). It is currently unclear how the bodies will interoperate and communicate—within Member States, across Member States, and with non-EU authorities—to provide sufficient certainty and consistency for businesses developing AI systems for use throughout the EU, to ensure that the appropriate body or regulator is tasked with enforcement and/or punitive action, and to avoid unnecessary and expensive duplication of effort.

Having multiple authorities oversee activities relating to AI has the potential to create additional bureaucracy with respect to approvals, jurisdictional conflicts, and potential over stepping of regulatory domains. It also has the potential to create multiple regulatory regimes which would need to be complied with and could result in over-regulation.

Accountability to one body or several bodies needs to be clearer. The potential exists for multiple fines to be imposed. For instance, for a single violation, fines could be imposed by EU Data Protection Authority, EU AI Authority, and a sectoral regulator in several countries. Too much regulatory overlap and the potential for multiple punitive fines and disciplinary action could undermine accountability, turning a significant but manageable problem into a company applying for insolvency and liquidation to avoid the extensive legal proceeding and expenses.

At the same time, the existence of multiple regulators could create an "enforcement paralysis" leaving each regulator relying on other regulators to act, and no enforcement taking place at all. Without clear accountability and allocation of duties, and coordination across Member States, there can be no effective accountability to regulators.

The AIA also permits differential treatment between manufacturers or providers who are EU members and those who are non-EU members in the extent of oversight, whether approached by one body for the whole of the EU or several member states bodies. This creates a potential for anti-competitive and exclusionist practice.

→ **Recommendation 2.3: We consider that if a multi-tiered regulatory compliance structure is to persist that there should be greater collaboration, coordination, cooperation, and communication, not just between the EU institutions and regulatory bodies across the EU but with governments and regulatory bodies outside the EU. Without such collaboration and coordination, the AIA may have the effect of potentially exporting EU values and the so-called "Brussels Effect" in an undesirable manner, raising some of the same extraterritoriality concerns for the AIA as has been seen as the EU GDPR has evolved.**

Roles and Responsibilities of Private Actors in the AI industry

⚠ **Caution 2.1: The draft AIA does not effectively account for AI supply chain and/or AI components (whether procured through "AI as a Service" offerings or not) despite further revisions provided for in the Presidency Compromise text. The lines between manufacturer, distributor, importer, provider, and user can become blurred. When AI is put onto the market without either a specifically intended purpose or where the AI can perform generally applicable functions (so-called "General Purpose" AI), demarcation of responsibility could be difficult to unpick and become confused (Recital 70a and Article 52a). Even where the user does not modify the AI (by only taking AI or a pre-trained model "off the shelf"), the regulatory burden could potentially fall on that user as if they were a provider. This may have a disproportionate effect.**

For example, Article 24 can, under some circumstances, place duplicative duties on providers and product manufacturers. Another example of the lack of clarity can be seen in Article 28. In Article 28, certain changes in the purpose or operation of a high-risk AI system can shift the provider-level duties to a distributor, importer, user, or other third-party. This begs further questioning: How, when, and why ought someone other than the provider (someone who puts AI on the market or puts AI into service) be held accountable? How much of a modification in intended purpose is needed to shift provider duties under Article 28(1)(b)? What constitutes a substantial modification under Article 28(1)(c)? While either of these changes might justify shifting some duties onto the actor making the modifications, why should a modification eliminate all duties of the provider of the AI system under Article 28(2)? What about the practicalities of actors other than the provider complying the technical documentation and conformity assessment requirements? How, who, and when will these issues be managed? What disclosures should a revised provider be able to expect, particularly concerning the original training data and understanding of bias that was taken into account during design and development?

Part of the difficulties arise from the draft AIA treating an AI system software, data, and data models as a "product" under product legislation, when it is, in most respects, an intangible good or service. With complex supply chains, the lack of a clear delineation of roles and responsibilities between ecosystem players can result in: (a) confusion as to who IS responsible for compliance, (b) unnecessary duplication of compliance and conformity assessments, and (c) an inefficient "clogging up" of the regulatory system.

The complexity of the AI lifecycle and AI ecosystem, and lack of clarity in roles and duties could also mean that risk apportionment and allocation of liability for accountability in contracts will be overly complicated and unwieldy to manage.

⚠ **Caution 2.2:** **Who bears primary responsibility under the AIA needs further clarity, particularly in respect of "AI as a Service" or MLopS, with regard to how responsibility and duties (and ultimately liability) do and do not change over time and as circumstances change. If this remains unclear in the AIA, this will create uncertainty for businesses, and may have an undesirable and/or innovation-stifling effect.**

→ **Recommendation 2.4:** **The AIA should provide a clearer picture in relation to various ecosystem roles and responsibilities and clearly demonstrate how these may change when the AIA adds or shifts provider-type duties to another actor or creates interdependent duties.**

We recognise that the EU is live to these concerns and may reflect updated liability laws in relation to AI in a new, related regulation. Yet as mentioned in our comments under Principle 1 of the Responsible AI Framework, we would welcome clear guidance and signposting from the EU as to the nature of the relative interdependency and how this is likely to be achieved.

General Purpose AI and Scientific Research and Development Exemptions

⚠ **Caution 2.3:** **The new Presidency Compromise text introduces two further exemptions from the scope of the AIA. These are AI developed and launched which is "general purpose" in nature (see the new proposed Article 52A) and AI which is either used for Scientific Research and Development purposes or any research on AI (to the extent such systems are not placed onto the market) (see Articles 2(6) and 2(7)). These new exemptions seem to us to create a significant gap in the ability of the AIA to ensure accountability across the EU AI ecosystem. The concerns stem from the fact that the exemption seems to (a) side-step the potential consequences of the use of such unregulated systems and (b) ignore the role that the actual developers of such systems (as opposed to legally identified "producers" under the AIA) in using best practice to ensure basic algorithmic hygiene, utilise clear system architectures, and manage training probity. This would appear to us to create a fundamental accountability problem in that these parties are presumed to escape liability or consequence for such systems despite the fundamental role they have had in their creation.**

Governance and Risk Management

Whilst the tools and mechanisms for AI governance and human oversight within organisations are not prescribed by the AIA, the AIA seems to expect such measures to be in place and to be effective for accountability purposes. This is not a safe assumption to make and it is not clear whether this originates from existing practices in the market. On the contrary, there is enough evidence to suggest that there are a lack of such accountability measures within organisations currently. There appears to be little scrutiny of what internal governance or oversight functions are actually in place in organisations, or whether or not they could reasonably perform the duties expected of them. Furthermore, there is no evaluation as to whether the human oversight itself is competent, capable, and has the capacity (time, money, effort, and resources) appropriately ascribed to it to help govern and meaningfully oversee AI systems deployed by it effectively.

→ **Recommendation 2.5: We consider that further guidance, and significant scrutiny of AI governance and oversight within providers and other actors, should form part of the AIA in order to ensure that organisations maintain effective accountability systems internally as well as comply with the requirements to permit external accountability.**

Holistic Disclosure

Nothing can be held to account without the necessary transparency and explainability to understand how the system is operating and why it operates in the manner that it does. But the information needed to hold organisations providing or using AI to account requires more than just source code access or making trade secrets available in an AI Database. The general standard in Article 11 that technical documentation contains "all the necessary information to assess the compliance of the AI system" needs elaboration on what information must be included. For example, clear and understandable documentation is needed as to:

- The intended purposes of an AI system, and its limitations (this is also needed to manage down-the-line liability risks of providers)
- process transparency
- decision transparency
- transparency as to people and diversity issues
- data representativeness transparency
- details that due diligence checks have been undertaken in respect of the supply chain
- details that bias has been identified and how it has been mitigated or managed
- the allocation of roles and responsibilities within the organisation for AI governance and oversight, with sufficient details to demonstrate that the organisation is governing and monitoring its AI well and responsibly

- details on how the organisation maintains human oversight which is meaningful and abates automation bias
- model transparency
- outcomes of the AI system have been checked as to their reasonableness and proportionality
- details of the economic drivers that led to certain decisions, risk appetites, or business models.

While illustrative, this is still not an exhaustive list.

Although code can be scrutinised/audited, the current draft AIA does not appear to require the code to be held in escrow and updated at intervals. Since AI is application agnostic, there is nothing to stop an organisation from using AI code in one context for a given purpose which is non-compliant or not legitimate, halt use in that context, and then re-purpose the AI into another context. Nothing as stands would prevent that "phoenix syndrome" from happening. It therefore does not appear to hold the organisations sufficiently accountable for the code, models, and training data they use in context. This kind of behaviour and potential poor practice could be further proliferated if a provider simply designated their AI "General Purpose AI" in order to further avoid regulatory burden.

⚠ **Caution 2.4: The right of a notified body to carry out periodic audits should be evaluated to enhance the role of regulators beyond the confines of mere management system quality assurance, and to identify and scrutinise key aspects of AI systems for potential violation of fundamental human rights.**

→ **Recommendation 2.6: The information that must be detailed in the technical documentation should be improved and tailored to a supervising regulator and/or notified body, given the commercial sensitivities of the disclosure. Original AI code and training datasets should be held in escrow to allow assessment of systems in their original, pre-deployment form and also to ward off phoenix syndrome–like behaviour from AI businesses.**

Conformity Assessment Process

The conformity assessment requirements set forth in Articles 19 and 43 and Annexes VI and VII play a crucial role in ensuring AI systems function in a compliant manner before they go into widespread use. While appropriate regulatory penalties and individual means of redress (as discussed below) are important parts of any regulatory scheme for AI, providing for a rigourous review of AI systems before they go into effect can help to avoid the harm that later consequences can only seek to deter.

⚠ **Caution 2.5: Unfortunately, the rigour of expected compliance with the conformity assessment requirements is questionable. What the AIA says should be done and what is actually done in practice may differ. Without any scrutiny of an organisation's conduct with respect to AI through a third-party "notified body" assessment or through consideration of user feedback, any assessment conducted internally by an organisation may result in a flawed assessment**

that is based on a closed feedback loop or is biased. Depending on organisations to conduct their own conformity assessments risks the compliance process itself lacking rigour.

Secondly, an approach that permits a self-determined conformity assessment for organisations which put into service or place on the market High Risk AI systems for any purpose stated in Annex III, except for biometric identification and categorisation of natural persons, is self-limiting and problematic. Some of the items in Annex III such as (5) access to and enjoyment of essential private and public services and benefits, (6) law enforcement, (7) migration, asylum, and border control management, and Annex III (8) the administration of justice and democratic processes are quite critical from a fundamental human rights perspective. To limit the process to self-assessment, without necessary checks by an independent authority may lead to function/mission creep and excessive use of powers. For instance, a police department being the law enforcement agency will not be able to independently self-assess the impact of their systems involving human rights. These are all areas where AI systems have the potential to cause substantial harm to a person, people groups, their livelihood and wider society, and to undermine the important institutions of government and public trust—just as would AI systems involving biometric identification.

We also have concerns regarding the provisions of notified bodies as to biometric identification systems. We appreciate that draft Article 43(1) requires that, when such systems are to "be put into service by law enforcement, immigration, or asylum authorities as well as EU institutions" and a notified body is involved, then the designated Market Surveillance Authority must act as the notified body. However, relying on a Market Surveillance authority and/or national competent data protection supervisory authority to oversee or approve an AI system in the context of law enforcement and other agencies that are part of the same government that pays that authority creates room to question the independence of the authority and the reliability of the conformity assessment it performed. Therefore, there must be adequate checks and balance to ensure the independence of the assessment.

→ **Recommendation 2.7: The categorisation of types of AI systems that require a conformity assessment involving an independent notified body should be expanded. The list should not just include biometric identification and categorisation systems (draft Annex III para 1) but should also include draft Annex III paragraphs 5 through to 8 as well. Independent use of notified bodies creates further accountability. If independent notified bodies are not suitable for reviews to be performed by a Market Surveillance Authority, then additional safeguards should be implemented to maximize the independence and reliability of the conformity assessment.**

High Risk vs. Other AI Systems

⚠ **Caution 2.6: The bifurcation of the proposed AIA for High Risk and other AI systems overlooks a range of circumstances in which intermediate levels of oversight and regulation may be needed to ensure proper accountability—particularly when the costs of regulatory compliance can create barriers to entry for start-ups, scale-ups, and SMEs—and invite the use of anti-competitive practices by increasingly large market players.**

Just as there are circumstances where more rigour is needed for systems designated as high-risk under the Regulation, there are also certain circumstances where less rigour should be required. There will also be AI systems that would be subject only to a voluntary code of conduct under Article 69 that may merit some level of required oversight. A regulatory system for AI cannot be tailored to the circumstances of every possible AI purpose or every possible provider or user of AI, but the system should have sufficient flexibility to provide some proportionality between the desirable level of regulatory oversight and the prospective harm that might result from a given AI system or version of a system.

At the same time, the costs of compliance cannot be set so high as to prevent the development of beneficial AI systems. The financial cost, and the opportunity cost of the time and effort spent on compliance, could easily form a barrier to entry or be cost-prohibitive to start-ups, scale-ups, and SMEs. Unduly high regulatory costs might also have an anti-competitive effect and drive some small actors to either (a) want to be acquired, or (b) not to innovate in high-risk, high-potential markets and to direct such activity to other (less regulated) markets and jurisdictions.

Further efforts are needed to identify when, and under what circumstances, regulatory objectives can be achieved with reduced costs of compliance. For example, it is not clear whether compliance with conformity assessment is iterative. That is, when there is a revision, expansion, or modification of an existing AI system, Article 28 would appear to require a new full-scale conformity assessment to take place when the costs of conformity (and the requirements to be undertaken) necessary to satisfy regulatory objectives could be significantly less, with testing focused on the deviations from the existing conforming model.

On the other hand, creating too loose a system for establishing iterative compliance for “licensing”/permission or certification purposes could put other market players at a disadvantage relative to competitors whose systems have already gone through the process. Processes need to be developed for conducting an algorithmic audit for each new iteration of an AI system depending on how the system operates and to systematically assess the possible impacts that changes in the system might have for the user and/or other stakeholders.

Intermediate measures for ensuring regulatory compliance might also be based on the use of standardised datasets or adherence to certain established protocols or approaches in development of a new AI system. Creating “safe harbours” or short-cuts for establishing conformity of an AI system—or certain components of a system—would encourage the development of a market for more efficient and compliant AI tools and allow oversight to focus on the novel aspects and implications of a new AI system.

→ **Recommendation 2.8: Measures and processes should be developed to reduce the burdens of regulatory compliance, particularly for smaller businesses, by identifying circumstances where reliance on existing systems, datasets, or approaches reduce the risks of new or modified AI systems. Regulators should be cognizant of the anti-competitive effects that the draft AIA and any revisions to the AIA might engender. The AIA should also consider if there are areas that do not qualify for inclusion as a “High-risk AI,” but that should not be left entirely to voluntary compliance by providers.**

Persistence of Duties Imposed on Actors

The persistence of duties imposed on a provider, distributor, or importer of an AI system needs to be tied to the market use of that system, and processes should be established for the sunsetting, decommissioning, and dependency of AI systems.

Factors and risks concerning sunsetting were raised in our Responsible AI Impact Assessment. The draft AIA, in some places, adopts standards for the duration of certain obligations that are divorced from the economic reality of the AI systems they seek to regulate. Most notably, Article 50 places a fixed 10-year obligation on providers to retain the specified information without consideration of whether an AI system may have been taken off the market years before. Conversely, a successful AI system might, perhaps with updates, remain active in the market for more than a decade. The AIA thus ought to address circumstances where decommissioning or sunsetting of a given AI system occurs either intentionally or unintentionally, whether because of non-conformity, due to a lack of user adoption, from problems in maintaining or sustaining the system, or simply because an AI system has become outdated and/or unreliable in the absence of sufficient support. Consideration should also be given to what steps providers can, after an appropriate period, take to bring their duties with respect to given models of AI systems to a close and to what obligations providers might still owe to users who wish to continue to operate legacy systems.

→ **Recommendation 2.9: The draft AIA needs to expressly provide for the decommissioning and sunsetting of AI systems, building in de minimis legal requirements. Practical and ethical steps are also needed to mitigate risks and the impact of business and social dependency in the event that an AI system is sunsetted or decommissioned.**

Effective and Proportionate Remedies

The deterrent effect of regulatory penalties for non-compliance must balance being sufficient to deter misconduct while being not so disproportionate to undermine the development and economic feasibility of socially beneficial AI systems.

Article 71 of the draft AIA authorize fines of up to 30 million Euros or 2%-6% of a company's worldwide annual turnover. While paragraph 1 of that Article does call for Members States to adopt rules for penalties that are "effective, proportionate, and dissuasive," where there are "big stick" approaches to fines, the threat of such fines can, despite good intentions, dissuade desired activity or, if imposed, destroy a technological platform and/or an ecosystem supporting that platform. Where business-to-business and/or business-to-consumer economic models have been established in a given technological platform, there is an increased risk of business and consumer dependency on that platform and its ecosystem. Imposing a disproportionate fine on such AI platforms and providers could have a destructive effect that ripples through the wider economy and society. The draft AIA also does not provide for any interim or gradients of fines that might be imposed prior to the larger fines proposed in the law. Leaving nearly unfettered discretion to Member States to develop penalty schemes risks unnecessary variation among Member States and may undesirably distort the behaviour of private actors seeking to minimize their liability.

→ **Recommendation 2.10: The draft AIA needs to expand on and develop the principle of proportionality in the application of administrative fines and provide a sufficient framework or process to ensure reasonably harmonised schemes of penalties. Penalty schemes should also take account of and/or provide guidance concerning situations of dependency and the downstream effects of imposing sizable fines on providers, importers, or distributors.**

Private Individual Remedies

⚠ **Caution 2.7: The draft AIA lacks a suitable "one stop shop"-style redress mechanism that can provide an effective and available means for all those injured by an AI system to seek compensation for their injuries.**

Whilst redress for businesses and individuals harmed by AI systems is not specified in the draft AIA, it appears they rely on the existence of unspecified existing or future laws to provide the essential vehicle or mechanism for such redress to be sought or achieved. Conformity assessments can help to prevent harm, and administrative penalties may have a deterrent effect, but neither can replace the need to compensate injured end users (whether they be a citizen, consumer, service/product user, or contributor) or a non-user or other stakeholders who are impacted or influenced by the AI system. AIs may also cause harm at a societal level, involving persons unknown to one another and who may even lack common characteristics but still share in a group harm. This could be in the form of violations of Human Rights, unequal or discriminatory treatment, infringements on privacy (data privacy outside of that concerning data protection), or claims for transgressions of consumer law, product liability law (particularly for tangible component parts), competition or anti-trust law, administrative public law, and/or industry-specific regulation. AI may also raise entirely new issues not encompassed by current legal regimes. To this end, a right to contestability and redress is essential.

Simply creating a legal right to recover is not enough to provide effective redress. If someone has a legal right to damages, the right is meaningless if they do not have: (a) the ability to pursue the claim due to language or jurisdictional barriers, (b) a realistic means of proving that the AI system failed or caused specific harm, (c) the resources needed to take advantage of the legal opportunities to hold AI systems to account (whether through funding litigation or by identifying others who have experienced similar harm and can share the burden), or (d) the knowledge of who to pursue because the party responsible for the challenged operation of an AI system may be difficult to identify, whether due to a multiplicity of components, white-labels remarking, or modifications made after the providers original creation of the system, or the use of general purpose AI. Even if an injured consumer or business can surmount these obstacles, there is no assurance that the court or regulator adjudicating a dispute would have the technical expertise needed to reach a just outcome.

→ **Recommendation 2.11: The draft AIA needs to provide (or make reference to via external legislation) (1) a mechanism for queries from potentially harmed individuals to be investigated across regulated and unregulated arenas, because deployment of an application-agnostic technology may arise in many varied fora which overlap sectors, and (2) a mechanism which**

> **makes justice, equity, restitution, and (if necessary and appropriate) damages for individual and/or group harm accessible to all.**

Although creating a scheme to provide effective redress will not be easy, the current AIA (or accompanying legislation) offer a real opportunity to address the power asymmetry between manufacturer/providers/ platform providers and user/end users, and to rectify outcomes which have caused biased, unfair, or unjust results. This could be through the use of evidentiary or legal presumptions, shifted burdens of proof, enhanced or streamlined discovery tools, and/or specialized tribunals.

Principle 3
Transparency and Explainability

Organisations that develop, make available or use AI systems, and any national laws or industry standards that govern such use, shall ensure that such use is transparent and that the decision outcomes of the AI system are explainable.

Principle 3 of the Responsible AI Framework mandates an approach to transparency and explainability which forms of the foundational base of the other principles of the Responsible AI Framework. In terms of AI transparency, it calls for decisions made by AI systems to be fair and impartial, to support human agency and autonomy, and to ensure that meaningful responsibility is passed on to the developers and users of AI systems. AI explainability should be embedded by design, in so far as practicable, and should be capable of iterative improvement as the state of the art improves. The Principle is application-agnostic: it applies to AI systems which create legal effects on the user or affects users in a similarly significant manner, and it requires a contextual approach to transparency and explainability which is commensurate with the impact the application could have on the user and their expertise. For the full text of this principle, please see page 98.

Clarification on Usage of "Transparency" and "Explainability"

We would note at the outset that there are significant industry-based differences in which the terms "transparency" and "explainability" are used.

For the purposes of this analysis, we refer to both concepts consistent with the definitions of Principle 3. AI transparency in that context refers to the use of AI being "transparent" in its usage to those who are the subject of an AI decision. AI explainability, in contrast, overlays onto transparent usage a requirement to ensure meaningful interpretability of the algorithmic logic of an AI system.

We distinguish this usage from the alternative industry approach of referring to AI transparency as a counterpoint to AI opaqueness. Generally speaking, AI transparency, when used in this context, is a technological design step which facilitates enhanced human interpretability of a machine learning system.

→ **Recommendation 3.1: Usage of the term "transparency" within the AIA varies depending on context and the relevant article. Articles 12 and 13, for example, refer to "transparency" but in fact appear to be more concerned with the intelligibility of machine learning decisions (or explainability as it is defined by Principle 3). Article 52, in contrast, also refers to "transparency" but in a manner which is more consistent with the Principle 3 definition. We recommend that this definitional uncertainty is corrected in future drafts of the AIA.**

Overall Approach to Transparency and Explainability

In general terms, we note that the concept of explainability (as that relates to improving the intelligibility of machine learning–based technologies) have been reflected in the draft AIA. Article 12 of the AIA introduces the concept of Record Keeping—in particular, the requirement for high-risk AI systems to automatically log events in an AI system's lifecycle to ensure levels of traceability in its functioning, as well as Article 13 which (whilst referring to transparency) mandates particular levels of explainability, both in general terms and specifically in relation to systems which are intended to interact directly with natural persons. We note that whilst the AIA concentrates on AI system explanability, there is a notable failure to legislate in the draft on wider AI transparency obligations—that is to say, an obligation to provide users with the parameters of the decision making affecting them in human-interpretable terms. This omission unfortunately coincides with the lack of individual personal remedies under the AIA, as we have mentioned elsewhere in this Green Paper.

⚠ **Caution 3.1: There are few substantive AI transparency obligations within the AIA within the definition of Principle 3. The AIA is light on information that must be disclosed to the people who are affected by AI systems. We would question the alternative focus in the draft measure on AI explainability which is not a complete substitute for transparency obligations and may in fact be technically very difficult to achieve in the context of some AI systems based in deep learning.**

High-risk AI systems include systems on biometric identification, education (including access to education), recruitment of employees, public services, law enforcement, etc. However, Article 13 only obliges providers of high-risk AI systems to make the AI system sufficiently transparent towards its users (i.e., to enable them to interpret the system's output and use it appropriately). We are concerned that an undue focus on AI system transparency seems to miss the point when it comes to the issue of AI transparency and the right to understand how decisions have been made.

If such systems are seen as a high risk to society, then why should the individuals that are the subjects of their outputs not be informed of the outcomes of use in their cases? For instance, why should a job applicant not know that the recruitment process is (partially) done by an AI system? In accordance with principles set out in the Responsible AI Framework, a person should know that he is undergoing an automated AI process. But the obligation towards that category of third parties is completely missing in the AIA. In contrast, the transparency obligation with respect to the three classes of AI system within Article 52 (i.e., systems intended to interact with natural persons, emotion, and/or biometric categorization systems and image manipulation systems) are designed in a way that individuals affected by their outputs are informed.

→ **Recommendation 3.2: Heightened transparency requirements should be generally applicable to all regulated high-risk AI systems and at least equivalent to those in Article 52.**

The AIA does not treat the algorithms used in social media, search, online retailing, app stores, mobile apps, or mobile operating systems as high risk. It should also be noted that Article 52 does not give the European Commission the power to change this list of AI systems, as opposed to other provisions which do so. Moreover, the transparency (communication) obligation only relates to a notification that an AI system is being used.

In the context of Article 13, we would argue that this provision is in general vaguely worded and open to interpretation. It is not clear in what way users can sufficiently interpret the system's output in order to use it appropriately. The AIA does not define what exactly the threshold of interpretability of an AI's outcome or decision should be.

⚠ **Caution 3.2:** **Article 52A on "General Purpose" system exemption seems to us to move the AIA further away from an approach which should extend at least some transparency obligations to all AI systems and consequently further away from the goal of enabling users of AI systems to understand the manner in which decisions have been made on a consistent basis. We view it as a retrograde step.**

→ **Recommendation 3.3:** **We would like to see some element of transparency by design as reflected in Principle 3 of the Responsible AI Framework extended to the development of all AI systems, regardless of classification. Standard, minimal transparency obligations could be designed for mid- and low-risk AI systems.**

Relationship between Transparency and Information Provision

A distinction is made in Article 13 between transparency and providing information to users. We could argue that the latter is part of the transparency requirement, cf. Principle 2.1 of the Responsible AI Framework (see page 96). Information provision should be seen as part of the transparency requirement. The correlation between the two obligations—information provision and transparency—is not clear. The text of the legal provision implies that the concepts are very related but still separate: the Article's title uses separate names and different parts of the Article are devoted to two concepts (paragraph 1 for transparency, paragraphs 2 and 3 for information provision). To avoid a lack of clarity in applying the two concepts, it could be argued that the AIA should continue the line taken in the previous AI policy documents and consider information provision as part of the general transparency obligation.

→ **Recommendation 3.4:** **A hierarchy should be established within the AIA to minimise the inconsistencies between transparency and information keeping concepts.**

Public Disclosure of Technical Documentation

It is arguable that the AIA should be more overt in allowing public release of technical documentation prepared in accordance with Article 11. Annex IV includes information that could be important to ensure transparency within a particular AI system. We would submit that more public access to the information could grow the public's trust and understanding of the system, but recognise that such disclosures would need to be necessarily constrained in some cases due to the need for appropriate intellectual property protection.

→ **Recommendation 3.5:** Further thought should be given to public disclosure issues by the EU. There are evident inconsistencies in disclosure levels in the current draft of the AIA which could harm public accessibility to justice for individuals harmed by AI decision making. For example, the documentation that arguably must contain a bias assessment does not need to be provided to users, the public, or those potentially affected by discriminatory algorithms. It is available only to regulators upon request. In contrast, the legislation clearly mandates assessments and disclosure of system accuracy.

Private Individual Remedies

In what is a running theme throughout our Green Paper commentary, we are concerned at the lack of individual recourse remedies provided for by the AIA. This theme, of course, runs through and is integral to the concept of transparency. To give a particular example, certain information on human oversight measures is included in the list of information to be provided in the instructions for use in Article 13.3 of the AIA. However, more information could be provided (e.g., how humans could appeal a decision).

→ **Recommendation 3.6:** The AIA does not outline any mechanisms by which those harmed by AI systems may seek recourse and redress from the user of AI systems (with the exception of Article 52) similar to those which can be found in equivalent legislative measures such as the GDPR. We call on the EU for urgent clarification as to how these essential requirements will be met.

Principle 4
Fairness and Non-Discrimination

Organisations that develop, make available or use AI systems and any national laws that regulate such use shall ensure the non-discrimination of AI outcomes, and shall promote appropriate and effective measures to safeguard fairness in AI use.

Principle 4 of the Responsible AI Framework calls for implementing organizations and governments to promote fairness and non-discriminatory outcomes. This includes providing the means to educate stakeholders on the limits of AI systems, designing them from the outset to be fair and non-discriminatory as well as ensuring continuous monitoring and mechanisms to ensure such outcomes during the use of an AI system. For the full text of this principle, please see page 100.

General Approach to Fairness and Non-Discrimination

There is a full acknowledgement in the recitals of the AIA as to the risks of bias and discrimination in AI systems, but in reality, the measure is surprisingly thin on concrete obligations to mitigate such risks. We focus on specific areas below but, before this, we would make some general, overarching observations.

Because the AIA is structured around a product liability framework and the classification of specific-use cases (i.e., in relation to prohibited AI and high-risk AI), it becomes very difficult to efficiently incorporate an outcomes-based approach which would adequately reflect issues such as bias and discrimination, which are inherently subjective in their interpretation.

The EU has included particular-use cases that may have high potential for bias and discrimination as "high risk" within the existing framework of the AIA in order to bring them within regulatory oversight. For example, we see in the most recent Presidency Compromise text the inclusion of insurance-related services within the scope of Annex III as high-risk systems. The associated commentary for this inclusion mentions that this is appropriate given the degree to which insurance-related AI applications could have unfair or discriminatory impacts on particular demographics.

⚠ **Caution 4.1: We consider taking a risk-based approach based on use cases to be a relatively unsustainable way to proceed as the EU is now faced with having to anticipate, on a case-by-case basis, which types of AI application could have an increased potential to cause bias or unfair outcomes. A far better way of achieving this is by way of an outcomes-based approach, which is necessarily use-case agnostic.**

It is mentioned elsewhere in this Green Paper that the AIA does not adopt the *"legal effects"* or *"similarly significantly affects"* structure of Article 22 of the GDPR (on Automated Decision Making and Profiling) which would allow for the consideration of the impact of particular use of an AI system on an person. As covered in the Article 29 working party guidance for Article 22 of the GDPR, any such impact needs to be subjectively assessed by reference to the individual upon which the impact rests. What may be trivial to an adult or white Caucasian may therefore be high impact (aka discriminatory as against a child or a transgender person).

→ **Recommendation 4.1: Whilst we recognize that an outcomes-based approach may necessarily involve a significant "rethink" of the manner in which AI systems are regulated by the AIA, a failure to adequately include a subjective mechanism of assessment in terms of bias and discrimination is a significant omission. We call upon the EU to reconsider the inclusion of a subjective user-based harm-impact mechanism, such as that used within Article 22 of the GDPR.**

Prohibited AI

We should acknowledge that the AIA makes efforts to prohibit certain AI systems that may give rise to unacceptable risks in relation to providing discriminatory outcomes. So for example, it prohibits the placing on the market or putting into service or use of AI systems that deploy certain specified subliminal techniques to distort behaviours, exploit vulnerabilities of certain groups on a discriminatory basis, or unduly monitors or scores social behaviour.[1]

We would, however, comment that these specific exclusions are themselves symptomatic of the product safety–related, use-case specific approach adopted by the AIA as we mention above. There is in their prohibition tacit acknowledgement on the part of the EU that such systems may cause harm in terms of bias and discrimination, but unfortunately a lack of recognition that these outcomes may occur in circumstances that are unrelated to prohibited-use cases. In this regard, we would refer to our wider recommendation under General Approach, above.

General Purpose AI and AI used for Scientific Research and Development

⚠ **Caution 4.2: The new Presidency Compromise text allows for "general purpose" AI to be exempted from the provisions of the AIA, as is AI used for Scientific Research and Development. Again, we see these exclusions as highly problematic and indicative of the use-case methodology that the EU appears to be adopting. In our view, it overlooks the consequences that the use of such systems could have on their users in terms of bias and discrimination.**

[1] See Article 5 AIA.

Education and Documentary Requirements

Part of the approach of Principle 4 of our Responsible AI Framework is to provide general awareness to the AI stakeholder community as to the potential risks and limitations in the use and deployment of AI systems. Again, we do not see significant measures in this regard in the AIA. We do note, however, that Article 14.4 (b) mentions the need for users of AI systems to be aware of the risk of "automation bias," i.e., the tendency for the user to automatically assume that the output of the AI system is correct.

Aside from the reference in Article 14.4 (b), the AIA does provide some granularity in relation to the documentary requirements that are presumed for high-risk AI systems, in Article 11, which references a specific list within Annex IV. Documentary requirements are a clear prerequisite and step towards the educated approach that is advocated by Principle 4 of the Responsible AI Framework. However, even within Annex IV we are concerned that it is only in section 3 that a general reference is made to documentation that refers to *"the foreseeable unintended outcomes and sources of risks to health and safety, fundamental rights and discrimination in view of the intended purpose of the AI system."*

→ **Recommendation 4.2: We would like to see a clearer inclusion of bias and fairness in the documentary requirements, particularly some specificity on the part of the EU in terms of key metrics which would assist organisations in determining the characteristics of biased outcomes and potential unfairness for the AI system in question. These need to be assessed together with the identification of biases or vulnerabilities in the AI's intended or unintended use and diversity of its application domain.**

Private Individual Remedies

⚠ **Caution 4.3: Principle 4 requires that parties who are harmed by the decisions of AI systems should have effective ways to seek remedies in discriminatory or unfair contexts. We find the AIA significantly lacking in this regard, and have raised this as a substantial concern in the context of different Responsible AI Principles elsewhere in this Green Paper.**

→ **Recommendation 4.3: In so far as the documentary requirements of the AIA are concerned, the Article 11 (Annex IV) requirements that arguably must contain a bias assessment does not need to be provided to users, the public, or those potentially affected by discriminatory algorithms. They are available only to regulators upon request. In contrast, the legislation clearly mandates assessments and disclosure of system accuracy. We would strongly recommend that some form of individual access right is introduced for individuals and groups who may have been discriminated against by AI decisions.**

Conformity Assessment Procedure

The draft AI Regulation provides little direct mention of fairness. It is also unclear about the requirement to conduct and publish impact assessments which evaluate algorithmic bias. There are references to bias in various provisions of the draft AI Regulation, e.g., Article 10.5 allows providers of AI systems to process data containing sensitive properties such as race, gender, and ethnicity to ensure *"bias monitoring, detec-*

tion and correction" and Article 15.3 requires that learning high-risk AI systems be developed in such a way to ensure that possible biased outputs due to feedback loops are duly addressed with appropriate mitigation measures. (See also Annex IV, Sections 2(g) and 3.)

However, these references do not require impact assessments to evaluate bias issues. Nor is it a requirement to make available conformity assessment documentation (which might arguably include bias assessments) to users, the public, or those affected by discriminatory algorithms. The result is that the information relating to algorithmic bias may be lacking.

→ **Recommendation 4.4: As with the technical documentation that needs to be produced in accordance with Article 11 and Annex IV, we would like to see explicit further mention in the Conformity Assessment Procedure set out in Article 43 and Annex VII to bias-related testing. The same considerations as to user access mentioned above under Documentary Requirements applies to wider Conformity Assessment documentation, and we would see the Conformity Assessment Documentation as forming a fundamental part of any evidence that could be accessed and used by individuals to seek remedy in cases of discriminatory harm.**

Embedding Fairness by Design and Continuous Monitoring

The AIA does make significant efforts to embed bias monitoring into provisions which are intended to cover the training of high-risk AI models. Article 10 requires the implementation of data quality and data governance mechanisms in relation to datasets that are used with high-risk AI systems, including requirements to ensure that such datasets are representative of the persons or groups of persons on which they are intended to be used and encompass a consideration of possible biases.[2]

In so far as continuous monitoring is concerned, we also note the provisions of Article 15, which refers to the resilience of AI systems in the context of accuracy and robustness. Article 15.3 in particular provides that *"High-risk AI systems that continue to learn after being placed on the market or put into service shall be developed in such a way to ensure that possibly biased outputs due to outputs used as an input for future operations ('feedback loops') are duly addressed with appropriate mitigation measures."*

→ **Recommendation 4.5: We would like to see more details from the EU in terms of the way in which the fairness and non-bias goals in Article 10 and Article 15 are likely to be achieved in practical terms. A failure to provide any such benchmarks makes the obligations essentially self-policing by the applicable AI provider and makes it harder to determine the extent to which standards have been met (or otherwise).**

[2] See Article 10.2 (f).

Emotion Recognition and Biometric Categorisation Systems

AI-enabled emotion recognition and biometric categorization systems raise specific and unique concerns are worth noting separately under Principle 4 of the Responsible AI Framework due to the particular manner in which they are treated under the AIA.

The definitions of 'emotion recognition system' and 'biometric categorisation system' build on the definition of 'biometric data'. 'Biometric data' is defined as *"personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of the natural person, which allow or confirm the unique identification of that natural person, such as facial images or dactyloscopic data."*

The requirement that the 'biometric data' must allow or confirm the unique identification of a natural person could potentially exclude more recent applications coming within the definition of 'emotion recognition system' and 'biometric categorisation system.'

Emotion recognition systems and biometric categorisation systems are not listed under the high-risk classification in Annex III 1 (a). Both of these systems have been defined, but it is not entirely clear which risk category emotion recognition systems and biometric categorisation systems would be classified under the AIA. Some commentators have classified emotion recognition systems and biometric categorisation systems as 'low risk' and only be subject to transparency requirements.

However, it appears that emotion recognition systems could be categorised as high-risk AI systems in the context of law enforcement, according to Annex III 6 (b). Emotion recognition systems could be categorised as 'AI systems intended to be used by law enforcement authorities as polygraphs and similar tools, or to detect the emotional state of a natural person.'

Biometric categorisation has the potential to cause discrimination and other harms by casting individuals into arbitrary and stereotyped boxes, and then to make predictions, inferences, or decisions about them on that basis.

Similarly, biometric categorisation systems used in the context of Annex III 6 (e) could qualify as high-risk AI systems 'intended to be used by law enforcement authorities for predicting the occurrence or reoccurrence of an actual or potential criminal offence based on profiling of natural persons as referred to in Article 3(4) of Directive (EU) 2016/680, or assessing personality traits and characteristics or past criminal behaviour of natural persons or groups.'

Notably, Article 52(2) also provides very broad exemption from the transparency requirement where AI systems used for biometric categorisation are permitted for the detection, prevention, and investigation of criminal offences.

→ **Recommendation 4.6: Further mechanism and accountability requirements should be added to the AIA to remove any potential ambiguity and uncertainty, as the harms and impacts vary considerably depending on the purpose, scope, and context of the use of biometrics systems.**

As illustrated above, these emotion recognition systems and biometric categorisation systems could be deployed as part of the seven other high-risk AI systems listed pursuant to Article 6(2) and Annex III.

Principle 5
Safety and Reliability

Organisations that develop, deploy or use AI systems and any national laws that regulate such use shall adopt design regimes and standards ensuring high safety and reliability of AI systems on one hand while limiting the exposure of developers and deployers on the other hand.

Principle 5 of the Responsible AI Framework calls for appropriate ethical and moral principles to underpin standards governing AI system use and deployment. In terms of behavioural standardisation, the principle recognizes that different societal, ethical, and moral considerations may apply locally and need to be accommodated. Thorough testing regimes and co-ordination with international standardization bodies is also recommended. For the full text of this principle, please see page 102.

Overall Approach to Safety and Reliability

The draft EU AIA adopts a multi-tiered, risk-based approach to safety and reliability which distinguishes between prohibited AI, high-risk AI, and other non–high risk AI. We consider this to be very much derived from the product safety approach that the AIA has adopted.

Two types of high-risk AI systems are identified: (a) AI systems covered by European Union harmonisation legislation listed in Annex II to the draft AI Regulation, the so-called New Legislative Framework; and (b) AI systems referred to in Annex III to the draft AI Regulation. On all products that are subject to this New Legislative Framework (e.g., medical devices, machinery, toys, personal protection equipment), the European Commission has thus decided to adjust the existing regulatory regimes by enriching those with AI-flavoured requirements. We consider this to be an elegant response to a complex regulatory problem—it prevents the creation of a new regulatory regime.

In general terms, the approach taken by the AIA leads to the application of different regulatory tools to cope with varying perils. Viewed positively, such a conceptual approach seems convincing as the depth of regulatory intervention depends upon a graduated scale reflecting the potential seriousness of use—in effect, balancing often conflicting interests between ease of use and the maintenance of safety and reliability concerns to determine whether and to what extent an intervention is necessary and appropriate. We do remain concerned, however, that this will lead to a potential inconsistency of approach across the three types of AI systems classified within the EU AIA.

Disparity in Risk Management Requirements

Article 9 of the AIA requires that a risk management system be established, implemented, documented, and maintained in relation to high-risk AI systems. Such a risk management system should consist of a continuous iterative process run throughout the entire lifecycle of the AI system, requiring regular systematic updating.

⚠ **Caution 5.1: We are concerned that the risk management requirements of Article 9 of the AIA only apply on a compulsory basis to systems that are classified as "High Risk" and thus not to low-/minimal-risk AI systems, not even if the AI system (i) interacts with humans, (ii) is used to detect emotions or determine association with (social) categories based on biometric data, (iii) generates or manipulates content, or (iv) is designed for autonomous decision making (and such decision making may negatively impact fundamental rights or ethical standards or principles).**

In addition, the risk management requirements of the AIA only apply to providers and product manufacturers of high-risk AI systems, and not overtly to governments and organizations merely using a high-risk AI system developed by others. (We do take note that there may be implicit coverage beyond this as the definition of "Provider" may already include certain "Users" in some circumstances, more specifically when referring to a party that develops an AI system or that has an AI system developed with a view to "putting it into service under its own name or trademark." The wording "putting into service" is defined as "the supply of an AI system for first use directly to the user or for own use on the Union market for its intended purpose." We would welcome more clarity as to when these conditions might apply.)

We also note a general absence of any measures guaranteeing the safety of AI systems developed "in house." By way of comparison, we would note that the Medical Devices Regulation (MDR) in principle excludes in-house developed medical devices, just as the in-house developed high-risk AI systems intended for in-house use fall partly outside the scope of the AIA. However, the MDR does include conditions for in-house-developed medical devices in order to prevent parties from developing and using in-house medical devices to circumvent the regulation. We would recommend a similar arrangement in the AIA to ensure that AI systems developed and used in house are safe and reliable.

Finally, we note that the AIA in the new draft Presidency Compromise text excludes general purpose AI and AI used for scientific research and development. We would prefer that further classes of AI system are included within the scope of the AIA rather than fewer.

→ **Recommendation 5.1: We would hope for (industry) codes of conduct to be developed (see Article 69) that voluntarily introduce a requirement to establish a risk management system for such AI systems, but optimally we would call for the scope of these obligations to be extended to other AI systems and also to users of such AI systems.**

Security and Cybersecurity Requirements

⚠ **Caution 5.2:** **Security, and specifically cybersecurity, are very important topics in the AIA. However, although these topics are mentioned throughout the AIA, neither security nor cybersecurity are defined in terms of the actual steps which should be taken by Providers. We are concerned by this as embedding cybersecurity (and hence security more widely) into an AI system effectively needs to be done at inception or, to borrow the phraseology of the GDPR, by design—it is not just an ongoing, post-implementation process.**

The draft AI Regulation uses the terms 'security' and 'cybersecurity' without any differentiation. It is not clear why and to what extent this differentiation may lead to different legal consequences. There are some references made to the Regulation (EU) 2016/1148 concerning network and digital services but no requirements as related to AI systems have been put in place based on the said Regulation. However, the said Regulation is focused on networks mainly and not on data. Furthermore, no clear distinction between high-risk systems and "normal" systems is made in the context of any (cyber) security requirements and safeguards (except that Article 15 only relates to high-risk systems, while this also seems to be relevant for "normal" systems).

Looking at other sources for comparative references, companies need to have a clear legal basis in determining what aspects of high-risk AI systems should be secure in security terms: different levels within the lifecycle of AI systems require different types of security requirements. By way of specific example, a high-risk AI system does not only consist of the system per se but also of training data, decisions on the evaluation of test results, partnership, etc.[1] All of these aspects need to be considered when setting up the security requirements for high-risk AI systems (see also ETSI report SAI 005[2]).

→ **Recommendation 5.2:** **The AIA strongly focuses on products and related AI systems, and links requirements to the EU product liability regime. However, AI systems may differ and mainly consist of services requiring different set-ups from a security perspective. These varying demands need to be reflected in the AI regulation, which is not yet apparent from the current draft. Therefore, we would strongly recommend that differentiation of security requirements depending upon the extent to which they are based on hardware, software, services, or product types should be considered and factored into the next draft.**

→ **Recommendation 5.3:** **In relation to security and cybersecurity aspects, the EU Commission should consider including additional remarks specifying the applicable standards to be complied with. Such specification would increase the clarity on respective requirements. In order to ensure a lasting adequacy, either a reference to non-static standards or a regular review and update thereof seems preferable. Specific definitions, requirements, and consequences in case of a breach related specifically to high-risk AI systems seem to be missing. This concerns mainly Articles 6, 8, 9, 10, 15, 16, and 27 of the AIA.**

[1] See, for example, ETSI considerations on AI security https://www.etsi.org/technologies/securing-artificial-intelligence?jjj=1633529023029 (visited on 6 October 2021).

[2] https://www.etsi.org/deliver/etsi_gr/SAI/001_099/005/01.01.01_60/gr_SAI005v010101p.pdf dated March 2021.

Other Obligations Supporting Risk Management

Besides the explicit risk management system obligation in Article 9, the AIA contains various supporting obligations to validate ongoing accurateness, such as:

- the record-keeping obligation for high-risk AI systems under Article 12 (to ensure a level of traceability of the AI system's functioning throughout its lifecycle that is appropriate to the intended purpose of the system);
- the human oversight obligation under Article 14 (aimed at preventing or minimizing the risks to health, safety, or fundamental rights that may emerge when a high-risk AI system is used in accordance with its intended purpose or under conditions of reasonably foreseeable misuse);
- the post-market monitoring obligation under Article 61 (to actively and systematically collect, document, and analyze relevant data on the performance of the high-risk AI system throughout its lifetime and to evaluate continuous compliance of the AI system); and
- the quality management system obligation under Article 17 (to ensure compliance with the AI Regulation, including in relation to aforementioned obligations).

We consider these additional obligations helpful in this respect but note that they are only targeted at providers of high-risk AI systems. Once more, we emphasize the meaningfulness of considering regulating the users of high-risk AI systems. This, of course, does not prevent the user from undertaking compliance on a voluntary basis.

Conformity Assessment Procedure Does Not Underpin Ethical Approach

⚠ **Caution 5.3: Although the AIA (according to Recital 5) supports the EU objective of being a global leader in the development of secure, trustworthy, and ethical artificial intelligence (as stated by the European Council), and ensures the protection of ethical principles (as specifically requested by the European Parliament), the Conformity Assessment procedure lacks a clear obligation for providers and users of (high-risk) AI systems in terms of adhering to ethical principles when designing, developing, putting on the market, putting into service, or using such systems.**

As a result, the conformity assessment procedure as referred to in Article 43 does not provide a direct means to validate the underpinning ethical principles of the AI system assessed. This may allow for high-risk AI systems to meet the requirements of the proposed AI Regulation without safeguarding recognized ethical principles within the EU. Please note, in this respect, that it is not uncommon for EU legislative measures to include clear obligations to adhere to recognized ethical principles. For example, Article 62(3) of the Medical Devices Regulation (MDR) requires clinical investigations of medical devices to be designed and conducted in such a way that the rights, safety, dignity, and well-being of the subjects participating in a clinical investigation are protected and prevail over all other interests. Clinical investigations, furthermore, need to be subject to ethical review, to be performed by an ethics committee in accordance with national law.[3]

→ **Recommendation 5.4: The European Commission should adopt an ethically underpinned approach within the AIA which will align it with the European Parliament Resolution of 20 October 2020 on a Framework of ethical aspects of artificial intelligence, robotics, and related technologies.[4]**

Scope of the Conformity Assessment Procedure Is Too Narrow

We are of the view that the conformity assessment procedure in the AIA is too narrow. Again, we draw analogies in this regard to the MDR—in particular Annex VII of the MDR, Section 4.9 (changes and modifications), which requires the respective notified body to have documented procedures and contractual arrangements with manufacturers of medical devices in place relating to the manufacturers' information obligations and the conformity assessment of (significant) changes.

These changes specified in the MDR are not limited to changes to the device itself (including changes in design, type, or incorporated/utilized substances), but also includes changes in circumstantial facts, such as changes to the associated quality management system (QMS), or changes to the intended use of or claims made for the device by the respective device manufacturer.

Under Article 43(4) of the AIA, however, a high-risk AI system only needs to undergo a conformity re-assessment whenever the AI system itself is substantially modified. A conformity re-assessment is not required when the AI system itself is not modified, but nevertheless the associated QMS (as required under Article 17 of the proposed AI Regulation) is changed.

A (significant) change to the QMS may lead to a situation where regulatory compliance of the AI system can no longer be ensured. A similar situation occurs when the AI system itself is not modified, but the underpinning human-defined objectives (being the intended use of the AI system developed) are changed. Again, Article 43 of the AIA does not require a conformity re-assessment in such situations.

We question why the European Commission has adopted a more limited approach in the AIA in comparison to equivalent procedures in the MDR.

→ **Recommendation 5.5: We recommend that the conformity re-assessment procedure in Article 43 is reassessed within the AIA. Currently, this procedure is only necessary when the AI system itself is modified (as stated in Article 43.4). We feel that the process would benefit from being widened to also apply when relevant circumstantial facts (such as the underpinning human-defined objectives of the system) change. If so, this would provide an opportunity to**

[3] A further clear obligation to adhere to recognized ethical principles when conducting a clinical investigation of a medical device is set out in Chapter I, Article 1 of Annex XV to the MDR.

[4] 2020/2012(INL).

raise subsequent questions as to whether such relevant changed circumstantial facts should also include an assessment of the extent to which the AI system conforms to (and is underpinned by) recognized ethical standards.

Documentation Requirements: Responsibility Rests with the Providers and Manufacturers, Not Users

We note that the documentation requirements under the AIA only applies to 'providers' and 'product manufacturers' of high-risk AI systems, in other words: the governments and organizations developing or making available the high-risk AI systems (Article 16 for providers and Article 24 for product manufacturers). Importers and distributors of high-risk AI systems only have to verify that providers have complied with this documentation requirement (Articles 26 and 27).

→ **Recommendation 5.6: There is a need for users of high-risk AI systems to define the ethical principles underpinning the high-risk AI system they are using. We encourage the EU to reassess this aspect and also to introduce an obligation for governments and organizations using high-risk AI systems to document the ethical principles underpinning the high-risk AI systems they use.**

→ **Recommendation 5.7: We call for the scope of the documentation obligation for high-risk AI systems under Article 11 to be extended to any AI system (high risk or not, including "general purpose" systems and those used for scientific research and development) that (i) interacts with humans, (ii) is used to detect emotions or determine association with (social) categories based on biometric data, (iii) generates or manipulates content, or (iv) is designed for autonomous decision making.**

→ **Recommendation 5.8: For any documentation obligations concerning underpinning ethical standards or principles, the AIA does not make a distinction between AI systems that are designed to autonomously make decisions affecting humans and AI systems that do not. Consequently, we further recommend that this obligation be extended to AI systems (high risk or not) that are designed for autonomous decision making, especially where such decision making may negatively impact fundamental rights and/or ethical standards or principles.**

Enhanced Flexibility Required for Local Ethical Flavours

The AIA lays down harmonised rules on artificial intelligence. The (harmonised) requirements of Title III, Chapter 2, are based on the guidelines formulated by the EU High Level Expert Group (HLEG) ethics guidelines, which—by their very nature—are EU-wide ethics guidelines.

We would therefore question whether there is any flexibility to better reflect local ethical priorities, e.g., on gender neutrality or religious aspects, particularly in cases where these local ethical priorities may provide more protection than the EU-wide ethics guidelines.

Notwithstanding the above comment, we note that Article 10.2 of the AIA allows the use of datasets which may reflect certain geographical, religious, or social considerations and traditions. Furthermore, the definition of AI systems under the draft AI Regulation (Article 3.1) explicitly refers to "human-defined objectives." This may allow for some room to maneuver in respect of applying local ethical standards within the framework of the underpinning European harmonised ethical standards.

Identification of Known and Foreseeable Risks

In relation to the Identification of known and foreseeable risks, also under conditions of reasonably foreseeable misuse for high-risk AI systems, the draft AI Regulation prescribes the establishment of a risk management system by the provider (Article 9). Such a risk management system comprises the "identification and analysis of the known and foreseeable risks" associated with the respective high-risk AI system. We would consider this to be a fair approach in that you can only pre-decide possible occurrences that are known or foreseeable. However, risks do not generally arise solely in possible occurrences where the AI system is used in accordance with its intended purpose, but also under conditions of reasonably foreseeable misuse. If new risks become known or foreseeable post-market, these should also be considered.

Testing and Logging Are Also Needed to Identify and Avoid Situations of Reasonably Foreseeable Misuse

We have mentioned that any risk management system for high-risk AI systems under Article 9 should comprise not only the estimation and evaluation of the risks that may emerge when the high-risk AI system is used in accordance with its intended purpose, but also under conditions of reasonably foreseeable misuse.

Our view is that the testing procedures under the AIA only seem to be limited to those that are suitable to validate the intended purpose of the relevant AI system and not to avoid reasonably foreseeable misuse (Articles 9.6 and 9.7). A similar approach is adopted in Article 12.2, where logging capabilities in relation to record keeping only need to ensure a level of traceability of the AI system's functioning that is appropriate to the intended purpose of the system, and not what is appropriate to log reasonably foreseeable misuse and undertake mitigative and corrective action.

→ **Recommendation 5.9: We would encourage the European Commission to broaden the scope of the testing and logging obligations under Articles 9 and 12 to include reasonably foreseeable misuse.**

Error-free Data (100% vs. as Much as Practicable)

We remain concerned that the requirements of accuracy in relation to datasets specified in Article 10.3 of the AIA do not reflect real-world dataset demographics. The practical consideration here is to ensure that sufficient steps have been taken to ensure that such data are as free from inaccuracies as can realistically and practicably be achieved, given the circumstances and the use cases proposed for the AI system concerned. This is especially important when given the wider context of the proposed penalties regime under the AIA which reserves its highest class of fines for non-compliance with its Data and Data Governance provisions.

⚠ **Caution 5.4: As for data used in high-risk AI systems, the AIA currently requires all training, validation, and testing data to be relevant, representative, free of errors, and complete (Article 10.3). We would question whether it is possible to ensure a "letter-perfect" standard in this regard. Instead we recommend that AI systems should be free of errors "so far as is practicable under the given circumstances and taking into account the intended purpose of the AI system."**

Our assumption is that any risks in this respect will be addressed as part of the risk management system as set out in Article 9.

Bias Detection: Use of Special Category Data only When Strictly Necessary (but for What Purpose?)

Data can never be wholly bias free, but to more adequately identify bias in data requires the processing of special categories of personal data, which is highly sanctioned under the GDPR. We therefore commend the European Commission's decision to explicitly address this regulatory challenge in Article 10.5, and to allow for such processing, within appropriate safeguards.

Under the current wording of Article 10.5, this is restricted to situations where such processing of special categories of personal data *"is strictly necessary for the purposes of ensuring bias monitoring, detection and correction in relation to the high-risk AI system."* We do, however, feel that this qualification is very broad, as such processing will always be necessary to ensure such matters as bias monitoring.

→ **Recommendation 5.10: Further specific limitations must be added to safeguard the processing of special categories of personal data, which should only be permitted to the extent strictly necessary and on a proportionate basis for the intended purpose and outcome of a high-risk AI system.**

Industry Self-regulation

Self-regulation and standard market practices are promoted under the AIA. We refer to Article 69, according to which the European Commission and the EU member states shall encourage and facilitate the drawing up of codes of conduct intended to foster the voluntary application to AI systems other than high-risk AI systems of the requirements set out in Title III of the AIA.

→ **Recommendation 5.11: Self-regulation should be extended further to cover codes of conduct related to environmental sustainability, accessibility for persons with a disability, stakeholder participation in design and development, and diversity (all topics that are not covered in the current draft of the AIA). We are of the view that it does not matter whether codes of conduct are drawn up by individual providers, by representative organizations, or by both—we would like to see this flexibility promoted across the industry.**

Sandboxing

Finally, we note that the AIA promotes the concept of AI regulatory sandboxes, which are encouraged to provide a controlled environment that facilitates the development, testing, and validation of innovative AI systems for a limited time before their placement on the market or putting into service (Article 53). The AIA additionally includes sandboxing conditions on the further processing of personal data for developing certain AI systems in the public interest (Article 54). We view this as is promising, as GDPR compliance often conflicts with innovative AI systems development.

Nevertheless, we would advocate a restrictive approach to this type of testing and evaluation mechanism where concessions are not prolonged indefinitely. Ideally, we would like to see full GDPR compliance being enforced in such systems after the temporary sandbox period—either when the AI system is placed on the market or put into service, also in relation to secondary use of personal data that may have been allowed during the sandbox period. We remain convinced that the best approach towards success often proves to be one where regulatory compliance is implemented 'by design' and becomes part of the provider's or user's DNA, from the outset.

Self-investigative Action by Providers

The AIA introduces several implementing rules to ensure comprehensive and transparent investigation of adverse and unanticipated outcomes of AI systems that may occur through their usage. The rules start with measures that providers and users of AI systems can or need to implement themselves. High-risk AI systems, for example, need to be designed and developed with capabilities enabling the automatic recording of events while the system is operating (logging); this logging should ensure a level of traceability of the system's functioning. To minimise risks to health, safety, or fundamental rights, high-risk AI systems need to be designed and developed in such a way that human oversight is possible (Article 14); this may involve a 'kill switch' in relation to the operation of the AI system where appropriate. Finally, providers of high-risk AI systems need to establish and document a post-market monitoring system to actively and systematically collect, document, and analyse relevant data provided by users or collected through other sources on the performance of the system (Article 61).

Market Surveillance: Adjusting the Regulatory Framework for Evaluation and Review

To aid a comprehensive and transparent investigation of adverse and unanticipated outcomes of AI systems that have occurred through their usage, providers of high-risk AI systems need to report any serious incident (e.g., incidents having lethal or injurious consequences) or any malfunctioning of those systems to the competent market surveillance authorities (Article 62). This is without prejudice to the operation of the competent market surveillance authority's own investigative powers. If a competent market surveillance authority finds that although an AI system is in compliance with the AI Regulation, it still presents a risk to the health, safety, or fundamental rights of persons, it may even demand a withdrawal or recall from the market (Article 67).

Unfortunately, we can find no direct link to any investigative action on adverse and unanticipated outcomes of types and classifications of particular AI systems and subsequent review-based changes to the regulatory framework.

→ **Recommendation 5.12:** In terms of other specific issues within the regulatory regime proposed by the AIA, we observe that high-risk AI systems are required to produce log files which are subject to specific rules. The fact that these log files can simply be requested to be handed over to regulatory authorities in the absence of specific justification does not seem to present to us sufficient legal grounds and does not appear to be transparent. Whilst the connection between a system failing in compliance terms and disclosure of a relevant log file is implicit, we would comment that there needs to be an appropriate legal mechanism to be put in place for authorities to request log files for a limited list of legal reasons. Furthermore, we are of the view that the duty to store log files needs to be limited in time and related to certain incidents.

→ **Recommendation 5.13:** In terms of the power by regulatory authorities to disconnect or shut down high-risk AI systems, we are of the view that this needs to be more specific and more precise to enable it to sit comfortably within the balanced systems of rights and protections envisaged by European law. As it stands, the right very simply expressed could create a wide range of potential further issues if exercised (including matters such as liability, data protection issues, violations of other AIA provisions, and potential removal from EU citizens of valuable technological infrastructure supporting critical services). In our view, consideration needs to be given (particularly in circumstances where the affected high-risk systems may be depended upon by large numbers of people due to their criticality) that alternative methods of enforcement are considered, such as enforcing alternative non-AI technological methods to achieve the same goals.

International Coordination and Standardization

International coordination and standardization are promoted by the AIA. This includes appropriate coordination and cooperation between notified bodies active in the conformity assessment procedures of AI systems pursuant to the draft AI Regulation (Article 38), the possibility to authorize conformity assessment bodies established under the law of a third country with which the European Union has concluded an agreement, such to carry out the activities of notified bodies under the draft AI Regulation (Article 39), but also the coordination and cooperation between EU member states in the context of AI regulatory sandboxes (Article 53).

It is to be expected that the Commission, when adopting harmonised standards or common specifications under Title III, Chapter 5 of the draft AI Regulation, will also try to adhere to international standards for the development and deployment of safe and reliable AI systems as much as possible and practicable. We would only comment that the degree of dependence upon third-party standards providers needs to be limited on the basis that such bodies do not themselves exhibit the characteristics of democratic institutions and have little accountability per se. This, of course, should always be balanced against the access to industry expertise that standards making bodies can attract, and that although many standards making organizations are located internationally, the significant numbers of European talent (amongst that of other continents) is often represented, enabling such institutions to obtain input from a potentially more diverse and inclusive range of experts and organisations. Whilst short term dependence on such providers is understood and acknowledged, we would hope and expect the EU to assess such internationally developed standards and adapt and adopt them, or in the longer term develop its own standards based on fundamental European values and democratically accountable strengths where existing standards fall short in this regard.

Principle 6
Open Data and Fair Competition

Organisations that develop, make available or use AI systems and any national laws that regulate such use shall, without prejudice to normal rules of intellectual property and privacy:

(a) foster open access to, and the portability of, datasets (where privately held), especially where such datasets are deemed significant and important or advance the "state of the art" in the development of AI systems;

(b) ensure that data held by public sector bodies are, in so far as is reasonably practicable, portable, accessible and open; and

(c) encourage open source frameworks and software for AI systems which could similarly be regarded as significant and important and advance the "state of the art."

AI systems must be developed and made available on a "compliance by design" basis in relation to competition/antitrust law.

Principle 6 of the Responsible AI Framework calls for open access to and portability of datasets, especially in cases where these are likely to promote the state of the art in AI system development. The Principle mandates that public sector authorities should lead by example in this regard and encourage open data and portability. These guidelines are made subject to the application of intellectual property law and competition/anti-trust laws. For the full text of this principle, please see page 104.

Overall Approach to Open Data and Fair Competition in the AIA

We would not categorise open data and fair competition as significant issues addressed by the AIA. Indeed, the issues at the heart of our Principle 6 lie on the outskirts of its scope; accordingly, very few of the AIA's provisions are directly related to open data and fair competition. Accordingly, our treatment of these concepts is necessarily limited in this Green Paper—although we would note that there are two related draft legislative measures being pursued by the EU in relation to Data where more comprehensive positions on open data and fair competition appear to reside, both driven by the EU Strategy for Data.

The European Strategy for Data focuses on how the EU can take a leading role in the data economy by creating a single European data space that provides easy access to high-quality data, thereby boosting growth and creating value. In addition to the publication of the White Paper, the European Commission also published a high-level document that previewed some of the initiatives and policies the European Commission intended to put in place. This included the creation of EU data spaces and an evaluation of the fitness of EU competition rules for the digital age.

On 1 October 2021, EU Member States agreed on a common position with respect to the proposal for an EU Data Governance Act (DGA) which aims to allow reuse, and sharing of reuse, of public sector data for commercial and non-commercial purposes by setting out conditions for the use of such data. The DGA refers to data as any digital representation of acts, facts, or information and any compilation of such acts, facts, or information, including in the form of sound, visual, or audiovisual recording. This definition covers both personal data under the GDPR and non-personal data.

In addition, the proposed EU Data Act and amended rules on the legal protection of databases (https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13045-Data-Act-&-amended-rules-on-the-legal-protection-of-databases_en) look at governments' access to business data for "public interest purposes," i.e., public bodies obtaining data from businesses for planning purposes, traffic analysis, or public health.

This initiative will aim to increase access to and further use of data, so that more public and private actors can benefit from techniques such as Big Data and machine learning. The conditions of access and further usage in B2B relationships are often regulated by private contracts. The initiative would look both at data usage rights in industrial value chains and particularly at a fair distribution of usage rights that allow all parties to benefit from data-driven innovation. It would also seek to ensure positive effects for the use of data in the public interest. In short, it is about ensuring fairness in the allocation of economic value among actors of the data economy.

Provisions Addressing Open Data and Fair Competition

Within the AIA itself, the provisions addressing open data and fair competition are limited to Articles 1 a and b, 3 (26), 6, 10, and 63. To a certain degree Article 53 "AI regulatory sandboxes" may also address open data.

Although open data and fair competition are only addressed in the AIA to a limited extent, Article 1 states that prohibitions of certain artificial intelligence practices (lit. a) and rules on market monitoring and surveillance are part of the subject matter of the AIA. These issues are then only addressed in the few articles mentioned above.

The AIA distinguishes between high-risk AI systems and other AI systems. An AI system is to be considered high risk if it creates a high risk to the health and safety or fundamental human rights of natural persons. This classification depends on the intended purpose of the AI system. It will be classified as high risk if it is intended to be used as a safety component of a product, or is itself a product covered by the legislation listed in Annex II of the AIA, and the product whose safety component is the AI system itself as a product, is required to undergo a third-party conformity assessment with a view to the placing on the market or the putting into service of that product according to the harmonised legislation listed in Annex II.

Furthermore, all AI systems listed in Annex III of the AIA are considered high risk. This creates a fairly broad scope of high-risk AI systems that have to comply with the proposed AIA safety or quality rules. For example, Article 10 sets out quality criteria for the training, validation, and testing data of high-risk AI systems. It requires that appropriate data governance and management practices must be implemented. In particular, these data governance and management practices have to relate to the relevant design choices; data collection; relevant data preparation processing operations; prior assessment of the availability, quantity, and suitability of the datasets that are needed; examination because of possible biases; the identification of any possible data gaps or shortcomings; and how those gaps and shortcomings can be addressed. The datasets also have to be relevant, representative, free of errors, and complete. Furthermore, the datasets have to take into account the characteristics or elements that are particular to the specific geographical, behavioural, or functional setting within which the high-risk AI system is intended to be used if necessary due to the intended purpose.

The AIA also provides that the market surveillance control of AI systems will be conducted by the market surveillance authority (Article 63), which means the national authority carrying out the activities and taking the measures according to Regulation (EU) 2019/1020 on market surveillance and compliance of products. The national supervisory authorities will have to report to the EU Commission regularly on the outcomes of their relevant market surveillance activities. In addition, the national supervisory authorities have to report to the Commission any information identified in the course of market surveillance activities that may be of potential interest for the application of EU law on competition rules. This being said, the AIA provides that the competent authority for the EU institutions will be the European Data Protection Supervisor, thereby strengthening the link between competition and data protection within the EU institutions.

Competition/Anti-trust Law Considerations

It is clear that the AIA, in particular the prohibitions and limitations concerning high-risk AI systems, has much in common with other regulatory initiatives which balance the positive effects of free markets against public interests (safety fundamental rights, data protection, consumer protection, etc.) similar to the Digital Services Act proposal to limit harmful content on the internet.

Protection of EU values and fundamental rights are, in other words, balanced against commercial freedom. It may be argued that the distinction between high-risk AI systems and other AI systems means that non-high risk AI systems become more competitive absent the competition from high-risk AI systems, since high-risk AI systems are prevented (or impeded as they will have to comply with stricter rules). This will create a level playing field and leave more room for non–high risk AI systems to compete without a disproportionate regulatory burden.

Market monitoring and surveillance is addressed in the AIA, but apart from a reference to traditional competition law and the obligation to inform competition authorities, there are no specific measures included that specifically address fair competition.

Market Monitoring and Surveillance

The market monitoring and surveillance provisions (Article 63) include a new regulatory body and new practices for testing, monitoring, and compliance processes. The market monitoring mechanisms are focused on the protection of fundamental rights. However, the market distortions caused by the AIA are not addressed directly, other than a general remark that the proposal is without prejudice to the application of Union competition law (Explanatory Memorandum 1.2). The regulatory powers of the AI regulatory authorities are limited to surveillance activities designated under the specific acts in Annex II section A (mainly product safety). There is no further guidance as to how the national regulatory authorities should take any effects on competition into account. The reporting obligation designates a one-way information duty, rather than a consultation procedure.

Recommendation 6.1: **We would recommend that the market monitoring and surveillance provisions of Article 63 are broadened to include a consultation procedure. This would be preferable since it would make competition law assessments more integrated with AI regulatory sector specific intervention, similar to the regulation in the electronic communication services sector.**

Sandboxes

The intention of the sandbox environment (Article 53) is to foster innovation and accelerate access to markets. The sandbox environment is a welcome approach to ensure regulatory testing and expedient removal of existing barriers in Member States as well as establishing a legal basis for using personal data required under GDPR.

Recommendation 6.2: **We would suggest that the sandbox environment be expanded to include open data access and thereby serve as an enabler to ensure fair competition and lift smaller players up by way of also offering access to data (e.g., openly available training sets).**

Open Data

Given the necessity for quality datasets that comply with the data governance and management practices, the need for open datasets becomes increasingly apparent to ensure fair competition. Access to data is key for AI innovation and will also affect the competition between AI systems. Since the Open Data Directive (2019/1024) addresses access to data, it is perhaps not surprising that the substantive provisions in the AIA do not cover access to data. Access to data as a prerequisite for developing high-risk AI systems is briefly commented in recital 45, but only by referring to public initiatives aiming to procure accessible high-quality data for the training, validation, and testing of AI systems.

Recommendation 6.3: **We consider that it would be preferable to see some alignment in the AIA with the principles in the Open Data Directive and also the forthcoming DGA and EU Data Act, including data altruism. Consistent with this, there should be recognition and alignment with new business models proposed by the DGA, such as data intermediation services. For example, Principle 6 of the Responsible AI Framework states that organisations that develop,**

deploy, or use data-driven systems and any national laws that regulate such use shall promote open source and decentralised frameworks. This means that an AI regulation should assess how the use of AI tech solutions and their outputs can be used in other situations or by other organisations, and private organisations should be encouraged or fostered through open access and portability of datasets. Public sector bodies in particular should be required or at least encouraged to ensure that data held by them and used within their AI systems are portable, accessible, and open if reasonably practicable.

Fair Competition

→ **Recommendation 6.4:** Notwithstanding the forthcoming DGA and EU Data Acts, we also are of the view that the topic of open data and fair competition should be expanded in the AIA (to again align it with a consistent EU approach). This principle is more than a policy consideration and needs to be embedded in the regulatory approach as well.

→ **Recommendation 6.5:** The principle of fairness is embedded in the AIA. That principle of fairness focuses on equality and non-biased discrimination, and does not seem to cover fair competition. There is little reference in the proposed Regulation to regulatory tools that can facilitate fair competition (i.e., access to data) and no mention of the negative effects AI systems may have on competition, such as algorithmic collusion. We would like to see these aspects covered in significantly more detail in updated drafts of the AIA (or anti-trust legislation that may accompany it).

Principle 7
Privacy

Organisations that develop, deploy or use AI systems and any national laws that regulate such use shall endeavour to ensure that AI systems are compliant with privacy norms and regulation, taking into account the unique characteristics of AI systems, and the evolution of standards on privacy.

Principle 7 of the Responsible AI Framework calls for an applied integration of privacy laws taking into account the unique characteristics of AI systems—navigating the inherent tensions between the use of AI and respect for private personal data. Accordingly, developers of such systems should have regard to the actual need for personal data given the context of their applications and use cases. For the full text of this principle, please see page 106.

Overall Approach to Privacy

In terms of the general approach, the AIA leaves the GDPR unaffected and complements it with harmonised rules on the design, development, and use of certain high-risk AI systems, as well as restrictions on certain applications of remote biometric identification systems. In this context, this means that the AIA and GDPR need to be read together to understand the full scope of privacy-related obligations that apply to the stakeholders of an AI system.

Role Allocation

Although under the AIA “distributors” and “users” of high-risk AI systems do have some obligations in respect of the AI systems they distribute and use respectively, the operational safeguards in respect of AI systems are only imposed on “providers” under Article 16.

Under the AIA, a “provider” is defined as “a natural or legal person, public authority, agency, or other body that develops an AI system or that has an AI system developed with a view to placing it on the market or putting it into service under its own name or trademark, whether for payment or free of charge.” Further, Article 28 broadens the definition of “providers” by stating that “any distributor, importer, user or other party shall be deemed a provider for the purposes of Article 16 where: (i) they place on the market or put into service a high-risk AI system under their name or trademark; (ii) they modify the intended purpose of a high-risk AI system already placed on the market or put into service; or (iii) they make a substantial modification to the high-risk AI system.”

This gets us so far; but how are these obligations to be shared in the likely multi-party environment of AI development and use? The definition of "provider" set out in the AIA does not capture the potential complexity of how an AI system might be "provided" and the different roles.

It would be extremely unusual for a single company to create an AI solution alone. The base technology will need configuration and then the end use will develop further; vendors increasingly must form strategic partnerships that give them access to all the necessary complementary technologies and data. In short, the authorship of AI systems is often multi-party in addition to the end user often being a further separate entity.

As such, any attempt to regulate AI systems needs to apply across all those organisations that develop, deploy, and use the AI systems to ensure appropriate operational safeguards are applied to protect privacy of data subjects across the lifecycle of an AI system, and across the stakeholder constellation.

⚠ **Caution 7.1: The AIA does not appear to provide for situations where there are multiple stakeholders involved in providing a high-risk AI system, for example, an IT company provides the engine on which it is built, a data analytics company overlays its expertise, and another company feeds in the data required to teach an AI system. It is not clear which party will be the "provider" under the AIA in this type of arrangement, so which party is required to comply with the requirements under Article 16?**[1]

In order to apply regulatory protection effectively, each of these parties should be subject to requirements to employ appropriate operational safeguards. However, if this is how "provider" is to be interpreted (as including all parties involved in the development of an AI system) it is not clear how the responsibility and liability for such development should be split between the different parties.

→ **Recommendation 7.1: The AIA needs to better address the practical realities of how an AI system may be developed, so it is clearer as to which stakeholder the relevant obligations should apply to and how. Alternatively, it could take a more outcomes-based approach, so that any party that engages with an AI system has an obligation to ensure appropriate operational safeguards are put in place to mitigate against adverse outcomes.**[2]

Anonymisation (vs. Pseudonymisation)

Although the AIA regulates AI from a more risk-based perspective, it does not even consider anonymisation or pseudonymisation as methods for AI systems that exclude or mitigate data protection concerns. It is only in the context of processing special categories of personal data (see Article 9.1 of the GDPR) that the AIA mentions privacy-preserving techniques such as pseudonymisation and encryption of data as adequate measures to protect data subjects' rights (see Article 10.5 of the AIA).

[1] In this regard, we would draw attention to the useful recommendations provided by UNESCO under its draft text on the ethics of artificial intelligence, dated July 2021.

[2] See also the Guidance on AI and Data Protection produced by the UK's Information Commissioner, which also recognises a multi-actor environment in the context of AI development and deployment.

Assuming that artificial intelligence is to be capable of learning, it requires a great deal of data with which to train, test, and evaluate it. If the data is personal, its use for these purposes creates data protection problems. Anonymisation could be a solution for many AI systems, even if it must be considered that in many jurisdictions the method of anonymisation is actually treated as data processing and has to comply with data protection rules.

The massive amounts of data, the numerous possibilities to analyse and combine data, as well as several parties being involved with only one technology has the potential to threaten the rights of individuals. AI has the distinct and clear potential to violate personal data, and anonymisation is one way to prevent breaches of data protection rights. However, many AI systems cannot be trained with only anonymous data, and need a certain reference to an individual in order to be able to be improved and better trained.

Another possibility for prevention of privacy infringements is pseudonymisation, under which method the data subject remains "individualisable" but not identifiable.

⚠ **Caution 7.2: Whilst the AIA sets rules for high-risk AI systems that use data for training (see Article 10), it fails to create a set of rules for the use of anonymous data, and even more so to create a clear distinction between pseudonymous and anonymous data. Our concern is that a significant proportion of AI models (not just high-risk AI systems) are trained with personal data often derived from other processing activities. Even if in this context the data are pseudonymised, the framework of Articles 5 and 6 of the GDPR will apply until such data are fully anonymous. This creates several legal questions because further processing of personal data in order to train, test, and evaluate AI systems will not be compliant in many situations. Therefore, anonymisation before change of purpose, or pseudonymisation thereafter with the consequence of exemptions from the GDPR, is a vital interest for providers of AI systems.**

⚠ **Caution 7.3: The AIA does not address the risk of certain AI systems (particularly General Adversarial Networks) to be used as tools to unpick privacy preserving techniques or to game a targeted AI system to create further vulnerability.**

→ **Recommendation 7.2: We would recommend that a supportive measure within the AIA might be to prescribe certain "compliant" methods of anonymisation (for example, that it should be done by a third-party processor). Subsequent provisions could clearly prohibit de-anonymisation and provide for penalties in the event of an infringement.**

We note that pseudonymised data are, within the regime of the GDPR, more or less treated identically to direct personal data. However, if implemented correctly, the risk for individuals could be significantly reduced. We would suggest that the use of pseudonymised data in an AI context (and properly managed) could be exempted from the GDPR (e.g., from fully fledged information or conducting a Data Protection Impact Assessment (DPIA)).

This would require, however, that general rules for pseudonymisation are given to AI providers in order to minimise risks for data subjects. Such a set of rules should at least cover the following parameters:

- details of a pseudonymisation procedure (especially which data may be used and which direct/indirect identification features must be removed)
- under which circumstances the addition of and linking with further pseudonymised data from third-party sources should be permissible
- how the pseudonymised data may be handled (for example, in the case of disclosure)
- the requirement to retain the additional information and the measures to regulate access to that information
- the conditions under which re-identification may or must take place and by whom, as well as the distinction from anonymous data

Lawful Basis and Consent

The AIA specifies that high-risk AI systems which make use of techniques involving the training of models with data are to be developed on the basis of training, validation, and testing datasets that meet the quality criteria referred to in Article 10.2 to 10.5. Article 10.2 also provides that training, validation, and testing datasets must be subject to appropriate data governance and management practices. Such data governance must also include the lawful basis for processing of personal data.

When using personal data to train AI systems, we would typically see that applicable data is collected by an AI user from customers, employees, or other data subjects. Such an AI user can lawfully process data harvested in this manner for the fulfilment of a particular contract or base their processing on the direct consent of the data subject. In a similar fashion, we would typically see the AI provider (in this case a sub-processor of the AI user) being permitted to use such data under the GDPR so long as it is for the same purpose (and indeed subject to the same legal basis). However, in what is a relatively common problem in the AI industry, AI providers typically use such data to train and develop AI models beyond the express purposes of the AI user, and again, typically rely on legitimate interests as a processing ground in such circumstances.

⚠ **Caution 7.4: There must be additional safeguards in place beyond the mere application of data protection principles to prohibit AI system training when the balance between the fundamental rights of the data subject and the legitimate interests of the AI provider is not met, especially with high-risk AI systems. We are of the view that specific quality criterion have been omitted from the AIA in this regard—that the processed data must be, before being utilised for AI system training, in the overriding interest of the data subject and/or the AI user. In no circumstances should training be undertaken when the overriding interest is not with the AI user.**

Purpose Limitation

The classification of an AI system under the AIA is based on its intended purpose. Different obligations derive from this classification. Changes of purpose change could lead to either additional or fewer (different) compliance issues. Under the GDPR, further processing beyond the initial purpose specified without a lawful basis is not allowed unless Article 89 GDPR might apply, which is especially critical under AI systems.

Article 89 of the GDPR provides an exception to further processing for archiving purposes in the public interest, scientific or historical research purposes, or statistical purposes, when specific guarantees are taken.

→ **Recommendation 7.3: We are concerned that there is an open question as to which processing activities can be subsumed under the term "scientific and statistical purposes," especially as AI could include statistical approaches (see Annex I of the AIA). Interpreted widely, this could conceivably permit any AI processing for further purposes which we argue would be an unintended consequence. We would therefore urgently recommend a clarification on the status of AI systems vis-à-vis Article 89 of the GDPR.**

Profiling and Automated Decision Making

⚠ **Caution 7.5: Profiling and automated decision making are activities typically reliant on AI and involve personal data. The extent to which and with what consequences the GDPR applies to AI systems/machine learning in this context is not specifically addressed by the AIA.**

Our concern is that providers of AI systems might be forced to inform data subjects properly about how the algorithms in AI systems interpret data through a kind of machine perception, labelling and clustering raw input, and what consequences result for the data subject from the output of such machine learning. In our view, it is questionable whether, on the one hand, such explanations would be transparent for the data subject, and on the other hand, whether such declarations would even be possible for the provider in a complete and correct manner.

→ **Recommendation 7.4: We would submit that the goal of any machine learning system is to focus on algorithms that are designed to find patterns in data and use these patterns to make predictions. Hence, in many AI systems (in particular in the field of data mining that deals with the prediction of future developments), profiling is an indispensable component. This is why we consider that it would enhance the AIA in this context to define a framework which allows profiling as part of machine learning and provides for a context-driven explanation of the consequences of this profiling, rather than just a blanket proscription.**

Biometric Identification

The use of 'real-time' biometric identification systems in publicly accessible spaces for purposes other than law enforcement is not prohibited by the AIA.

In light of the current concerns in relation to the use of biometric identification systems, the AIA should provide greater clarity on its interplay with the existing EU data protection legislation including the prohibition in Article 9 of the GDPR, the exemptions in Article 9(2) of the GDPR, and automated decision making (including profiling) in Article 22 of the GDPR. This is an opportunity to bring in complementary rules in the unique context of AI which will reinforce and strengthen the provisions in the GDPR and alignment with the other EU laws.

Recommendation 7.5: For clarity, it would also be beneficial to provide further provisions to regulate private use of remote biometric identification in publicly accessible spaces in a way that is consistent with Article 9 of the GDPR. The AIA should explicitly clarify that existing EU data protection legislation applies to any processing of personal data falling under its scope regarding biometric data, including the GDPR. As the AIA would be a global standards-setting piece of legislation—much like the GDPR—it would assist if any potential ambiguities in this regard were removed.

Principle 8
AI and Intellectual Property

Organisations that develop, deploy or use AI systems should seek to strike a fair balance between benefiting from adequate protection for the intellectual property rights for both the AI system and the AI output and allowing availability for the wider societal benefit. Governments should investigate how AI systems and AIcreated output may be afforded adequate protection whilst also ensuring that the innovation is sufficiently disclosed to promote progress.

Principle 8 of the Responsible AI Framework calls for a balance between protecting the intellectual property rights of rights holders and the potential benefit to society of wider disclosure. Caution should be used in the introduction of new laws which impact Intellectual Property Rights (IPRs), and the interests of all relevant stakeholders should be taken into account in this regard. For the full text of this principle, please see page 108.

Access to Data, Information, and Documentation

The AIA provides specific access rights to national competent authorities and notified bodies to documents and data (*"full access to the training, validation and testing datasets used by the provider"*) relating to high-risk AI systems. Such rights enable such authorities to access also the source code of the AI system *"where necessary to assess the conformity of the high-risk AI system with the requirements set out in Title III, Chapter 2 and upon a reasoned request."*

According to Article 70 AIA, *"national competent authorities and notified bodies involved in the application of this Regulation shall* ***respect the confidentiality*** *of information and data obtained in carrying out their tasks and activities in such a manner as to protect, in particular … intellectual property rights, and confidential business information or trade secrets of a natural or legal person, including source code …."*

The AIA does not specify what appropriate measures should be implemented in order to *"respect the confidentiality"* of data and documents received by the national competent authorities or notified bodies. In particular, the AIA makes no reference whatsoever to the levels of security that these authorities should ensure in the exercise of their powers regarding access to documentation and data.

Directive (EU) 2016/943[1] provides, in this respect, trade secret protection to the extent that the information to be protected are *"subject to reasonable steps under the circumstances, by the person lawfully in control of the information, to keep it secret."*[2]

In this respect, it is noted that the European Commission's Intellectual Property Action Plan 2020[3] announced a clarification of the Trade Secrets Directive together with a review of the Database Directive (Directive 96/9/EC of 11 March 1996). The European Commission's Work Programme 2021 confirmed the review of the Database Directive as part of its broader Data Package. The Work Programme states that action is being taken across these areas with legislation to cover "the safety, liability, fundamental rights and data aspects of artificial intelligence ... to set right conditions for better control and conditions for data sharing for citizens and businesses."[4]

⚠ **Caution 8.1: Intuitively, allowing access to the information specified in Article 70 AIA to a public third party in the absence of specific protection requirements of such information has the potential to create a potential (and unnecessary) point of failure that, if exploited for example by a malicious third-party agent, could undermine the effective protection as trade secrets of the data and information disclosed.**

The issue is relevant not only from the trade secrets perspective: unlawful access to data, information, and documents disclosed to national competent authorities and notified bodies involved in the application of this Regulation could jeopardise the potential patentability of any invention described in such data and documents, therefore destroying the patent novelty requirement. Moreover, also database rights might also be put at risk in absence of specific security safeguards.

Ultimately, the AIA correctly poses the issue of intellectual property protection for developers of AI systems in the event of access to data and documents by competent bodies, but without specifying "how" and "by what means" such an effective protection can be actually obtained.

⚠ **Caution 8.2: In light of its disclosure requirements, there is a risk that the AIA may act to disincentivise and/or dissuade development and/or exploitation of AI systems in the territory of the European Union (EU), thereby depriving EU citizens and/or undertakings based in the EU of any benefit afforded by such AI systems, possibly to the detriment of EU society.**

→ **Recommendation 8.1: The introduction of specific requirements regarding processes for the secure management of information by the national competent authorities and notified bodies involved in the application of the AIA is considered highly appropriate, also by reference to existing international standards and ISMS frameworks (e.g., ISO 27001, NIST, etc.) and audit**

[1] Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use, and disclosure.

[2] Article 2(1)(c) Directive (EU) 2016/943.

[3] European Commission Communication COM(2020) 760 final: Making the most of the EU's innovative potential—An intellectual property action plan to support the EU's recovery and resilience.

[4] Commission Work Programme 2021 (COM(2020) 690 final): 2.2. A Europe fit for the digital age.

best practices, also including a specific requirement regarding time limitations in the storage of or access to data and documents, limited to what is strictly necessary for the performance of their duties under the AIA.

→ **Recommendation 8.2: Standard access and disclosure agreements should be elaborated by the Commission—as has happened, for example, with Standard Contractual Clauses under the GDPR—in order to facilitate standardisation of the access procedures and ensure the effectiveness of confidentiality measures among Member States, and to provide an efficient allocation of responsibilities in case of intentional or non-intentional breach of confidentiality of data and information accessed by the national competent authorities and notified bodies.**

Such an approach would also likely boost *"the voluntary application to AI systems other than high-risk AI systems"* of the AIA requirements, in line with the objectives under Article 69 AIA, dedicated to code of conducts.

The foregoing discussion recognises and suggests steps that may be implemented to improve upon measures taken to respect the confidentiality of data and documents submitted to national competent authorities and notified bodies applying the proposed regulation. Notwithstanding that the AIA could be more specific with respect to the processes for securing the management of information, it is nevertheless welcomed that the value of the data and documents discussed above is recognised and that it is appropriate to keep it confidential. This should be contrasted with proposals under discussion within and between major national and supranational intellectual property offices to require the submission of the training set data for patent applications seeking protection for an AI invention in which training data is to be utilised. It is suggested that such submissions may be made mandatory or may be made voluntary, although it is expected that even in a voluntary system the submission of training data will in effect become mandatory as it will be expected by patent examiners, and its absence will have a negative impact on the outcome of the application process.

A deposit system for training data, similar to the deposit system for microorganisms[5] necessary for the working of an invention the subject of a patent application, has been suggested which will provide access to the training data to any persons who wish such access. However, this would clearly be incompatible with the notion that the data may be kept confidential. It is beyond the scope of this paper to discuss whether or not such training data is necessary to enable AI patent applications or whether issues concerning the requirements for granting a patent may render the proposals for submission of training data unfeasible and even incompatible with the innovation process. However, such a training data deposit system is inconsistent with the need to respect the confidentiality of training data as recognised in the proposed AIA.

→ **Recommendation 8.3: We would like to see the respective bodies responsible for assessing AI innovation and systems consult with each other with respect to the principles they are to implement in performing their duties and to ensure that there is a consistency of principle and approach between them.**

[5] Budapest Treaty—International Microorganism Deposit System.

Striking a Fair Balance between Algorithmic Transparency, AI Explainability, and Trade Secrets Protection

High-risk AI systems are subject to explainability requirements under the AIA. In particular, Article 13.1.1 provides that high-risk AI systems should be *"designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately."* In addition, Article 13.2 states that high-risk AI systems should be *"accompanied by instructions for use in an appropriate digital format or otherwise that include concise, complete, correct and clear information that is relevant, accessible and comprehensible to users."* Furthermore, Recital 47 of the AIA provides that in order *"to address the opacity that may make certain AI systems incomprehensible … a certain degree of transparency should be required for high-risk AI systems. Users should be able to interpret the system output and use it appropriately. High-risk AI systems should therefore be accompanied by relevant documentation and instructions of use and include concise and clear information, including in relation to possible risks to fundamental rights and discrimination, where appropriate."*[6]

The explainability requirements set in the AIA are formulated as general obligation, since the proposal itself is silent on what specific measures should be taken in order to ensure that high-risk AI systems are sufficiently interpretable to the user. Such identification appears to be largely left to a self-assessment of the provider before the AI system is placed on the market or put into service, so that the question of how to achieve effective AI explainability is left to the discretion of AI systems providers.[7]

We provide a fuller analysis of the tension between AI explainability and AI transparency in our commentary under Principle 3.

A strict approach on the explainability issue may, however, lead to the interpretation of the provisions of the AIA as a lock pick to force AI developers to disclose, at least partially, trade secrets in their AI algorithms, which are often natural "black boxes" where even their creators cannot easily explain how they work.

The AIA is not the only piece of legislation dealing with AI transparency and explainability requirements, mainly in connection with algorithmic decisions affecting individuals.

In the first place, a vibrant debate exists on whether a right to explanation is actually encompassed in the GDPR. Those who support the existence of such a right of explanation in the GDPR emphasise the conjunction between Article 22,[8] Article 15 on access rights (according to which the data controller must provide individuals with *"meaningful information about the logic involved"* in automated decision-making),

[6] Commission Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, COM/2021/206 final (2021).

[7] Ebers, Martin, "Regulating Explainable AI in the European Union: An Overview of the Current Legal Framework(s)," (9 August 2021). Liane Colonna/Stanley Greenstein (eds.), *Nordic Yearbook of Law and Informatics 2020: Law in the Era of Artificial Intelligence.* Available at SSRN: https://ssrn.com/abstract=3901732 or http://dx.doi.org/10.2139/ssrn.3901732.

[8] Providing, in connection with automated decisions affecting individuals, that *"the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision."*

and Recital 71, which points out that safeguards *"should include specific information to the data subject"* including *"an explanation of the decision reached after such assessment."*[9]

Explainability and transparency requirements in connection to algorithms are also provided by the EU Digital Service Act (DSA) proposal,[10] which refers to algorithms under the definition of "recommender systems," defined as *"fully or partially automated system used by an online platform to suggest in its online interface specific information to recipients of the service, including as a result of a search initiated by the recipient or otherwise determining the relative order or prominence of information displayed."* The recipients of the service have the right, under Article 29(1) DSA, to know the main parameters of recommender systems, and to have options to influence/modify those parameters, including at least one option which is not based on profiling.[11]

Many authors have, however, pointed out that the explainability of AI/algorithmic decisions is affected by several unresolved issues, including both technical obstacles as well as legal limits that may impair explainability itself.[12] One of the main challenges is that, due to the unavailability of alternative intellectual property protections and their often opaque and inscrutable status, AI algorithms are often protected as trade secrets.

⚠ **Caution 8.3: The explainability requirements of the AIA appear to be built upon the assumption that a technology creator is able to understand how the technology functions and to explain how an algorithm makes decisions based on the dataset parsed by it. These assumptions (particularly in cases where Deep Neural Networks are involved) may in certain circumstances be inaccurate, since such algorithms are not just protected as trade secrets under a legal perspective, but also constitute in most cases technological "black boxes" where the algorithm is secret by definition.[13] The notion of black box AI refers to scenarios in which we can see only input data and output data for algorithm-based systems, without having insight into exactly what happens in between.[14] In this respect, the complexity of the AI system/algorithm leads to a high level of difficulty in providing effective explanations.[15]**

[9] Since the GDPR mentions a right to explanation only in its non-binding recitals, it should be highlighted that some scholars agree that GDPR does not provide for a right to explanations of individual decisions.

[10] Proposal for a Regulation of The European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC, COM/2020/805 final.

[11] Huseinzade, Nazrin, "Algorithm Transparency: How to Eat the Cake and Have It Too," available at https://europeanlawblog.eu/2021/01/27/algorithm-transparency-how-to-eat-the-cake-and-have-it-too.

[12] Brkan, M., & Bonnet, G. (2020). "Legal and Technical Feasibility of the GDPR's Quest for Explanation of Algorithmic Decisions: of Black Boxes, White Boxes and Fata Morganas." *European Journal of Risk Regulation,* 11(1), 18-50. doi:10.1017/err.2020.10.

[13] Tschider, Charlotte, "The Consent Myth: Improving Choice for Patients of the Future," 96 *Wash. Univ. L. Rev.* 1505.

[14] Ebers, Martin, "Regulating Explainable AI in the European Union: An Overview of the Current Legal Framework(s)," (9 August 2021). Liane Colonna/Stanley Greenstein (eds.), *Nordic Yearbook of Law and Informatics 2020: Law in the Era of Artificial Intelligence.* Available at SSRN: https://ssrn.com/abstract=3901732 or http://dx.doi.org/10.2139/ssrn.3901732.

[15] Tschider, Charlotte, Beyond the Black Box (May 1, 2021). "Beyond the Black Box," 98 *Denv. L. Rev.* 683 (2021), available at SSRN: https://ssrn.com/abstract=3855782.

As emphasised among scholars, explaining AI is potentially an extremely complex task since it does not involve solely the algorithm itself[16] but also algorithmic design choices and algorithmic training (since if data and training processes are rendered or left opaque and obscure, being subjected to trade secrecy status means that the overall opacity of the system increases).

The *dynamic inscrutability* of AI algorithms should be taken into due consideration, with particular reference to complex AI systems such as Machine Learning (ML) and Neural Networking AI.

Generally speaking,[17] ML unsupervised learning (which leverages a lack of pre-arranged structure that permits an AI utility to create its own relationships and structures between data elements) might develop highly complex algorithms which may be also dynamic, changing over time based on new data supplied. This means that the algorithm might change the minute after the decision has been made, so it is therefore impossible to define the algorithm's status at the exact moment of the decision. Explanation would therefore only be achievable temporarily before the algorithm changes again. Moreover, neural networks—which are based on unsupervised ML—add additional hundreds or thousands of hidden layers of computation that increase the outputs' accuracy while increasing the obscurity of the underlying logic.[18]

An AI system is, therefore, more than "just" an algorithm, and the quest for AI explainability cannot concentrate on algorithmic transparency: any interpretation of the AIA provision on explainability for high-risk AI systems aimed at "opening" the black boxes (therefore jeopardising trade secrets rights in the algorithm) should be discouraged.

It should be noted, in this respect, that EU national courts are already interpreting disparate law provisions to force organisations into disclosing trade secrets in their algorithms and software source code.[19] In 2015, the French *Commission d'accès aux documents administratifs* obliged the *Direction générale des finances publiques* to release the source code of the computer program used to estimate the income tax of natural persons.[20] In Italy, the TAR (an administrative law court) stated that an algorithm is a digital administrative act and therefore, under the freedom of information regime, the citizens have the right to access it.[21] Also the Italian State Council recently stated that *"in order to allow the full knowledge of the module used and the criteria applied with the algorithm, it is necessary to guarantee a wide transparency, which must invest every aspect of the training and use of the IT medium, so as to guarantee the knowledge*

[16] Bob Violino, "Designing and Building Artificial Intelligence Infrastructure," TechTarget (5 April 2018), https://searchenterpriseai.techtarget.com/feature/Designing-and-building-artificial-intelligence-infrastructure.

[17] Eda Kavlakoglu, "AI vs. Machine Learning vs. Deep Learning vs. Neural Networks: What's the Difference?," IBM Cloud (27 May 2020), https://www.ibm.com/cloud/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks.

[18] Jenna Burrell, "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms," *Big Data & Soc'y* 1, 5–9 (Jan-June 2016), https://doi.org/10.1177/2053951715622512.

[19] Noto La Diega, Guido, "Against the Dehumanisation of Decision-Making: Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection, and Freedom of Information" (31 May 2018). 9 (2018) JIPITEC 3 para 1, available at SSRN: https://ssrn.com/abstract=3188080.

[20] Commission d'accès aux documents administratifs, avis 20144578 – 8 January 2015, https://joinup.ec.europa.eu/collection/egovernment/document/france-opens-source-code-tax-and-benefits-calculators-increase-transparency.

[21] TAR Lazio, chamber III bis, 22 March 2017, no 3769.

of the identity of the its authors, the procedure used for its full knowledge of a rule expressed in a language different from the legal one."

Despite the position of the case law briefly recalled above, trying to "open the black box" (by forcing organisations to give access to the underlying trade secrets, including source code) might be however impossible or at least highly ineffective, since the information (potentially) gathered though this process would be unlikely to provide the kind of information needed to evaluate risks relating to unfairness, discrimination, and safety.[22]

Such an approach would be therefore highly ineffective in protecting individual rights, as well in ensuring AI safety and fairness. Moreover, it would be extremely difficult (or even impossible), and undesirable from an IP perspective, to require organisations to explain their algorithms by giving access to trade secrets underlying and protecting the algorithm itself.

This also collides with one of the most relevant findings under Responsible AI Principle 8 (Intellectual Property), according to which an adequate balance of the interests of all relevant stakeholders should be sought when implementing or amending laws having impacts on intellectual property rights.

→ **Recommendation 8.4: A more effective approach to determining the operation of an algorithm when implementing it in an AI system would be to test the operation of the AI system by utilising a reference input to determine a satisfactory output. In this way, the issue of what goes on in the "black box" is irrelevant providing the output fulfils the necessary criteria when tested against a particular input. This may be considered an outcome-based approach rather than a design-based approach where the design of the algorithm and its configuration is to be transparent and explainable. Such an outcomes-based approach may obviate the need for access to data and documents to assess an AI system provided it behaves within what are deemed to be acceptable parameters.**

Different approaches to the explainability requirements under the AIA should be therefore explored in order to strike a fair balance between such requirements and IP protection of the algorithms themselves. In the balance between IP and trade secrets protection and the right to an explanation of the functioning of algorithms, a fair balance should be sought in order to provide meaningful and useful information to the recipients, without jeopardising the IP rights and trade secrets of the AI system developers/right holders.

While such a balance between explainability and protection of IP rights may be difficult to achieve, it should be made clear that requirements under Article 13 should be without prejudice to trade secrets rights in the algorithm/AI system. In this respect, while many authors proposed different approaches to algorithmic explainability, such as for example utilising counterfactual explanations,[23] the development

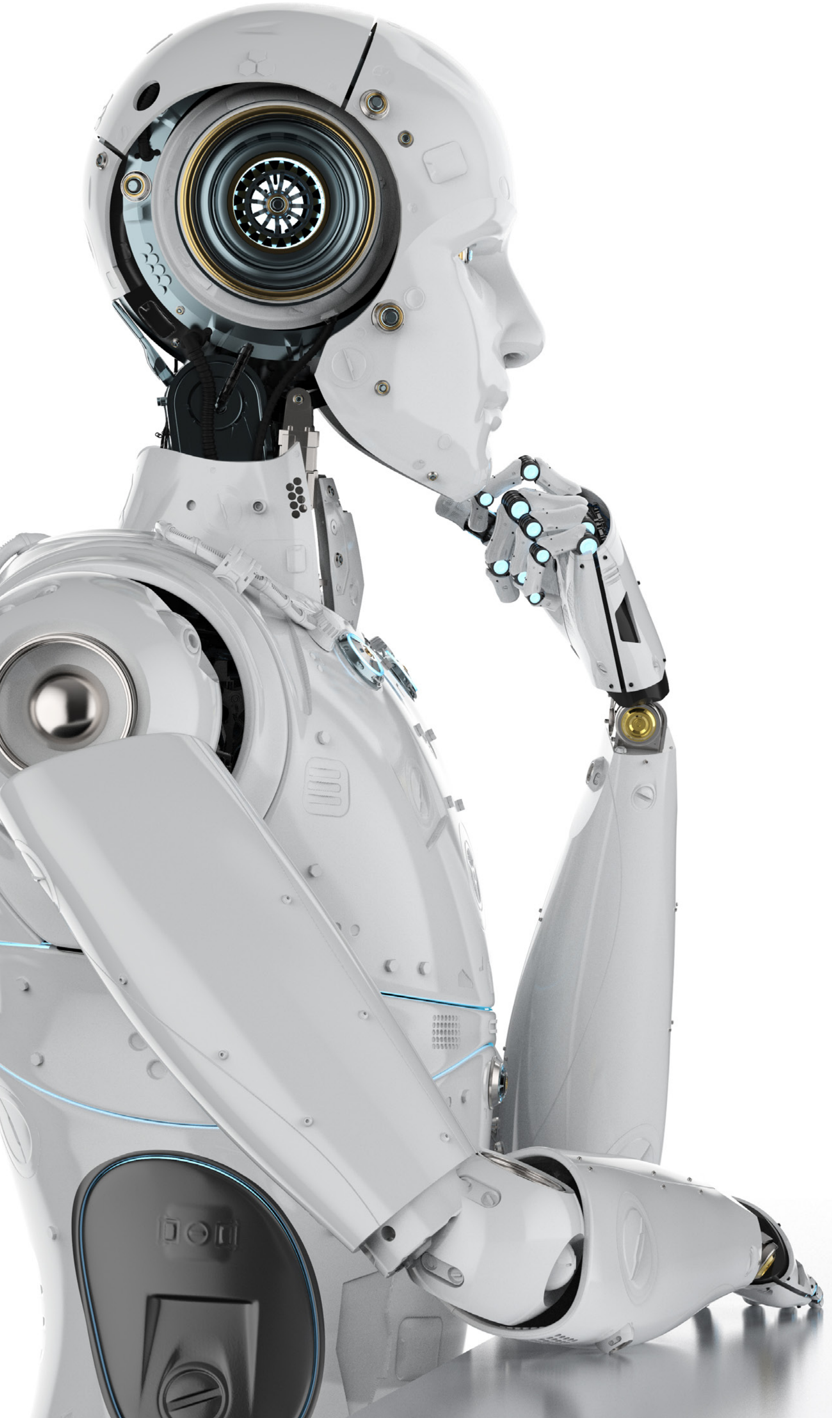
[22] Lilian Edwards & Michael Veale, "Slave to the Algorithm? Why a 'Right to Explanation' Is Probably Not the Remedy You Are Looking For," 16 *Duke L. & Tech. Rev.* 18, 18 (2017).

[23] Wachter, Sandra and Mittelstadt, Brent and Russell, Chris, *Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR* (6 October 2017). Harvard Journal of Law & Technology, 31 (2), 2018, Available at SSRN: https://ssrn.com/abstract=3063289 or http://dx.doi.org/10.2139/ssrn.3063289.

of standardised best practice enabling algorithmic explainability without disclosing the underlying trade secrets is of paramount importance for achieving a fair balance between explainability and IP protection.

PART C

Responsible AI 2021 Policy Framework



Responsible AI

A GLOBAL POLICY FRAMEWORK

Principle 1
Ethical Purpose and Societal Benefit

Organisations that develop, make available or use AI systems and any national laws or industry standards that govern such use should require the purposes of such implementation to be identified and ensure that such purposes are consistent with the overall ethical purposes of beneficence and non-maleficence, as well as the other principles of the Policy Framework for Responsible AI.

1 Overarching principles

1.1 Organisations that develop, make available or use AI systems should do so in a manner compatible with human agency, human autonomy and the respect for fundamental human rights (including freedom from discrimination).

1.2 Organisations that develop, make available or use AI systems should monitor the implementation of such AI systems and act to mitigate against consequences of such AI systems (whether intended or unintended) that are inconsistent with the ethical purposes of beneficence and non-maleficence, as well as the other principles of the Policy Framework for Responsible AI set out in this framework.

1.3 Organisations that develop, make available or use AI systems should assess the social, political and environmental implications of such development, deployment and use in the context of a structured Responsible AI Impact Assessment that assesses risk of harm and, as the case may be, proposes mitigation strategies in relation to such risks.

2 Human Agency and Autonomy

2.1 Organisations that develop, make available or use AI systems that surveil human behavior shall put in place appropriate safeguards to promote the right to be let alone (the right not to be subject to arbitrary interference with his privacy, family, home or correspondence), informed human agency and autonomy and to avoid destructive self-censorship, loss of individuality and identity, loss of freedom of expression and the loss of human ability to think freely and independently. Such safeguards shall include conducting a responsible AI ethical risk assessment of the technology as part of an accountable governance process prior to deployment of the AI System and ensuring that any such deployment is consistent with respect for other principles of the Policy Framework for Responsible AI such as Transparency and Explainability, Fairness and Non-Discrimination, and Privacy

2.2 Organisations that develop, make available or use AI systems that surveil human behavior using sensitive personal data (such as data collected in non-public spaces such as the home), facial-recognition data or biometric data shall apply the Transparency and Privacy principles with particular rigour, including as regards the reasonable purpose, limited collection, limited use, limited disclosure and limited retention principles, as well as by providing full transparency as to whether and when a device's voice, movement or image surveillance features have been activated. Sensitive personal data such as biometric data and genetic data collected locally by IoT devices (such as fitness monitors and smart phones) and natural language, movement and image data collected by "always on" IoT devices (such as personal assistants and smart home devices) shall, to the greatest extent possible, securely store such data, in encrypted format, only locally on the device in a manner that allows for the maximal level of autonomy and control over the data by the individual(s) to whom it relates.

2.3 Organisations that develop, make available or use AI systems that predict and influence human behavior shall put in place appropriate safeguards to promote informed human agency and autonomy and to avoid destructive psychological and behavioural manipulation, addiction, dependency and attention deficit. Such safeguards shall include conducting a responsible AI ethical risk assessment of the technology as part of an accountable governance process prior to deployment of the AI System and ensuring that any such deployment is consistent with respect for other principles of the Policy Framework for Responsible AI such as Transparency and Explainability, Fairness and Non-Discrimination, and Privacy.

3 Work and automation

3.1 Organisations that implement AI systems in the workplace should provide opportunities for affected employees to participate in the decision-making process related to such implementation.

3.2 Consideration should be given as to whether it is achievable from a technological perspective to ensure that all possible occurrences should be pre-decided within an AI system to ensure consistent behaviour. If this is not practicable, organisations developing, deploying or using AI systems should consider at the very least the extent to which they are able to confine the decision outcomes of an AI system to a reasonable, non-aberrant range of responses, taking into account the wider context, the impact of the decision and the moral appropriateness of "weighing the unweighable" such as life vs. life.

3.3 Organisations that develop, make available or use AI systems that have an impact on employment should conduct a Responsible AI Impact Assessment to determine the net effects of such implementation.

3.4 Organisations that develop, make available or use AI systems that surveil or influence employee behavior in the workplace shall put in place appropriate safeguards to promote the informed human agency, autonomy and dignity of employees and to avoid inappropriate or destructive impacts on the emotional or psychological health of employees (monotony of tasks, excessive surveillance, gaming of behavior, continuous exposure to horrific content). Such safeguards shall include conducting a responsible AI ethical risk assessment of the technology as part of an accountable governance process prior to deployment of the AI System and ensuring that any such deployment is consistent with respect for other principles of the Policy Framework for Responsible AI such as Transparency and Explainability, Fairness and Non-Discrimination, and Privacy.

3.5 Governments should closely monitor the progress of AI-driven automation in order to identify the sectors of their economy where human workers are the most affected. Governments should actively solicit and monitor industry, employee and other stakeholder data and commentary regarding the impact of AI systems on the workplace and should develop an open forum for sharing experience and best practices.

3.6 Governments should promote educational policies that equip all children with the skills, knowledge and qualities required by the new economy and that promote life-long learning.

3.7 Governments should encourage the creation of opportunities for adults to learn new useful skills, especially for those displaced by automation.

3.8 Governments should study the viability and advisability of new social welfare and benefit systems to help reduce, where warranted, socio-economic inequality caused by the introduction of AI systems and robotic automation.

4 Environmental impact

4.1 Organisations that develop, make available or use AI systems should assess the overall environmental impact of such AI systems, throughout their implementation, including consumption of resources, energy costs of data storage and processing and the net energy efficiencies or environmental benefits that they may produce. Organisations should seek to promote and implement uses of AI systems with a view to achieving overall carbon neutrality or carbon reduction.

4.2 Governments are encouraged to adjust regulatory regimes and/or promote industry self-regulatory regimes concerning market-entry and/or adoption of AI systems in a way that the possible exposure (in terms of 'opportunities vs. risks') that may result from the public operation of such AI systems is reasonably reflected. Special regimes for intermediary and limited admissions to enable testing and refining of the operation of the AI system can help to expedite the completion of the AI system and improve its safety and reliability.

4.3 In order to ensure and maintain public trust in final human control, governments should consider implementing rules that ensure comprehensive and transparent investigation of such adverse and unanticipated outcomes of AI systems that have occurred through their usage, in particular if these outcomes have lethal or injurious consequences for the humans using such systems. Such investigations should be used for considering adjusting the regulatory framework for AI systems, in particular to develop, where practicable and achievable, a more rounded understanding of how and when such systems should gracefully handover to their human operators in a failure scenario.

4.4 AI has a particular potential to reduce environmentally harmful resource waste and inefficiencies. AI research regarding these objectives should be encouraged. In order to do so, policies must be put in place to ensure the relevant data is accessible and usable in a manner consistent with respect for other principles of the Policy Framework for Responsible AI such as Fairness and Non-Discrimination, Open Data and Fair Competition and Privacy.

5 Weaponised AI

5.1 The use of lethal autonomous weapons systems (LAWS) should respect the principles and standards of and be consistent with international humanitarian law on the use of weapons and wider international human rights law.

5.2 Governments should implement multilateral mechanisms to define, implement and monitor compliance with international agreements regarding the ethical development, use and commerce of LAWS.

5.3 Governments and organisations should refrain from developing, selling or using lethal autonomous weapon systems (LAWS) able to select and engage targets without human control and oversight in all contexts.

5.4 Organisations that develop, make available or use AI systems should inform their employees when they are assigned to projects relating to LAWS.

6 The weaponisation of false or misleading information

6.1 Organisations that develop, make available or use AI systems to filter or promote informational content on internet platforms that is shared or seen by their users should take reasonable measures, consistent with applicable law, to minimise the spread of false or misleading information where there is a material risk that such false or misleading information might lead to significant harm to individuals, groups or democratic institutions.

6.2 AI has the potential to assist in efficiently and pro-actively identifying (and, where appropriate, suppressing) unlawful content such as hate speech or weaponised false or misleading information. AI research into means of accomplishing these objectives in a manner consistent with freedom of expression should be encouraged.

6.3 Organisations that develop, make available or use AI systems on platforms to filter or promote informational content that is shared or seen by their users should provide a mechanism by which users can flag potentially harmful content in a timely manner.

6.4 Organisations that develop, make available or use AI systems on platforms to filter or promote informational content that is shared or seen by their users should provide a mechanism by which content providers can challenge the removal of such content by such organisations from their network or platform in a timely manner.

6.5 Governments should provide clear guidelines to help organisations that develop, make available or use AI systems on platforms identify prohibited content that respect both the rights to dignity and equality and the right to freedom of expression.

6.6 Courts should remain the ultimate arbiters of lawful content.

Principle 2
Accountability

Organisations that develop, make available or use AI systems ought to be accountable for the consequences of their actions and shall designate an individual or individuals who are accountable for the organisation's compliance with the principles of the Policy Framework for Responsible AI or other adopted principles (including analogous principles that may be developed for a specific industry) with the objective of keeping humans behind the machines and AI Human centric.

1 Accountability

1.1. The identity of the individual(s) designated by the organisation to oversee the organisation's compliance with the principles shall be made known upon request.

1.2. Organisations that develop, make available deploy or use AI systems shall use human oversight to carry out determination of the situations in which to carry out delegation to AI decision-making, while ensuring that such use is to accomplish human-chosen objectives. Human oversight can be achieved through three mechanisms, i.e., human-in-the-loop (where humans retain full control to intervene in every decision-making cycle), human-on-the-loop (where humans can intervene during the design cycle of the system and may carry out monitoring) and human-in-command (where humans can oversee the overall activity of the AI system and decide the situations and manner in which it may be used).

1.3. Organisations that develop, make available deploy or use AI systems shall implement policies and practices to give effect to the principles of the Policy Framework for Responsible AI or other adopted principles (including analogous principles that may be developed for a specific industry), including:

i. establishing processes to determine whether, when and how to implement a "Responsible AI Impact Assessment" process;

ii. establishing and implementing "Responsible AI by Design" principles;

iii. establishing procedures to receive and respond to complaints and inquiries;

iv. training staff and communicating to staff information about the organisation's principles, policies and practices; and

v. developing information to explain the organisation's principles, policies and procedures.

2 Government

2.1. Governments should seek to work collaboratively and in a coordinated manner across the international landscape to apply the principles of this Policy Framework for Responsible AI or other analogous internationally recognised principles to ensure consistency of approach and application when holding AI systems to account.

2.2. Governments that assess the potential for "accountability gaps" in existing legal and regulatory frameworks applicable to AI systems should adopt a balanced approach that encourages innovation while mitigating against the risk of significant individual or societal harm.

2.3. Any such legal and regulatory frameworks should promote the eight principles of the Policy Framework for Responsible AI or encompass similar considerations and consider appropriate legal and regulatory enforcement and redress mechanisms.

2.4. Governments should not grant distinct legal personality to AI systems, as doing so would undermine the fundamental principle that humans should ultimately remain accountable for the acts and omissions of AI systems.

2.5. Governments should be transparent and put appropriate human oversight mechanisms in place when utilising AI systems for products or services which are in the public interest, and ensure that the objective and outcomes of such AI Systems are understood by its subjects or citizens.

3 Contextual approach

3.1. The intensity of the accountability obligation will vary according to the degree of autonomy and criticality of the AI system and its potential to cause individual or societal harm. The greater the level of autonomy of the AI system and the greater the criticality of the outcomes that it may produce, the higher the degree of accountability that will apply to the organisation that develops, deploys or uses the AI system ("High Risk AI").

3.2. Where an AI system is deemed to be High Risk AI, a Responsible AI Impact Assessment ("RAIIA") should be conducted and clearly identify the accountable person(s).

Principle 3
Transparency and Explainability

Organisations that develop, make available or use AI systems, and any national laws or industry standards that govern such use, shall ensure that such use is transparent and that the decision outcomes of the AI system are explainable.

1 Purpose

1.1 The Transparency and Explainability principle aims to promote and maintain public trust in AI systems by requiring organisations that develop, make available and use AI systems to provide sufficient information to demonstrate whether decisions made by the AI systems are fair and impartial, support human agency and human autonomy and establish meaningful responsibility and accountability of an AI system's developers and users.

1.2 The Transparency and Explainability principle supports the Ethical Purpose and Societal Benefit principle, the Accountability principle, the Fairness and Non-Discrimination principle, the Safety and Reliability principle and the Privacy principle.

2 Transparency

2.1 Organisations that make available or use an AI system in decision-making processes which produce legal effects concerning an individual or similarly significantly affects an individual shall make readily available meaningful information regarding: (a) the fact that an AI system is being used in a decision-making process; (b) the intended purpose(s); (c) the types of datasets that are used and generated by the AI system; and (d) whether and to what extent the decision-making process may include human participation.

2.2 The information set forth in Section 2.1 should be made readily available to the affected individual before such automated decision-making process occurs in order to provide the individual with an opportunity to assess whether or not to seek a human-centric alternative decision-making process.

3 Explainability

3.1 Organisations that make available or use an AI system in decision-making processes which produce legal effects concerning an individual or similarly significantly affects an individual shall make readily available to such individuals information in objectively clear terms that explains how a decision/outcome was reached, with, at a minimum: a) the information set forth in Section 2.1 above; b) information that offers meaningful interpretability of the algorithmic logic of the AI system; c) meaningful information to understand the decision/outcome; and d) information regarding how the individual may contest the decision or outcome.

3.2 The information set forth in Section 3.1 should be made readily available to an affected individual promptly after such automated decision-making process occurs in order to provide the affected individual with an opportunity to assess whether or not to challenge the decision or outcome.

4 Gradual and contextual approach

4.1 The intensity of the transparency and explainability obligations will depend on a variety of factors, including the nature of the data involved, lack of human participation in the decision-making, the result of the decision and its consequences for the affected individual.

4.2 Ultimately, transparency and explainability must balance the rights, interests and reasonable expectations of the person subject to the decision with the legitimate interests of the organisation making the decision and considerations of overall societal benefit.

4.3 The intensity of the transparency and explainability obligations will generally be higher where the AI system is made available or used in relation to lay persons who are unlikely to understand the technology rather than with an expert whose understanding of the system may be more easily established. Moreover, the intensity of the transparency and explainability obligations will generally be higher where an AI system is used by a public sector organization in the context of enforcing legal obligations rather than by a private sector organisation in the context of offering services.

4.4 The intensity of the transparency and explainability obligations will generally be higher where sensitive personal data is used or where the outcome of the decision will have a material impact on the affected individual's legal or human rights or similarly significantly affects an individual. The intensity of these obligations will generally be lower where non-sensitive personal data or de-personalised data is used or where the impacts on the affected individual's legal or human rights are relatively inconsequential.

4.5 In situations giving rise to high intensity transparency and explainability obligations, organisations that make available or use an AI system in decision-making processes affecting individual rights should, in addition to the information set forth in Sections 2.1 and 3.1 above, make readily available to such individuals meaningful information regarding: a) the traceability and auditability of the algorithmic logic of the AI system, and b) the testing methods used to promote the principles within this policy framework.

5 Transparency and explainability by design

5.1 Organisations that develop AI systems should ensure that the system architecture, algorithmic logic, datasets, testing methods, and all related development and operational policies and procedures serve to incorporate and embed transparency and explainability by design in accordance with national laws and consistent with relevant industry standards. In so far as is reasonably practicable, such systems should aim to be designed from the outset and maintained to promote meaningful transparency and explainability that complements the intended purpose(s) of the AI system.

4.2 The design and development methodologies adopted in Section 5.1 should have the flexibility to embrace evolving industry standards, providing ongoing iterative improvements in transparency and explainability in parallel with advancement in the state of the art during the lifecycle of the AI system.

4.3 Since embedding transparency and explainability into AI system design requires extensive planning and multi-disciplinary expertise, organisations should develop frameworks to assist programmers and developers to design and develop AI systems that possess the desired values and to help reconcile the tensions that exist between accuracy, cost and explainability.

6 Technological neutrality

6.1 The use of an AI system by an organisation does not increase or reduce the procedural and substantive requirements that would otherwise apply if the decision-making process were controlled by a human.

Principle 4
Fairness and Non-Discrimination

Organisations that develop, make available or use AI systems and any national laws that regulate such use shall ensure the non-discrimination of AI outcomes, and shall promote appropriate and effective measures to safeguard fairness in AI use.

1 Awareness and education

1.1 Awareness and education on the possibilities and limits of AI systems is a prerequisite to achieving fairer outcomes.

1.2 Organisations that develop, make available or use AI systems should take steps to ensure that users are aware that AI systems reflect the goals, knowledge and experience of their creators, as well as the limitations of the datasets that are used to train them.

2 Technology and fairness

2.1 Carefully designed AI systems offer the possibility of more consistently fair and non-discriminatory outcomes than are achievable in systems that rely on human decision-making.

2.2 Decisions based on AI systems should be fair and non-discriminatory, judged against the same standards as decision-making processes conducted entirely by humans.

2.3 The use of AI systems by organisations that develop, make available or use AI systems and Governments should not serve to exempt or attenuate the need for fairness, although it may mean refocusing applicable concepts, standards and rules to accommodate AI.

2.4 Users of AI systems and persons subject to their decisions must have an effective way to seek remedy in discriminatory or unfair situations generated by biased or erroneous AI systems, whether used by organisations that develop, make available or use AI systems or governments, and to obtain redress for any harm. Taking into consideration the societal impacts of unfair AI, collective remedies could be a useful tool to address bias or unfairness.

3 Development and monitoring of AI systems

3.1 AI development should be designed to prioritise fairness and non-discrimination. This would involve addressing algorithms and data bias from an early stage and continuously throughout the entire lifecycle of the AI system with a view to ensuring fairness and non-discrimination.

3.2. Before making available or using an AI system, organisations should systematically assess the expected performance of the AI system with respect to potentially unlawful or unfair discrimination as compared to the performance of the processes currently in use.

3.3. Organisations that develop, make available or use AI systems should remain vigilant to the dangers posed by bias. This could be achieved by establishing ethics boards and codes of conduct, and by adopting industry-wide standards and internationally recognised quality seals.

3.4. AI systems with an important social impact could require independent reviewing and testing on a periodic basis.

3.5. In the development and monitoring of AI systems, particular attention should be paid to disadvantaged groups which may be inadequately or unfairly represented in the training data.

4 A comprehensive approach to fairness

4.1 AI systems can perpetuate and exacerbate bias, and have a broad social and economic impact in society. Addressing non-discrimination and fairness in AI use requires a holistic approach. In particular, it requires:

i. the close engagement of technical experts from AI-related fields with statisticians and researchers from the social sciences; and

ii. a combined engagement between governments, organisations that develop, make available or use AI systems and the public at large.

4.2 The Fairness and Non-Discrimination Principle is supported by the Transparency and Accountability Principles. Effective fairness in use of AI systems requires the implementation of measures in connection with both these Principles.

Principle 5
Safety and Reliability

Organisations that develop, make available or use AI systems and any national laws that regulate such use shall adopt design regimes and standards ensuring high safety and reliability of AI systems on one hand while limiting the exposure of developers and deployers on the other hand.

1 Require and/or define explicit ethical and moral principles underpinning the AI system

1.1 Governments and organisations developing, making available or using AI systems should define the relevant set of ethical and moral principles underpinning the AI system to be developed, deployed or used taking into account all relevant circumstances. A system designed to autonomously make decisions will only be acceptable if it operates on the basis of clearly defined principles and within boundaries limiting its decision-making powers.

1.2 Governments and organisations developing, making available or using AI systems should validate the underpinning ethical and moral principles as defined periodically to ensure ongoing accurateness.

2 Standardisation of behaviour

2.1 Governments and organisations developing, making available or using AI systems should recall that ethical and moral principles are not globally uniform but may be impacted e.g., by geographical, religious or social considerations and traditions. To be accepted, AI systems might have to be adjustable in order to meet the local standards in which they will be used.

2.2 Consider whether all possible occurrences should be pre-decided in a way to ensure the consistent behaviour of the AI system, the impact of this on the aggregation of consequences and the moral appropriateness of "weighing the unweighable" such as life vs. life.

3 Ensuring safety, reliability and trust

3.1 Governments should require and organisations should test AI systems thoroughly to ensure that they reliably and robustly adhere, in operation, to the underpinning ethical and moral principles and have been trained with data which are curated and are as 'error-free', 'bias-free' as practicable, given the circumstances. This includes requirements on procedural transparency and technical transparency of the development process of the AI system and the data uses in that respect, as well as the explainability of the decision-making process an AI system will apply when in operation.

3.2 Governments are encouraged to adjust regulatory regimes and/or promote industry self-regulatory regimes for allowing market-entry of AI systems in order to reasonably reflect the positive exposure that may result from the public operation of such AI systems. Special regimes for intermediary and limited admissions to enable testing and refining of the operation of the AI system can help to expedite the completion of the AI system and improve its safety and reliability.

3.3 In order to ensure and maintain public trust in final human control, governments should consider implementing rules that ensure comprehensive and transparent investigation of such adverse and unanticipated outcomes of AI systems that have occurred through their usage, in particular if these outcomes have lethal or injurious consequences for the humans using such systems. Such investigations should be used for considering adjusting the regulatory framework for AI systems; in particular to develop a more rounded understanding of how such systems should gracefully handover to their human operators.

4 Facilitating technological progress at reasonable risks

4.1 Governments are encouraged to consider whether existing legal frameworks such as product liability require adjustment in light of the unique characteristics of AI systems.

4.2 As AI systems might be partially autonomous, organisations developing, deploying or using such systems should pursue continuous monitoring of systems deployed and/or used, allowing human operators to interrupt unanticipated alterations.

4.3 Governments should support and participate in international co-ordination (through bodies such as the International Organisation for Standardisation (ISO) and the International Electrotechnical Commission (IEC)) to develop international standards for the development and deployment of safe and reliable AI systems. Governments are further encouraged to contemplate requirements on continuous monitoring with human oversight as part of their regime balancing encouragement of progress vs. risk avoidance.

Principle 6

Open Data and Fair Competition

Organisations that develop, make available or use AI systems and any national laws that regulate such use shall, without prejudice to normal rules of intellectual property and privacy:

(a) foster open access to, and the portability of, datasets (where privately held), especially where such datasets are deemed significant and important or advance the "state of the art" in the development of AI systems;

(b) ensure that data held by public sector bodies are, in so far as is reasonably practicable, portable, accessible and open; and

(c) encourage open source frameworks and software for AI systems which could similarly be regarded as significant and important and advance the "state of the art."

AI systems must be developed and made available on a "compliance by design" basis in relation to competition/antitrust law.

1 Supporting effective competition in relation to AI systems

1.1 Governments should support and participate in international co-ordination (through bodies such as the OECD and the International Competition Network) to develop best practices and rigourous analysis in understanding the competitive impact of dataset control and AI systems on economic markets.

1.2 Governments should undertake regular reviews to ensure that competition law frameworks and the enforcement tools available to the relevant enforcement authorities are sufficient and effective to ensure sufficient access to necessary inputs, and adequate choice, vibrant rivalry, creative innovation and high quality of output in the development and deployment of AI systems, to the ultimate benefit of consumers.

2 Open data

2.1 Governments should foster and facilitate national infrastructures necessary to promote the portability of and open access to, datasets, especially those that are significant and important, to all elements of society having a vested interest in access to such datasets for research and/or non-commercial use to further advance the "state of the art" in relation to such technology and to ensure the efficacy of existing AI systems. In this regard, governments should give serious consideration to two-tier access models which would allow for free access for academic and research purposes, and paid-for access for commercialised purposes.

2.2 Governments should support open data initiatives in the public or private sector with guidance and research to share wide understanding of the advantages to be gained from open access data, the structures through which datasets can be shared and exchanged, and the processes by which data can be made portable and suitable for open access (including API standardisation, pseudonymisation, aggregation or other curation, where necessary).

2.3 Governments should ensure that the data held by public sector bodies are accessible and open, where possible and where this does not conflict with a public sector mandate to recover taxpayer investment in the collection and curation of such data. Private sector bodies such as industry organisations and trade associations should similarly support and promote open data within their industry sector, making their own datasets open, where possible. The degree of relative influence that private sector organisations have on applicable markets should be assessed on a continuous basis by regulators.

2.4 Organisations that develop, make available or use datasets, especially those which could be regarded as significant or important or which could be regarded as advancing the "state of the art" are similarly encouraged to open up access to, and/or license, such datasets, where possible via chaperoned mechanisms such as Data Trusts.

2.5 Any sharing or licensing of data should be to an extent which is reasonable in the circumstances and must be in compliance with legal, regulatory, contractual and any other obligations or requirements in relation to the data concerned (including privacy, security, freedom of information and other confidentiality considerations). In addition, all stakeholders involved in such sharing or licensing should be very clearly identified in terms of legal roles, duties and responsibilities.

3 Open source AI systems

3.1 Organisations that develop AI systems are normally entitled to commercialise such systems as they wish. However, governments should at a minimum advocate accessibility through open source or other similar licensing arrangements to those innovative AI systems which may be of particular societal benefit or advance the "state of the art" in the field via, for example, targeted incentive schemes.

3.2 Organisations that elect not to release their AI systems as open source software are encouraged nevertheless to license the System on a commercial basis.

3.3 To the extent that an AI system can be subdivided into various constituent parts with general utility and application in other AI use-cases, organisations that elect not to license the AI system as a whole (whether on an open source or commercial basis) are encouraged to license as many of such re-usable components as is possible.

4 Compliance by design with competition/antitrust laws

4.1 Organisations that develop, deploy or use AI systems should design, develop and deploy AI systems in a "compliance by design" manner which ensures consistency with the overarching ethos of subsisting competition/antitrust regimes to promote free and vibrant competition amongst corporate enterprises to the ultimate benefit of consumers.

Principle 7
Privacy

Organisations that develop, make available or use AI systems and any national laws that regulate such use shall endeavour to ensure that AI systems are compliant with privacy norms and regulations, taking into account the unique characteristics of AI systems, and the evolution of standards on privacy.

1 Finding a balance

1.1 There is an inherent and developing conflict between the increasing use of AI systems to process private data, especially personal data; and the increasing regulatory protection afforded internationally to personal and other private data This protection typically applies principles of purpose limitation, data minimisation and storage limitation.

1.2 Governments that regulate the privacy implications of AI systems should do so in a manner that acknowledges the specific characteristics of AI and that does not unduly stifle AI innovation.

1.3 However, governments should foster the privacy principles, in particular of purpose limitation, for personal data within the use of AI systems.

1.3 Organisations that develop, make available and use AI systems should analyse and constantly check their current processes to identify whether they need be updated or amended in any way to ensure that the respect for privacy is given as a central consideration. This includes consideration as to whether and to want extent AI systems actually require the processing of personal (as opposed to, e.g., anonymous) data.

2 The operational challenges ahead for AI users

2.1 AI systems create challenges specifically in relation to the practicalities of meeting of requirements under a number of national legislative regimes, such as in relation to consent and anonymisation of data. Likewise AI systems create challenges as to data subject rights and legal certainty for all parties involved. Accordingly, organisations that develop, deploy or use AI systems and any national laws that regulate such use, shall make provision for alternative lawful bases for the collection and processing of personal data by AI systems, such as a rightful use of the input and output data.

2.2 Organisations that develop, deploy or use AI systems should identify the level of responsibility when they use input or output data for AI systems (e.g., to avoid unlawful discrimination). Organisations should then consider the resulting consequences and obligations, including implementing operational safeguards to protect privacy such as privacy by design principles that are specifically tailored to the specific features of deployed AI systems.

2.3 Organisations that develop, deploy and use AI systems should appoint an AI Ethics Officer, in a role similar to Data Protection Officers under the GDPR, but with specific remit to consider the ethics and regulatory compliance of their use of AI.

3 AI as a tool to support privacy

3.1 Although there are challenges from a privacy perspective from the use of AI, in turn the advent of AI technologies could also be used to help organisations comply with privacy obligations.

Principle 8
AI and Intellectual Property

Organisations that develop, make available or use AI systems should seek to strike a fair balance between benefiting from adequate protection for the intellectual property rights for both the AI system and the AI output and allowing availability for the wider societal benefit. Governments should investigate how AI systems and AI-created output may be afforded adequate protection whilst also ensuring that the innovation is sufficiently disclosed to promote progress.

1 Supporting incentivisation and protection for innovation

1.1 Innovation is of greatest value when it benefits society. Funding is necessary to develop innovation to a level where it can be disseminated and utilised by society. Those from whom funding is sought require a return on their investment. Consequently, there must be incentivisation and protection for innovation if it is to attract investment and be brought to the greater good of society.

1.2 Organisations must therefore be allowed to protect rights in works resulting from the use of AI, whether AI-created works or AI-enabled works.

1.3 However, care needs to be taken to ensure consistency with the policy objectives of existing intellectual property regimes in order to avoid inconsistencies between respective regimes.

1.4 There should be a balance between the protection of innovation and disclosure of innovation.

2 Protection of IP rights

2.1 The possibility of the creation of works by autonomous AI is likely to require amendments to existing IP laws.

2.2 Organisations that develop, deploy or use AI systems should have the option to take necessary steps to protect the rights for the AI system and in the resulting works. Where appropriate these steps should include asserting or obtaining copyrights, obtaining patents, when applicable, and seeking contractual provisions to allow for protection as trade secrets and/or to allocate the rights appropriately between the parties.

2.3 Nevertheless, the protection of IP rights should not be at the expense of allowing open availability to facilitate development for the wider societal benefit.

3 Development of new IP laws

3.1 Governments should be cautious with revising existing IP laws or seeking to introduce new laws.

3.2 Governments should explore the introduction of appropriate legislation (or the interpretation of existing laws) to clarify IP protection of AI-enabled as well as AI-created output.

3.3 When amending existing or implementing new IP laws, governments should seek adequately to balance the interests of all relevant stakeholders.

3.4 Governments should also explore a consensus in relation to AI and IP rights to promote the unhindered data flows across borders and the rapid dissemination of new technologies and seek to address these issues through an international treaty balancing protection with disclosure.

Contributors

EDITORIAL TEAM

John Buyers | Osborne Clarke LLP, UK

Patricia Shaw | Beyond Reach Consulting, UK

Susan Barty | CMS Cameron McKenna Nabarro Olswang LLP, UK

LEAD AUTHORS

Christian Frank | Taylor Wessing Partnerschaftsgesellschaft mbB, Germany

Alexander Tribess | Weitnauer Partnerschaft mbB Rechtsanwälte Steuerberater, Germany

Sonja Dürager | bpv Hügel Rechtsanwälte GmbH, Austria

AUTHORS AND OTHER CONTRIBUTORS

Richard Austin | Deeth Williams Wall, Canada

Phil Catania | Corrs Chambers Westgarth, Australia

Carmen De La Cruz | LEXcellence AG, Switzerland

Steven De Schrijver | Astrea, Belgium

Marco Galli | Gattai, Minoli Partners, Italy

Licia Garotti | Gattai, Minoli Partners, Italy

Jason Haas | Ervin, Cohen & Jessup LLP, USA

Sheena Jacob | CMS-Holborn Asia, Singapore

Louis Jonker | Van Doorne, Netherlands

Kit Lee | Corrs Chambers Westgarth, Australia

Cornelia Mattig | MLL Meyerlustenberger Lachenal Froriep Ltd, Switzerland

Stuart Meyer | Fenwick & West LLP, USA

Charles Morgan | McCarthy Tetrault, Canada

Nikhil Narendran | Trilegal, India

Salvatore Orlando | Ughi e Nunziante Studio Legale, Italy

Ana Pavlović | Endava, Slovenia

Michael Peeters | DAC Beachcroft LLP, UK

Julian Potter | WP Thompson, UK

Alesch Staehelin | Digital Counsel, Switzerland

Per-Kaare Svendsen | Kvale, Advocatfirma, Norway

Padraig Walsh | Tanner De Witt, Hong Kong

Thomas de Weerd | Houthoff, Netherlands

Anthony Wong | AGW Legal & Advisory, Australia

Nicole Beranek Zanon | Härting Rechtsanwälte AG, Switzerland

Alesch
Staehelin,
Digital
Lawyer
&
Counsel (DLC)®
Data, IT/IP & Media

ANTHONY WONG

astrea

Beyond Reach Consulting Limited
Thinking Beyond Reach to Reach Beyond. Applying Digital Ethics by Design into your DNA

bpv HÜGEL

CMS
law·tax·future

CORRS
CHAMBERS
WESTGARTH

DAC BEACHCROFT

DW2
Deeth Williams Wall

endava

ERVIN COHEN & JESSUP LLP

FENWICK

GATTAI, MINOLI, PARTNERS
STUDIO LEGALE

HÄRTING

HOUTHOFF

KVALE

LEXcellence
Legal | Compliance | Regulatory

mccarthy
tetrault

ML

Osborne
Clarke

Tanner
DeWitt
solicitors

TaylorWessing

TRILEGAL

UGHI E NUNZIANTE
STUDIO LEGALE

VANDOORNE

Weitnauer

WP THOMPSON
INTELLECTUAL PROPERTY

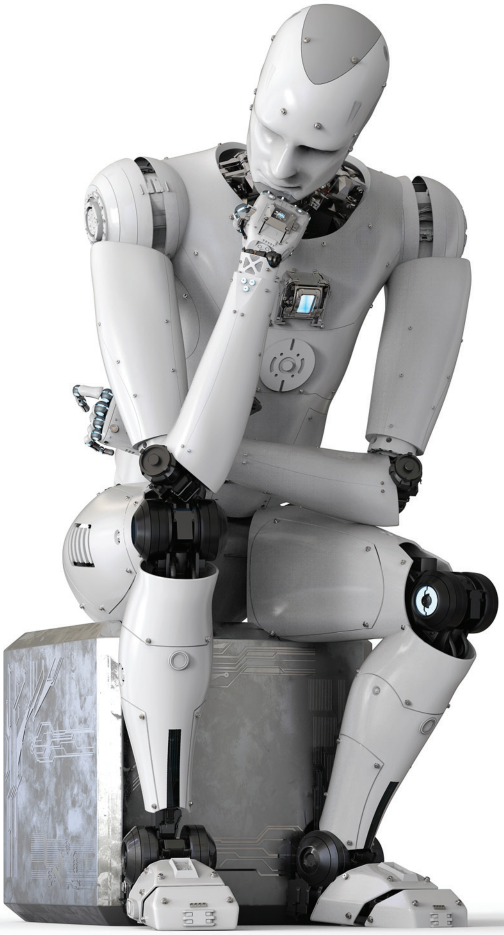
Responsible AI

A GLOBAL POLICY FRAMEWORK

ITechLaw is the leading global organisation for legal professionals focused on technology and law. It has been serving the technology law community worldwide since 1971.

ITechLaw has a global membership base representing more than 70 countries on six continents. Its members reflect a broad spectrum of expertise in the technology law field. Our mission is to create unparalleled opportunities for international networking and for the exchange of knowledge with experts and colleagues around the world.

ITechLaw is a thought leader in many areas of technology law, including as regards the responsible development, deployment and use of artificial intelligence.



ITECHLAW®
INTERNATIONAL TECHNOLOGY LAW ASSOCIATION

itechlaw.org

ISBN 978-1-7339931-2-8
90000
9 781733 993128