Response to European Commission White Paper on Artificial Intelligence

By
Karen Yeung

Professor of Law, Ethics and Informatics
Birmingham Law School & School of Computer Science
University of Birmingham

Member of the EU High Level Expert Group on AI

Summary of Recommendations Concerning the Proposed Regulatory Framework for AI

Recommendation 1 Clarify the conceptual basis of a risk-based regulatory framework

For the purposes of designing a risk-based regulatory framework for AI, **I recommend** that the concept underpinning the idea of a 'risk-based' regulatory framework is clearly defined and clarified. In particular, the essential concept underpinning the idea of risk-based regulation rests on the principle that the level of protection offered to those who are 'at risk' from the regulated activity should be proportionate to the severity, scale and magnitude of that risk. Accordingly, the stringency of the safeguards and level of scrutiny applied to the regulated activity should be proportionate to the attendant risks of that activity. A risk-based regulatory approach also implies that if the relevant activity generates risks considered by the community to be socially unacceptable, then activities that generate risks of this kind should be prohibited and prevented outright (so-called 'red lines') while activities that do not generate any of the relevant risks ought not attract regulatory intervention. The statement on page 10 that "A regulatory framework should concentrate on how to minimise the various risks of potential harm, in particular the most significant ones" does not accurately reflect what a risk-based regulatory framework is intended to do because it reflects an unduly simplified view that the aim of regulation is merely to minimise the 'most significant' risks (p 2)

Recommendation 2 Clarify the regulatory framework's policy objectives

I **recommend** that the Commission identify and specify in much clearer and more precise terms the core policy objectives which the regulatory framework is intended to promote, in light of the risks which AI technologies may generate. These substantive policy goals will be of critical importance, and should inform the design of the regulatory framework, the rules, rights, obligations and powers arising under it, and the manner in which it should be implemented and enforced.

Recommendation 3 Refine and clarify the scope of regulatory framework

All automated decision-making systems (whether fully automated or otherwise) that operate at scale are potentially capable of generating risks of the relevant kind, and which may be exacerbated in various ways through the use of machine learning (ML) and other forms of AI. Accordingly, I **recommend** that:

(a) the scope of the regulatory framework should in principle apply to all software-driven automated decision-making systems that are intended to operate at scale that may generate risks of the relevant kind, whether or not they utilise ML or representational AI (hereafter 'ADM/AI systems');

(b) explicit consideration be given to the relationship between the overlapping operation of the GDPR and the proposed regulatory framework, including attention to the institutional dimensions through which these regimes are implemented and enforced.

Recommendation 4 Traceability and auditability

The White Paper's analysis of the potential adverse effects of AI technologies due to their opacity, complexity, unpredictability and partial autonomy go beyond making the effective enforcement of existing EU law more difficult, but make it considerably more difficult to identify precisely why an AI system generated a particular output, or behaved in a particular way. This, in turn, creates acute difficulties in seeking to allocate and attribute responsibility for any adverse impacts, concomitantly, exacerbating difficulties faced by those adversely affected in obtaining effective redress. It is important that these new obstacles which AI technologies create for processes of responsibility attribution and allocation are taken into account in the design and content of the proposed regulatory framework. In particular, I **recommend** that the regulatory framework should include requirements that all ADM/AI systems that generate significant risks to others must be designed and built to secure 'auditability', thereby ensuring that they can be subject to meaningful review (including but not limited to requirements concerned with automated logging and the retention of log records), thus providing a concrete evidential trail to help securing due accountability and responsibility for the impact and effects of ADM/AI systems.

Recommendation 5 Risks to *all* fundamental rights, systemic risks and risks to collective goods and interests should fall within the purview of the regulatory framework

I recommend that the Commission's proposals for a risk-based regulatory framework for ADM/AI technologies should widen the range of recognised risks to include:

(a) explicit acknowledgement that these includes risks to ***all*** fundamental rights, and not merely legal rights currently protected under existing EU legislation.

(b) that any meaningful assessment of safety risk must also consider ***all*** risks to humans, property and the environment, and not merely those who voluntarily 'use' of the technology;

(c) ***systemic risk*** and **adverse societal level impacts,** including imparts on **collective goods and values**. These include risks to the broader cultural and political environment which fundamental rights and freedoms necessarily presuppose and in which they are anchored.

Recommendation 6 A more fine-grained risk-based scale for the assessment of risk is needed to ensure proportionate safeguards and compliance obligations

I strongly recommend that the proposed regulatory framework adopt a much more fine-grained scale for the assessment of risk, rather than adopting a binary classification system for determining whether or not a particular ADM/AI application falls within the scope of the regulatory framework or not. In particular, I recommend a 5 point scale: this would provide a sufficient level of granularity for which an accompanying set of proportionately demanding regulatory requirements could be attached, along the following lines:

- Level 1: no risk => negligible regulatory requirements
- Level 2: low risk => modest, simple regulatory requirements (voluntary labelling?)
- Level 3: medium risk => significant, but not unduly onerous regulatory requirements
- Level 4: high risk => demanding regulatory safeguards including ex ante approval
- Level 5: unacceptable risk => prohibited

Recommendation 7 Achieving policy congruence

In order to achieve policy congruence, thereby avoiding serious problems of over- and under-inclusion of ADM/AI applications, I **recommend** that

(a) the criteria for inclusion should be based solely on whether or not the proposed application generates a significant individual or collective risks to human health, property the environment or fundamental rights;

(b) there should be no threshold requirement that AI applications fall within a designated sector in order to fall within the scope of the regime.

(c) there should be no need for a residual category of 'exceptional instances' if the proposed criterion for inclusion specified in (a) is adopted.

Recommendation 8 Risk-based regulatory requirements

A risk-based regulatory framework must seek to impose regulatory requirements and safeguards that are proportionate to the risk generated by the regulated activity. Accordingly, **I recommend** that for AI applications, those which generate no or insignificant risks of the relevant kind, no regulatory requirements should be imposed. For higher levels of risk, the stringency of the safeguards and the degree of oversight should become proportionately more demanding, and require increasingly higher levels of transparency and opportunities for public participation and input. AI applications that generate risks that are regarded as unacceptable should be prohibited from deployment. An illustration is provided in Figure 1 (below).

Figure 1: Risk-based regulatory framework:

Risk level	**Seriousness of risk**	**Regulatory requirements**
Level 0	No risk	No requirements
Level 1	Insignificant risk	No requirements
Level 2	Low risk	Mandatory risk assessment via self-evaluation. Obligation to retain the risk assessment documentation and available for inspection and review by a competent authority. No need to lodge risk assessment with public register nor publish. Competent authority may issue a notice of 'insufficiency' if the assessment fails to provide a reasonable assessment of the nature, seriousness and magnitude of risks.
Level 3	Medium risk	Mandatory risk assessment via self-evaluation prior to deployment Obligation to lodge risk assessment on an open public register. Medium risk AI applications can be deployed provided that a published risk mitigation plan setting out the explaining the safeguards that will be put in place to mitigate these risks and keep them under continual review. Mandatory periodic review of risk assessment required. Competent authority empowered to conduct inspection to review quality of risk mitigation plan, and to confirm that the risk mitigation plan is being satisfactorily implemented in practice. If unsatisfactory, competent authority may issue and publicise a notice of improvement.
Level 4	High risk	Mandatory risk assessment via self-evaluation prior to deployment. Obligation to lodge risk assessment on an open public register. Public entitled comment on and make representations identifying potential adverse impacts of the proposed system. AI application cannot be deployed unless and until proposed deployer can 'make the case' to the competent authority, demonstrating why despite the high risks associated with the system, it should nonetheless be permitted because of its expected benefits and the systematic safeguards in place to

		mitigate the identified risks to an acceptable level, and adequately address concerns raised during the public consultation process. Once deployed, AI application is subject to mandatory periodic review of risk assessment, risk mitigation plan, and implementation of risk mitigation programme to confirm that risks are being adequately mitigated and managed.
Level 5	Unacceptable risk	Deployment prohibited

Recommendation 9 Mandatory algorithmic risk assessment

I strongly **recommend** that the proposed regulatory framework should include a mandatory obligation for all those seeking to develop and/or deploy ADM/AI technologies that generates risks that are more than *de minimis* (i.e. Level 2 or above) to undertake an 'algorithmic risk assessment' which assesses the extent to which the proposed application generates significant individual or collective risks to human health, property the environment or fundamental rights. Whether or not that risk assessment must be published, the powers of competent authorities to evaluate its quality, and whether the general public should be entitled to comment upon the proposal in light of the published risk assessment should depend upon whether the application is classed as medium risk (Level 3) or high risk (Level 4), with the highest risk class attracting the most demanding transparency and consultation requirements.

Recommendation 10 Human Rights-Centred Design, Deliberation and Oversight

The various requirements identified in the White Paper which are proposed for AI technologies deemed high risk represent a good starting point, but need to be refined and supplemented by a larger suite of requirements that attach to ADM/AI technologies that have been subject to a more fine-grained risk assessment and classification regime along the lines described above. I **recommend** that these requirements should be considered in light of the core requirements of a risk-based approach to the health, safety and environmental risks posed by such technologies in addition to, and integrated with, a 'human-rights centred approach' to regulatory governance. The latter approach is one that is:

1. anchored in human rights norms and a human rights approach;
2. utilises a coherent and integrated suite of technical, organisational and evaluation tools and techniques, that is
3. subject to legally mandated external oversight by an independent regulator with appropriate investigatory and enforcement powers, and
4. provides opportunities for meaningful stakeholder and public consultation and deliberation

It requires that AI systems must, in proportionate to the seriousness, magnitude and scale of the risk to fundamental rights which they generate, demonstrably comply with the following four families of regulatory requirements:

1. Design and deliberation
2. Assessment, testing and evaluation
3. Independent oversight, investigation and sanction
4. Traceability, evidence and proof

These requirements are elaborated on more fully in Annexe A.

Detailed Analysis and Response

1. General comments

My submission focuses solely on the proposed regulatory framework discussed in section 5 of the White Paper entitled 'An Ecosystem of trust: Regulatory Framework for AI' at page 5. As a general matter, I welcome the White Paper's explicit recognition that AI brings both potential benefits (opportunities) and risks (potential adverse impacts). However, it is important to recognise that AI applications might generate dangers that cannot be meaningfully quantified and are thus are more appropriately referred to as 'threats', rather than 'risks', given that the statistical concept of risk assumes that meaningful probabilistic quantification is possible[1]. Nevertheless, in keeping with the implicit understanding of 'risk' adopted in the White Paper, I will use the term risk in a 'loose' sense to include threats that cannot be meaningfully quantified for the purposes of this submission.

2. What is a 'risk-based regulatory approach' to AI and what does it require?

Although the White Paper does not offer a definition of a risk-based regulatory approach, clues about how this concept appears to be understood within the White Paper can be found in the following statements:

> "A regulatory framework should concentrate on how to minimise the various risks of potential harm, in particular the most significant ones' (p 10)
>
> "A risk-based approach is important to help ensure that the regulatory intervention is proportionate" (p 17)

Provide more clarity on the conceptual basis of risk-based regulation: For the purposes of designing a risk-based regulatory framework for AI, **I recommend** that the concept underpinning the idea of a 'risk-based' regulatory framework is clearly defined and clarified. In particular, the essential concept underpinning the idea of risk-based regulation rests on the principle that the level of protection offered to those who are 'at risk' from the regulated activity should be proportionate to the severity, scale and magnitude of that risk. Accordingly, the stringency of the safeguards and level of scrutiny applied to the regulated activity should be proportionate to the attendant risks of that activity. A risk-based regulatory approach also implies that if the relevant activity generates risks considered by the community to be socially unacceptable, then activities that generate risks of this kind should be prohibited and prevented outright (so-called 'red lines') while activities that do not generate any of the relevant risks ought not attract regulatory intervention. The statement on page 10 that "A regulatory framework should concentrate on how to minimise the various risks of potential harm, in particular the most significant ones" does not accurately reflect what a risk-based regulatory framework is intended to do because it reflects an unduly simplified view that the aim of regulation is merely to minimise the 'most significant' risks.

2.1 What are the policy goals underpinning the proposed regulatory framework?

Critical to the successful development and implementation of a risk-based regulatory framework is the need to:

(1) Identify the nature of the relevant 'risks' that the activity might generate;

[1] Knight, F. H. (1921) *Risk, Uncertainty, and Profit.* Boston, MA: Hart, Schaffner & Marx, Houghton Mifflin Company.

(2) Identify the regulatory framework's core policy objectives in light of those risks. These foundational policy objectives can then inform the design of the regulatory framework, the rules, rights, obligations and powers arising under it, and the manner in which it should be implemented and operationalised.

Yet in respect of both these pre-requisites, the White Paper's analysis is incomplete and unclear.

In relation to the overarching policy objectives of its proposed regulatory framework the White Paper states on page 9

> 'a clear European regulatory framework would build trust among consumers and businesses in AI and therefore speed up the uptake of the technology. Such a regulatory framework should be consistent with other actions to promote Europe's innovation capacity it must ensure socially environmental and economical optimal outcomes and compliance with EU legislation, principles and values. This is particularly relevant in areas where citizens' rights may be most directly affected, for example in the case of AI applications for law enforcement and the judiciary'

As a statement of the overarching policy goals of the proposed regulatory framework, it suffers from a considerable lack of clarity and conceptual coherence, while giving very little indication of how the notion of 'risk' is understood. It appears to imply that the overarching goal of the regulatory is twofold: to 'speed up the uptake of the technology' and to ensure 'socially, environmental and economically optimal outcomes'. However, compliance with 'EU legislation, principles and values' is recognised as a *constraint* on the achievement of those goals, and hence a third policy goal is to ensure that the pursuit of these two primary goals does not entail violation of existing EU legislation or 'EU principles and values.' This characterisation of the proposed regulatory framework's underlying policy-goals is problematic for several reasons:

a) It assumes that the 'uptake of AI technology' is inherently worthwhile, but this is a deeply questionable viewpoint that requires proper justification in order to demonstrate why the pursuit of AI uptake for its own sake will necessarily serve the public interest and human flourishing;

b) The concept of 'socially, environmentally, and economically optimal' outcomes is unclear and indeterminate. It provides no indication of the underlying substantive values that are indeed to inform the notion of 'optimality'. Moreover, assuming that each 'social', 'environmental' and 'economic' dimension represents a distinct domain, there will be inevitable tension and conflict between attempts to 'optimise' outcomes in each of these domains, yet no clues are provided about whether they are to be ranked equally vis-à-vis each other, or how any such tensions and conflicts should be traded-off and resolved;

c) There is a lack of clarity concerning which particular 'EU principles and values' are relevant, given that there are many possible candidate values and principles;

d) The statement makes no reference whatsoever to the notion of 'risks' that have been identified as associated with AI technologies that have prompted recognition of the need for a regulatory framework in the first place, making it even more difficult to understand what the relevant potential adverse impacts associated with AI are that justify the need for regulatory oversight and intervention.

Clarify the regulatory framework's policy objectives: I **recommend** that the Commission identify and specify in much clearer and more precise terms the core policy objectives which the regulatory framework is intended to promote, in light of the risks which ADM/AI technologies may generate. These substantive policy goals will be of critical importance, and should inform the design of the regulatory framework, the rules, rights, obligations and powers arising under it, and the manner in which it should be implemented and enforced.

To this end, it may be helpful to draw upon the way in which the dual objectives of the EU's General Data Protection Regulation (GDPR) are widely understood (and are recognised as sometimes coming into conflict), that is, in seeking to ensure that personal data is collected and processed in a manner that respects fundamental rights while seeking to promote a flourishing data economy within the single European market.

2.2 What are the 'risks' generated by AI technologies?

Assuming that the Commission seeks to establish a 'risk-based' regulatory framework, then identifying more precisely the nature, content and severity of the relevant 'risks' which AI technologies could generate is a matter of critical importance. This would then allow the Commission to identify the appropriate scope of the regulatory framework so that only those technologies and applications that are likely to generate risks of the relevant kind will fall within its remit. The White Paper acknowledges this by stating (under heading of 'problem definition') that

> "…AI can also do harm. This harm might be material (safety and health of individuals, including loss of life, damage to property) and immaterial (loss of privacy, limitations to the right of freedom of expression, human dignity, discrimination for instance in access to employment, and can relate to a wide variety of risks" (p 10)

And later on page 10 states that:

> "The main risks related to the use of AI concern the application of rules designed to protect fundamental rights (including personal data and privacy protection and non-discrimination) as well as safety and liability issues (p 10)

These statements appropriately recognise that AI technologies have the potential to cause both tangible and intangible harm (hereafter 'adverse impacts') and therefore pose 'risks' from which individuals and society will not be adequately protected from in the absence of a legitimate and effective regulatory regime. In other words, it is the capacity of these advanced digital technologies to adversely affect *others*, that justifies the need for regulation. It is helpful, however, to identify in more precise terms what it is that is distinctive about these technologies that could result in adverse impacts. To this end, I suggest that these risks can be more fully understood by attending to the capacity of these automated digital technologies systems to adversely affect *others* in ways to which those others:

- have **not consented**; and
- may be **unaware**, due to the **opacity and complexity** of these technologies and systems; and
- even if they are aware of that these technologies are in use in ways that might have adverse impacts, may lack the **practical capacity to challenge or oppose** the way in which those technologies **are being deployed in relation to them,** including the capacity to seek **correction** of any errors or seek a swift and **effective remedy and recourse** in the case where errors occur.

Moreover, these risks are **radically magnified** due to the capacity of these technologies to operate:

- **automatically**
- at **scale**; and
- in **real** time;
- **controlled remotely** at a **distance** in both time and space relative to its direct impacts;
- amass **large data sets** at a highly granular level in relation to the activities of individuals;

So understood, it is apparent that these risks are generated by ***all*** automated decision-making (ADM) systems that operate at scale, whether or not they utilise machine learning or other forms of artificial intelligence, although the latter are likely to introduce additional **complexity, unpredictability** and **opacity**.

At the same time, there is also clearly overlap with the relevant activities and technologies of interest and the scope and applicability of the GDPR.

Moreover, it is important to recognise that many of the indirect consequential effects of these risks may be unintended yet very serious. Take for example, the tragic case of a 28-year old Australian man who committed suicide having been wrongly billed with a debt of $28,000 by Centrelink, Australia's social security agency by an automated debt collection system.[2] It also reveals the fallacy of treating the automation of administration of benefit and welfare payments as a 'mere administrative task' rather than recognising the profound and direct implications such systems have on the lives and lived experience of many thousands of individuals.

Refine and clarify the scope of regulatory framework: All automated decision-making systems (whether fully automated or otherwise) that operate at scale are potentially capable of generating risks of the relevant kind, and which may be exacerbated in various ways through the use of ML and other forms of AI.

Accordingly, I **recommend** that:

(a) the scope of the regulatory framework should in principle apply to all software-driven automated decision-making systems that are intended to operate at scale that may generate risks of the relevant kind, whether or not they utilise ML or representational AI;

(b) explicit consideration be given to the relationship between the overlapping operation of the GDPR and the proposed regulatory framework, including attention to the institutional dimensions through which these regimes are implemented and enforced.

The White Paper appears to locate the source of these risks posed by AI technologies risks to:

- flaws in the overall design of AI systems (incl human oversight); OR
- use of data without correcting possible bias (p 11).

However, there are many other ways in which AI technologies may generate adverse impacts for individuals and for society. For example, an AI system may be specifically and expressly designed for purposes which necessarily entail serious interferences with the rights of individuals at scale – such as, for example, data-driven smart home 'assist' devices that are configured to continuously identify, detect and track the various activities of those living within the household in order to provide feedback to each of them with personalised ads and alerts and to on-sell the data thereby collected. In this example, the privacy invasive nature of the technology is not due to flawed 'design' but is inherent in the proposed functionality of the technology. Likewise, an algorithmic system may be developed in order to inform particular kinds of decisions, but is built on the basis of a set of data that is a poor proxy for the particular outcomes that the system is seeking to model and predict. For example, various attempts have been made to create 'recidivism risk' predictors that are intended to generate predictions about whether an individual is likely to commit a criminal offence within a specified time future time period, but are trained on the basis of arrest data. Because not all individuals who are arrested are convicted nor even charged with a criminal offence, and because many crimes are committed for which no arrests are made, the resulting risk predictor will predict the likelihood that an individual will be *arrested*, rather than whether an individual will in fact commit a criminal offence. As a result, individuals who are evaluated on the basis of such a model are at

[2] C Karp, (2020) 'Grieving mother whose son, 28, killed himself after he was incorrectly billed $28k by Centrelink breaks her silence on PM's move to repay 'unlawful debts'. 3 June. Available at https://www.dailymail.co.uk/news/article-8383213/Jennifer-Millers-son-killed-received-Centrelink-debt-28-00.html (accessed 10.6.20)

serious risk of being wrongly evaluated with concomitant impacts on their fundamental rights and freedoms (for example, unjustified imprisonment).

However, the White Paper subsequently goes on to recognise that the opacity, complexity, unpredictability and partly autonomous behaviour of AI might create difficulties in verifying compliance with existing laws. It states at p 12:

> The specific characteristics of many AI technologies, including 'opacity (black box effect), complexity, unpredictability and partially autonomous behaviour) 'may make it hard to verify compliance with, and may hamper the effective enforcement of rules of existing EU law meant to protect fundamental rights. Enforcement authorities and affected persons might lack the means to verify how a given decision was made with the involvement of AI was taken and therefore, whether the relevant rules were respected. Individuals and legal entities may face difficulties with effective access to justice in situations where such decisions may negatively affect them' (p 12)

It is indeed the case that the opacity, complexity, unpredictability and partial autonomy of AI create potentially serious risks. These traits together can make it in practice very difficult, if not impossible, to identify precisely why an AI system generated a particular output. Such difficulties in identifying the underlying *sources* and *causes* of a particular AI decision or behaviour may generate a multiplicity of risks, which include *but are not limited to* hampering the effective enforcement of EU laws due to difficulties in verifying compliance with existing legal rules. Rather, it creates *new risks* making it harder (if not impossible in some cases) to verify how and why the system in fact behaved in particular way, with concomitant adverse impacts for ensuring that responsibility and liability for the adverse impacts of an ADM/AI system are duly allocated and that effective redress is provided to those adversely affected[3].

Traceability and auditability: the White Paper's analysis of the potential adverse effects of AI technologies due to their opacity, complexity, unpredictability and partial autonomy go beyond making the effective enforcement of existing EU law more difficult, but make it considerably more difficult to identify precisely why an AI system generated a particular output, or behaved in a particular way, thus creating acute difficulties in seeking to allocate and attribute responsibility for any adverse impacts. This, in turn, exacerbates difficulties faced by those adversely affected in obtaining effective redress. It is important that these new risks to the process of responsibility attribution and allocation for the impacts of AI technologies are taken into account in the design and content of the proposed regulatory framework. In particular, I **recommend** that the regulatory framework should include requirements that, all AI systems that generate significant risks to others must be designed and built to secure 'auditability', thereby ensuring that they can be subject to meaningful review (including but not limited to requirements concerned with automated logging and the retention of log records), thus providing a concrete evidential trail to help securing due accountability and responsibility for the impact and effects of AI systems.

2.3 Risk assessment and evaluation

The essential concept underpinning the idea of risk-based regulation rests on the principle that the level of protection offered to those who are 'at risk' from the regulated activity should be proportionate to the severity, scale and magnitude of that risk (see above). Such an approach necessarily requires that the

[3] K Yeung (2019) *A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework.* Council of Europe Study DGI2019(05), prepared for the Expert Committee on human rights dimensions of automated data processing and different forms of artificial intelligence (MSI-AUT). Available at https://www.coe.int/en/web/freedom-expression/msi-aut/-/highest_rated_assets/SLGLXeB0VOD8/content/focus-on-responsible-ai-a-new-council-of-europe-study-draws-attention-to-the-responsibility-challenges-linked-to-the-use-of-artificial-intelligence?_194_INSTANCE_SLGLXeB0VOD8_viewMode=view/ (Accessed 12.6.20).

relevant 'risks' be identified and attempts made to evaluate their likelihood, and the magnitude and severity of the resulting adverse impacts should those risks materialise.

2.3.1 Tangible risks and intangible risks

The White Paper duly acknowledged that the 'harms' from AI can be both tangible and intangible (including violations of fundamental rights' (p 10). This implicit classification of relevant risks into two kinds, those concerned with tangible risks to health/safety, damage to property and the environment, on the one hand, and intangible risks to fundamental rights on the other, is also reflected in the subsequent statement on page 10 that

> 'The main risks related to the use of AI concern the application of rules designed to protect fundamental rights (including personal data and privacy protection and non-discrimination) as well as safety and liability issues.'

2.3.2 Risks to health, safety and the environment

The White Paper appropriately acknowledges that AI tech may present new safety risks for 'users' when they are embedded in products and services (at p 12) but fails to acknowledge that it may also present safety risks to those who are best understood as 'innocent bystanders' and not merely those who voluntarily choose to use AI technology. Accordingly, any meaningful assessment of safety risk must also consider all risks to humans, property and the environment, and not merely 'users' of the technology.

2.3.3 Risks to fundamental rights

I warmly welcome the White Paper's recognition at page 11 that the use of AI technologies could

> affect the values on which the EU is founded and lead to breaches of the of fundamental rights, including rights to freedom of expression, freedom of assembly, human dignity, non-discrimination based on sex, racial or ethnic origin, religion or belief, disability, age or sexual orientation, as applicable in certain domains, protection of personal data and private life or the right to an effective judicial remedy and a fair trial as well as consumer protection'

and its reference to useful examples, such as the increasing use of ADM systems to evaluate individual's entitlements and obligations, the potential for mass surveillance by states and employers, and its use in content moderation systems that may affect rights to free speech, data protection, privacy and political freedom.

Unfortunately, however, the White Paper then proceeds effectively to ignore these risks to fundamental rights when setting out the core elements of its proposed regulatory framework and its operative requirements. Instead, it appears to limit its concern to the risked posed to existing rights protected under existing EU law, rather than the risks which AI technologies may pose to fundamental rights more generally. This is reflected in various ways. For example, the White Paper identify ways in which AI technologies may undermine the effective enforcement of *existing* law, which forms the focus of suggested improvements to existing EU legislative framework specified on page 14. Although I welcome these suggested improvements, it cannot be assumed that existing legislation exhausts the range of risks which AI technologies may generate (see above). In addition, the following statement appears at page 10

> 'compliance with EU legislation, principles and values is particularly relevant' when citizens' rights are most directly affected, in the case of AI applications for law enforcement and the judiciary (p 10)".

In addition, on page 17 where the following text is suggested as the basis for identifying whether an AI application should fall within the scope of the regime:

> if the uses of AI applications produce *legal or similarly significant effects* for the rights of an individual or a company; that pose risk of injury, death or significant material or immaterial damage; that produce effects that cannot reasonably be avoided by individuals or legal entities

These statements suggests that the relevant 'risks' that will bring an AI application within the scope of the regime are only those which either affect safety/health *or* those which 'pose threats to 'citizens rights most directly affected' or which produce significant effects on the 'legal rights' of an individual or company, rather than the broader set of *fundamental rights* protected under the European Charter of Rights and Freedoms.

Moreover, the analysis presented in the White Paper appears to focus only on highly individualised understandings of 'risk', that is, risks to particular, identifiable individuals or concrete instances of threats to health, safety and the environment in individual cases. In so doing, the **systemic risk** and **adverse societal level impacts** appear to have been completely ignored. This is a particularly serious oversight, given the capacity of AI and automated decision making more generally to operate at *scale*, in *real time*, and to be *controlled at a distance* thereby radically magnifying the seriousness and reach of the risks of both tangible and intangible adverse impacts which AI technologies[4].

Risks to *all* fundamental rights, systemic risks and risks to collective goods and interests should fall within the purview of the regulatory framework: I recommend that the Commission's proposals for a risk-based regulatory framework for ADM/AI technologies should widen the range of recognised risks to include:

(a) explicit acknowledgement that these includes risks to ***all*** fundamental rights, and not merely legal rights currently protected under existing EU legislation.

(b) that any meaningful assessment of safety risk must also consider ***all*** risks to humans, property and the environment, and not merely those who voluntarily 'use' of the technology;

(c) ***systemic risk*** and **adverse societal level impacts,** including imparts on **collective goods and values**. These include risks to the broader cultural and political environment in which fundamental rights and freedoms necessarily presuppose and in which they are anchored.[5]

2.4 What does a proportionate 'risk-based regulatory approach require?

I have already recommended (above) that the Commission clarify the conceptual basis of its proposed risk-based approach to AI regulation. As I have explained, the essential concept underpinning the idea of risk-based regulation rests on the principle that the level of protection offered to those who are 'at risk' from the regulated activity should be proportionate to the severity, scale and magnitude of that risk.

2.4.1 A binary classification system is not a 'proportionate' risk-based system

In addition, although the Commission describes its proposed approach as 'risk-based' and proportionate, the substance of its proposals do *not* in fact reflect a proportionate risk-based approach to regulation. In particular, the White Paper proposes a binary approach, **based on a distinction between AI technologies classed as 'high risk** (and for which ex ante clearance will be required before such technologies can be

[4] K Yeung (2019) 'Why Worry About Decision-Making by Machine?' in K Yeung and M Lodge (eds.) *Algorithmic Regulation*. Oxford University Press: Oxford.

[5] Yeung (2019) *above* n.3.

deployed) **as distinct from all other AI technologies** which will fall outside the scope of the regulatory framework. Hence all other AI technologies will in effect be regarded as 'no risk' although the White Paper contemplates that AI developers and deployers might wish voluntarily participate in some kind of 'voluntary labelling' regime for these applications. This binary approach therefore essentially regards AI technologies as falling within one of two classes and dealt with accordingly: those which are 'high risk' are subject to onerous regulatory requirements, while all other AI technologies are fall outside the regime and therefore effectively regarded as 'no risk'. This binary approach cannot plausibly be characterised as reflecting a 'proportionate, risk-based' approach. It is too blunt an approach and will therefore fail to ensure that the level of protections and safeguards which apply to ADM and AI system development and deployment is proportionate to the 'risks' that they generate. The result will be inadequate protection where needed, while potentially excessive protection will be required where it is not needed. In short, reliance on a simple binary classification system cannot be regarded as proportionate and 'risk-based' and therefore fails to live up to the Commission's intention of devising a 'risk-based' regulatory framework.

A more fine-grain risk-based scale for the assessment of risk is needed to ensure proportionality in the required safeguards and regulatory obligations: I strongly recommend that the Commission adopt a much more fine-grained scale for the assessment of risk, rather than adopting a binary classification system for determining whether or not a particular AI application falls within the scope of the regulatory framework. I recommend that a 5 point scale would provide a sufficient level of granularity in order to devise a set of proportionately demanding regulatory requirements, based on the following:

Level 1: no risk => negligible regulatory requirements
Level 2: low risk => modest, simple regulatory requirements (voluntary labelling?)
Level 3: medium risk => significant, but not unduly onerous regulatory requirements
Level 4: high risk => demanding regulatory safeguards including ex ante approval
Level 5: unacceptable risk => prohibited

2.4.2 Exhaustively listing a catalogue of putative high risk sectors is not proportionate and risk-based

Although the White Paper emphasises the desirability of clarity and certainty in identifying 'high risk' AI applications, its attempt to achieve such clarity through the use of a pre-specified, exhaustive list of 'sectors' falling within the scope of the regulatory framework does so at the unjustified expense of policy congruence while failing to secure the desired level of clarity and certainty. In particular, the White Paper proposes that 'high risk' AI applications will fall within the scope of the regulatory framework if they meet two criteria: first, that they fall within one of the designated 'sectors' that are exhaustively listed as 'high risk' sectors (page 17) and secondly, if they have an 'impact on affected parties.' Both these criteria are unsuitable for establishing an appropriately calibrated risk-based framework for the following reasons.

(a) a sector-by-sector exhaustive list of 'high risk' sectors

Four sectors are identified as exemplars in which AI applications will be regarded as 'high risk' - healthcare, transport, energy and 'parts of the public sector' (p 13). Yet no clear reasons are provided concerning why *these* sectors are singled out. My suspicion is that healthcare, transport and energy were identified as high risk in light of their potential to generate 'safety' risks, while 'parts of the public sector' was recognised as generating risks to fundamental rights, at least rights to privacy and freedom from unlawful discrimination. However, if the Commission's risk-based approach is based on recognition that AI technologies generate risks to both health/safety and property on the one hand, and intangible risks to fundamental rights and values on the other, then why is the media and communications sector excluded, particularly when the threats posed by AI systems to fundamental rights such as the right to freedom of expression, the right to privacy, and the right to freedom from discrimination are clearly placed at risk by the application of AI tools in these sectors in ways that are now recognised as the cause of serious threats to individual rights, freedoms and the fabric of democracy itself?

This example illustrates that if the Commission's proposed approach is adopted, then the policy decision whether or not to list a designated sector as within scope will be regarded as dispositive. Yet this generates two further problems that will result in a failure of policy congruence. Firstly, there is no necessary and sufficient policy congruence between the sector in which the AI application operates and whether it generates a significant risks of adverse impacts. Hence there may well be AI applications that fall within designated sectors, such as the use of ADM/AI systems for weather forecasts that are used in the transport sector but do not generate any serious risk to safety or fundamental rights. Secondly, there are likely to be many AI applications that do *not* fall within the designated sectors yet *do* generate serious risks and yet will remain outside the scope of the regulated framework and thus not subject to adequate safeguards and oversight. At the same time, the focus of debate will then centre on whether or not a particular application is intended to operate within the relevant 'sector', given that the way in which any particular social or economic sector is defined and understood within a given community is a matter of social convention which may change over time and may be highly unstable. In short, although the White Paper assumes that pre-specifying the relevant sectors will provide a 'simple' rule, it will do more harm than good, leading to inappropriately bringing low-risk technologies within the potential scope of the regime, while high risk applications are excluded. The proposed approach is inconsistent with the stated goal of achieving a proportionate risk-based approach, is open to 'gaming' by AI developers and deployers, will generate unproductive disputes about whether a particular AI application is intended to operate within a designated sector, and is likely to significantly reduce the effectiveness of the regulatory framework in achieving its policy objectives and undermine its legitimacy in the eyes of affected stakeholders.

(b) 'impact on affected parties'

The White Paper then proposes that if an AI application falls within a specified sector, it will only fall within the scope of the regulatory regime if it is likely to generate 'significant risks', suggesting that one possible criteria for evaluating the level of risk might be' **'**impact on affected parties' (p 17). It suggests that one possibility might be to bring within the scope of the framework uses of AI applications that

> 'produce legal or similarly significant effects for the rights of an individual or a company; that pose risk of injury, death or significant material or immaterial damage; that produce effects that cannot reasonably be avoided by individuals or legal entities

However, the proposed criteria is too narrowly drawn, capturing only two relevant variable that indicates the level of 'risk' posed by AI technologies (risk of injury, or risk to legal rights which cannot reasonably be avoided). In particular, it suffers from the following shortcomings:

- appears to refers only to *direct impact* on individuals, failing to take account of *societal harms and systematic risks*. Yet these are especially important in relation to AI technologies due to their capacity to operate at scale and in real time.

- fails to recognise that AI applications may pose risks to *fundamental rights*, referring only to (legal) rights of an individual or company. This is a serious deficiency, given that scale, speed and distance at which AI technologies may be controlled, making it possible for a single AI application to violate the fundamental rights of an entire population

- fails to take account of the potential adverse effects on the physical, cultural and political environment or 'commons' that provide the essential social foundations which enables individual rights and freedoms, including democratic participation, to be meaningfully expressed and exercised.

Rather than attempt to formulate some kind of test that would act as a proxy for the relevant risks, it would be more appropriate and effective simply to rely directly upon 'risk' as the touchstone for determining whether any given AI application falls within the scope of the regime. In particular, given that the GDPR already impose mandatory requirements that data controllers undertake a data protection impact assessment in relation to proposed personal data processing proposals that pose a 'high risk' to 'fundamental rights (See GDPR Article 35, Recital 4 and Recital 75), it would be highly beneficial if the same 'high risk to fundamental rights' formulation, supplemented by significant risk to 'health, safety and the environment' was also applied to determine whether a given AI application falls within the scope of the regulatory framework, in which 'risk' is defined in both 'individual and collective' terms.

(c) 'Exceptional' cases

The White Paper then proposes that "exceptional instances" (p 18) may arise in which AI applications might be "considered to be high-risk irrespective of the sector concerned. It gives two examples of such applications, notably those

- affecting employment equality
- remote biometric identification

Although I agree that these two examples would clearly generate a 'high' level of risk to fundamental rights, the White Paper does not explain, however, the basis upon which 'exceptional' instances are to be identified. This not only significantly detracts from the capacity of the proposed regime to offer clear guidance, but also opens up the possibility for arbitrary determinations of which instances should be regarded as 'exceptional'. This is a woefully inadequate basis for bringing particular AI applications within the scope of the regime, allowing the prospect of arbitrary and unprincipled criteria that are without foundation. This outcome would be avoided entirely simply by requiring an assessment of AI applications by reference to risk in the manner specified above.

Achieving policy congruence: In order to achieve policy congruence, thereby avoiding serious problems of over- and under-inclusion of AI applications, I **recommend** that

(a) the criteria for inclusion should be based solely on whether or not the proposed application generates a significant individual or collective risk to human health, property the environment or fundamental rights;

(b) there should be no threshold requirement that AI applications fall within a designated sector in order to fall within the scope of the regime; and

(c) there should be no need for a residual category of 'exceptional instances' if the proposed criterion for inclusion specified in (a) is adopted.

3. Types of requirements: What legal obligations apply to various levels of 'risk'?

3.1 What would a meaningful 'risk-based' approach entail?

As I have already explained (see Recommendation 1 above), the essential concept underpinning the idea of risk-based regulation rests on the principle that the level of protection offered to those who are 'at risk' from the regulated activity should be proportionate to the severity, scale and magnitude of that risk. Accordingly, the stringency of the safeguards and level of scrutiny applied to the regulated activity should be proportionate to the attendant risks of that activity. Thus, if the relevant activity generates risks considered by the community to be socially unacceptable, then activities that generate risks of this kind should be prohibited and prevented outright (so-called 'red lines'). By adopting a more fine-grained classification scheme for assessing the risk of AI applications based on a 5-point scale, it then becomes possible to impose proportionate requirements and safeguards for each class of application, thereby establishing a proportionate, risk-based regulatory framework. Although there is considerable room for variation concerning the specific set of requirements and safeguards that should be attached to each respective risk class, the proposal set out below in Figure 1 provides a useful example.

Figure 1: Risk-based regulatory framework:

Risk level	**Seriousness of risk**	**Regulatory requirements**
Level 0	No risk	No requirements
Level 1	Insignificant risk	No requirements
Level 2	Low risk	Mandatory risk assessment via self-evaluation. Obligation to retain the risk assessment documentation on file which may be inspected and reviewed by a competent authority. No need to lodge risk assessment with public register nor publish. Competent authority may issue a notice of 'insufficiency' if the assessment fails to provide a reasonable assessment of the nature, seriousness and magnitude of risks.
Level 3	Medium risk	Mandatory risk assessment via self-evaluation prior to deployment Obligation to lodge risk assessment on an open public register. Medium risk AI applications can be deployed provided that a published risk mitigation plan setting out the explaining the safeguards that will be put in place to mitigate these risks and keep them under continual review. Mandatory periodic review of risk assessment required. Competent authority empowered to conduct inspection to review quality of risk mitigation plan, and to confirm that the risk mitigation plan is being satisfactorily implemented in practice. If unsatisfactory, competent authority may issue and publicise a notice of improvement.
Level 4	High risk	Mandatory risk assessment via self-evaluation prior to deployment. Obligation to lodge risk assessment on an open public register. Public entitled comment on and make representations identifying potential adverse impacts of the proposed system. AI application cannot be deployed unless and until proposed deployer can 'make the case' to the competent authority, demonstrating why despite the high risks associated with the system, it should nonetheless be permitted because of its expected benefits and the systematic safeguards in place to mitigate the identified risks to an acceptable level, and adequately address concerns raised during the public consultation process. Once deployed, AI application is subject to mandatory periodic review of risk assessment, risk mitigation plan, and implementation of risk mitigation programme to confirm that risks are being adequately mitigated and managed.
Level 5	Unacceptable risk	Deployment prohibited

Risk-based regulatory requirements: A risk-based regulatory framework must seek to impose regulatory requirements and safeguards that are proportionate to the risk generated by the regulated activity. Accordingly, **I recommend** that for AI applications, those which generate no or insignificant risks of the relevant kind, no regulatory requirements should be imposed. For higher levels of risk, the stringency of the safeguards and the degree of oversight should become proportionately more demanding, and require increasingly higher levels of transparency and opportunities for public participation and input. For AI applications that generate risks that are regarded as unacceptable, deployment should be prohibited. An illustration is provided in Figure 1 (above).

3.2 **Regulatory requirements**

Because the White paper proposes a binary classification system pursuant to which only those deemed 'high risk' are subject to regulatory requirements requiring ex ante approval before they can be deployed, it fails to offer an appropriately proportionate set of regulatory requirements and safeguards that reflect the severity, scale and magnitude of the risks which ADM/AI applications pose. A more fine-grained set of regulatory requirements is required that are proportionately more demanding the greater the severity, magnitude and scale of the risks of the proposed AI application, along the lines suggested in Figure 1.

Moreover, in order to implement *any* risk-based regulatory framework for ADM/AI technologies, there must be a mandatory obligation upon all those seeking to develop and/or deploy such technologies that generate risks that are more than *de minimis* (ie Level 2 or above) to undertake an 'algorithmic risk assessment' which assesses whether a proposed technological application generates significant individual or collective risks to human health, property the environment or fundamental rights. Yet the White Paper does not propose any obligations on anyone to undertake a proper risk-assessment, even for those wishing to deploy 'high risk' applications, merely referring to obligations concerning the 'keeping of records' (p 19).

Mandatory algorithmic risk assessment: I strongly **recommend** that the proposed regulatory framework should include a mandatory obligation for all those seeking to develop and/or deploy AI technologies that generate risks that are more than *de minimis* (i.e. Level 2 or above) to undertake an 'algorithmic risk assessment' which evaluates the extent to which the proposed application generates significant individual or collective risks to human health, property, the environment or fundamental rights. Whether or not that risk assessment must be published, and the powers of competent authorities to evaluate its quality, and whether the general public should be entitled to comment upon the proposal in light of the published risk assessment should depend upon whether the application is classed as medium risk (Level 3) or high risk (Level 4), with the highest risk class attracting the most demanding transparency and consultation requirements.

The White Paper then identifies a range of possible requirements, which it acknowledges will require further specification (p 18) including requirements concerning training data, record and data keeping, information provision, robustness and accuracy and human oversight. These are all useful starting points, although I make the following brief observations:

First, in relation to both **training data** and the **keeping of data and of records,** I welcome the White Paper's recognition that:

- the quality, integrity, provenance and appropriateness of the training data will affect the range and severity of risks to both safety and fundamental rights associated with AI applications; and

- that the opacity of AI systems and related difficulties may make it difficult to verify that AI applications comply with applicable rules and requirements, so that record-keeping may be need to be mandated concerning programming of the algorithm, training data, and keeping of the data themselves: to enable trace back and verification (ie methods, processes and techniques to build, train, test and validate) their safety and whether they avoid bias.

However, it is important to recognise that the risks to fundamental rights posed by training data, and the opacity of systems extend *beyond* risks to privacy and equality, potentially generate risks for a much broader range of fundamental rights and the broader environment which should also be taken into consideration;

Secondly, in relation to **information provision** and **robustness and accuracy** I welcome the White Paper's recognition that:

- proactive transparency requirements may be needed, including mandatory requirements to disclose 'systems capabilities and limitations', intended purpose, conditions under which it is expected to function, and expected level of accuracy; and

- that, at least for high risk systems, there is a need to ensure that they are robust and accurate during all life cycle phases, that outcomes are reproducible, and that errors and inconsistencies deal with during all life cycle phases, and resilient to attack/manipulation.

It is not clear, however, why both sets of requirements should apply only to high-risk AI systems, and not more widely, at least also to medium risk systems.

Thirdly, I welcome the White Paper's recognition of the importance of ensuring **human oversight** is in place for high risk systems, but emphasise that these risks must be situated within a larger context of the risks to individuals that AI systems might generate and the concomitant need to ensure that there **are adequate mechanisms for securing meaningful contestation, accountability, responsibility and redress** for those who might be adversely affected by such systems in ways that are unjustified. These requirements should apply to *all* AI and ADM systems which have legal effect or other 'significant impact' on individuals.

Fourthly, I welcome the White Paper's recognition that **biometric identification systems** (esp FRT) pose specific risks for fundamental rights and its proposal to launch broad European debate on circumstances in which specific common safeguards may be justified.

Human Rights-Centred Design, Deliberation and Oversight: The various requirements identified in the White Paper which are proposed for AI technologies deemed high risk represent a good starting point, but need to be refined and supplemented by a larger suite of requirements that attach to AI technologies that have been subject to a more fine-grained risk assessment and classification regime along the lines described above. I **recommend** that these requirements should be considered in light of the core requirements of a risk-based approach to the health, safety and environmental risks posed by such technologies in addition to, and integrated with, a 'human-rights centred approach' to regulatory governance.

The latter approach is one that is:

5. anchored in human rights norms and a human rights approach;
6. utilises a coherent and integrated suite of technical, organisational and evaluation tools and techniques, that is
7. subject to legally mandated external oversight by an independent regulator with appropriate investigatory and enforcement powers, and
8. provides opportunities for meaningful stakeholder and public consultation and deliberation

It requires that AI systems must, in proportionate to the seriousness, magnitude and scale of the risk to fundamental rights which they generate, demonstrably comply with the following four families of regulatory requirements:

1. Design and deliberation
2. Assessment, testing and evaluation
3. Independent oversight, investigation and sanction
4. Traceability, evidence and proof

These requirements are elaborated on more fully in Annexe A.

3.3 Addressees

I welcome the White Paper's recognition of the need to distribute legal obligations among economic operators involved in AI system development and deployment, which may involve multiple actors. While there is merit in imposing the primary legal responsibility on the actors 'best placed to address any potential risks', there is a strong case for applying a responsibility regime in which all those involved in the development and deployment of AI systems are jointly and severally responsible for any resulting adverse impacts in order to ensure that those adversely affected are not left without remedy or recourse, while the relative distribution of responsibility can then be apportioned based on the respective contributions of those involved.

3.4 Compliance, enforcement and redress

The White Paper proposes that for AI applications deemed 'high risk', ex ante conformity assessment could apply to verify that requirements have been met which might include procedures for testing, inspection or certification. While I agree that ex ante evaluation and approval should be required in relation to 'high risk' AI applications, these should be integrated within a larger set of requirements that vary in their level of protection to help secure the proportionate management of risk, based on a more fine-grained approach to risk assessment (see above). It is particularly important that requirements concerning proper documentation and logging are introduced, at least for medium risk systems and above, and periodic inspection for high risk systems in particular (including post-approval regulatory vigilance). To this end, competent authorities must have sufficient resources, investigative and enforcement powers and the technical expertise and capacity to undertake meaningful inspections of ADM/AI applications and systems to verify that adequate safeguards are in fact in place and being implemented in a satisfactory manner. The White Paper does not consider the range of inspection, intervention, enforcement and sanctioning powers which should be available to competent authorities, but these will be of vital importance if the regulatory framework is to deliver successfully on its underlying policy objectives.

In addition, the White Paper does not consider the formulation of the appropriate legal obligation to observe due care in order to ensure that the potential adverse impacts which 'high risk' systems may generate are avoided so that the accompanying level of risk is reduced to a level that is as low as practicable. This would, in effect, require those deploying such systems to do whatever is reasonably practicable to identify and control all relevant risks in order that such technologies may be lawfully deployed. Yet historical experience of 'safety case' regulation demonstrates that effective regulation depends critically on the imposition of general legal obligation on regulated parties to reduce risks 'as low as reasonably practicable' (ALARP) in order to require, in effect, that those deploying such systems must employ best practice safeguards in the deployment of high-risk AI technologies[6].

4. **Voluntary labelling** (p 24)

[6] A Hopkins (2012) 'Explaining "Safety Case"'. *National Research Centre for OHS Regulation*, Australian National University, Working Paper 87. Available via SSRN Network.

While there may be a useful role for a voluntary labelling regime, it is only appropriate for 'low risk' applications. As already indicated, the binary approach currently proposed that classes AI applications as either 'high risk' or 'no risk' does not establish a risk-based framework because it is too blunt and fails to provide for regulatory safeguards and obligations that are proportionate to the risks posed.

5. **Governance** (p 24)

The White Paper proposes a European governance framework for cooperation of national competent authorities to avoid fragmentation of responsibilities, increase member state capacity and ensure Europe progressively equips itself with increase in capacity to test and certify AI-enabled products and services. It also proposes that regular exchange forum (presumably modelled along the lines of the European Data Protection Board) be established and empowered to issue guidance, opinion and expertise, via network of national authorities and at EU level, supported by committee of experts assisting the Commission. Furthermore, these governance structure should guarantee maximum stakeholder participation, while avoiding duplication with existing functions. I cautiously welcome these proposals, although emphasise the need to clarify which member state authorities would qualify as relevant 'national competent authorities', and suggest that further consideration given to the relationship of the proposed regulatory regime with the obligations arising under the GDPR and the potential role that might be played by national data protection agencies, given the shared concern for the protection of fundamental rights and the adoption of a 'risk-based' approach to regulation.

I am deeply sceptical of the efficacy, integrity and robustness of EU certification and conformity assessment mechanisms, particularly following the VW Dieselgate scandal. However, as the White Paper indicates that these assessments would be 'without prejudice to monitoring compliance and ex post enforcement by competent national authorities' (p 24) this might help to mitigate my concerns. It will be vital that the relevant 'competent national authorities' will have the appropriate resources, legal and investigatory powers and technical skills and expertise necessary to ensure that the regulatory framework will secure its underlying policy goals effectively in a manner that would command the confidence of all stakeholders and the broader European public.

Karen Yeung
13.6.20
Birmingham, UK

Annex A

Set out below is an extract describing the core requirements of a regulatory governance regime that has been described as 'human-rights centred design, deliberation and oversight':

K Yeung, A Howes and G Pogrebna (2020) 'AI Governance by Human Rights-Centred Design, Deliberation and Oversight: An End to Ethics Washing' in MD Dubber, F Pasquale and S Das (eds) *The Oxford Handbook of Ethics of AI*, in press. Also available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3435011

Our proposal seeks to …[ensure] conformity with human rights norms, as the basis for a comprehensive design and governance regime constructed around the following four core principles in which human rights norms provide the foundational ethical standards which AI systems must demonstrably comply with…

… It requires that human rights norms are systematically considered at every stage of system design, development and implementation (making interventions where this is identified as necessary), drawing upon and adapting technical methods and techniques for safe software and system design, verification, testing and auditing in order to ensure compliance with human rights norms, together with social and organisational approaches to effective and legitimate regulatory governance. The regime must be mandated by law, and relies critically on external oversight by independent, competent and properly resourced regulatory authorities with appropriate powers of investigation and enforcement, requiring input from both technical and human rights experts, on the one hand, and meaningful input and deliberation from affected stakeholders and the general public on the other. This approach draws upon variety of methods and techniques varying widely in their disciplinary foundations which, suitably adapted and refined to secure conformity with human rights norms, could be drawn together in an integrated manner to form the foundations of a comprehensive design and governance regime. Its foundational ethical standards are comprised of contemporary human rights norms, designed around four principles, namely (a) design and deliberation (b) assessment, testing and evaluation (c) independent oversight, investigation and sanction, and (d) traceability, evidence and proof….

1. Design and deliberation: Central to our approach is a requirement that AI systems should be designed and configured to operate in ways that are compliant with universal human rights standards (such as those, for example, set out in the ECHR), and that, at least for systems identified during the design and development phase as posing a 'high risk' of interfering with human rights, affected stakeholders are consulted about the proposal and given opportunities to express their views about the proposed system's potential impact, in discussion with the system's designers. Where the risks to human rights are assessed as 'high' or 'very high'[7] this would trigger an obligation on system designers to reconsider and redesign the system and/or proposed business model[8] in order to reduce those risks to a form and level regarded as tolerable (understood in terms of a human rights approach to the resolution of conflict between rights and collective interests), in ways that duly accommodate concerns expressed by affected stakeholders and in recognition of the individual and collective benefits which the system is expected to generate.[9]

[7] Jasanoff, S. (2016) *The ethics of invention: technology and the human future*. WW Norton & Company. Raso, F. A. et al, (2018) *Artificial Intelligence & Human Rights: Opportunities & Risks*. Berkman Klein Center Research Publication.

[8] On the potential discriminatory impact of data-driven business models, see Ali, M., et al (2019) 'Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes.' Retrieved from https://arxiv.org/pdf/1904.02095.pdf.

[9] The participatory approach to social impact assessment referred to in the Council of Europe's AI Guideline strongly resonates with the role that our approach ascribes to public deliberation: see The Council of Europe,

2. Assessment, testing and evaluation: Users and others affected by the operation of AI systems (including the general public) can only have justified confidence that AI systems do in fact comply with human rights standards if these systems can be subjected to formal assessment and testing to evaluate their compliance with human rights standards, and if these occur regularly throughout the entire lifecycle of system development: from the initial formulation of a proposal through to design, specification, development, prototyping, and real world implementation, and which includes periodic evaluation of the data sets upon which the system has been trained and upon which it operates.[10] These evaluations form a core element of an overarching 'human rights risk management' approach, which aims to identify potential human rights risks *before* the deployment of AI and other relevant automated systems, and which occurs within a larger 'meta-regulatory' approach to AI governance in which AI system developers and owners are subject to legal duties to demonstrate to a public regulatory authority that their system is human rights compliant.[11] If significant risks to human rights compliance are identified, system developers must reconsider the design specification and system requirements with a view to modifying them in order to reduce those risks to a level that satisfies the tests of necessity and proportion[12] – or, in cases where the threats to human rights are disproportionate and thus unacceptably high, to refrain from proceeding with the development of the system in the form proposed. Once the system has been implemented, periodic review must be undertaken and test and assessment documents duly filed with the public authority. A system of 'AI vigilance' is also needed, entailing the systematic recording of adverse incidents arising from system operations, including potential human rights violations reported by users or the wider public, triggering an obligation on the system provider to review and reassess the system's design and operation, and to report and publicly register any modifications to the system undertaken following this evaluation. Systematic and periodic post-implementation monitoring and vigilance is needed to ensure that AI systems continue to operate in a human-rights compliant manner because, once implemented into real-world settings, AI systems will invariably display emergent effects that are both difficult to anticipate, and may scale very rapidly. Accordingly, there is also an accompanying need for more systematic and sustained research concerned with modelling social systems, in order to better anticipate and predict their unintended adverse societal effects.

Guidelines on AI, (19 Feb 2019) https://rm.coe.int/guidelines-on-artificial-intelligence-and-data-protection/168091f9d8 at 23-24.

[10] Kroll, Joshua A., Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. "Accountable algorithms." *U. Pa. L. Rev.* 165 (2016): 633; Borgesius, Frederik Zuiderveen *Discrimination, Artificial Intelligence and Algorithmic Decision-Making,* Council of Europe, Directorate General for Democracy at 51. Available at https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73, accessed 3.6.2019; Rieke, A., Bogen, M. and Robsinson, D.G. (2018) *Public Scrutiny of Automated Decisions: Early Lessons and Emerging Methods*, An Upturn and Omidyar Network Report. Available at https://www.omidyar.com/insights/public-scrutiny-automated-decisions-early-lessons-and-emerging-methods (Accessed 3.6.19).

[11] Also called 'management-based' regulation and 'enforced self-regulation', meta-regulation refers to a strategy in which regulators do not prescribe how regulated firms should comply, but instead require them to develop their own systems for compliance with legally mandated goals and to demonstrate that compliance to the regulator: Black, J. (2012) 'Paradoxes and Failures: "New Governance" Techniques and the Financial Crisis." *Modern Law Review* 75: 1037, 1045-1048.

[12] The formulation of the appropriate legal standard would need to reflect the established proportionality assessment that is well-established in addressing human rights conflicts and conflicts between human rights and legitimate collective interests, operating as the human rights equivalent as the 'as low as reasonably practicable' ('ALARP') requirement that applies to legal duties to ensure the safety of complex systems, per Hopkins, A. (2012) "Explaining Safety Case". Regulatory Institutions Network Working Paper 87. Available via SSRN network; Thomas, M. (2017). *Safety Critical Systems*. *Gresham Lectures.* London.

3. Independent oversight, investigation and sanction: In order to provide meaningful assurance that AI systems are in fact human rights compliant, rather than merely *claiming* to be human rights-compliant, independent oversight by an external, properly resourced, technically competent oversight body invested with legal powers of investigation and sanction is essential.[13] Because the operation of market-forces cannot provide those who design, develop and deploy AI systems with sufficient incentives to invest the required resources necessary to ensure that AI systems are human rights compliant, our proposed approach must operate within a *legally mandated* institutional structure, including an oversight body with a duty to monitor and enforce substantive and procedural (regulatory) requirements, including those concerning robust design, verification, testing, and evaluation (including appropriate documentation demonstrating that these requirements have been fulfilled) supported by legally mandated stakeholder and public consultation where proposed AI systems pose a 'high risk' to human rights.

We suggest that independent oversight is best designed within a meta-regulatory framework, in which legal duties are placed on AI system developers and operators to demonstrate to a public authority that their systems are human rights compliant.[14] Although there a wide variety of approaches that can be understood as meta-regulatory in form[15] the so-called 'safety case', properly implemented, is considered to have significantly contributed to ensuring the safety of complex systems in several domains, including safety regulation for off-shore petroleum drilling through to the regulation of workplace safety adopted in several Anglo-Commonwealth legal systems.[16] In his discussion of offshore petroleum drilling, Hopkins highlights five basic features of a safety case approach: (1) All operators must prepare a systematic risk (or hazard) management framework which identifies all major hazards and provides detailed plans for how these hazards will be managed, specifying the controls that will be put in place to deal with the identified hazards, and the measures that will be taken to ensure that controls continue to function as intended; (2) A requirement for the operator to 'make the case' to the regulator, that is, to demonstrate to the regulator that the processes that have been undertaken to identify hazards, the methodology they have used to assess risks, and the reasoning (and evidence) that has led them to choose one control rather than another, should be regarded as

[13] Borgesius, Frederik Zuiderveen (2018) *Discrimination, Artificial Intelligence and Algorithmic Decision-Making,* Council of Europe, Directorate General for Democracy, Available at https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73, (Accessed 3. 6.19).

[14] See n.37 above.

[15] See Black, J. (2006). 'Managing Regulatory Risks and Defining the Parameters of Blame: A Focus on the Australian Prudential Regulation Authority.' *Law and Policy* 28: 1-26; Gilad, S. (2010). 'It runs in the family: Meta-regulation and its siblings.' *Regulation & Governance* 4: 485-506. Coglianese, C. and E. Mendelson (2010). Meta-Regulation and Self-Regulation. In R. Baldwin, C. Hood and M. Lodge. (eds.) *The Oxford Handbook of Regulation*. New York, Oxford University Press**:** 146-168.

[16] The so-called 'safety case' movement emerged in the early 1990s in both the UK and USA as an approach to safety certification involving approval and oversight of complex systems, such as aircraft, nuclear power plants and offshore oil exploration. A Hopkins (2012). 'Explaining the "safety case"', *Regulatory Institutions Network*, Working Paper 87. Available at http://www.csb.gov/assets/1/7/WorkingPaper_87.pdf. There have, however, been criticisms of a safety case approach, including concerns about problems of confirmation bias, the need to consider worst case scenarios, reliance on probabilistic assessment to provide assurances of safety rather than the opposite goal of identifying unrecognised hazards, and examples of highly successful process based (rather than performance-based) approaches to securing safety in relation to submarines (eg the SUBSAFE programme): see N Leveson (2011) 'The Use of Safety Cases in Certification and Regulation.' *MIT Engineering Systems Division Working Paper Series*. Available at http://sunnyday.mit.edu/SafetyCases.pdf (Accessed 12 June 2019) 7-9. Leveson observes that the British Health and Safety Executive has applied a safety case regime widely to UK industries, pursuant to which responsibility for controlling risks is placed primarily on those who create and manage hazardous systems, based on three principles: (a) those who create the risks are responsible for controlling those risks, (b) safe operations are achieved by setting and achieving goals rather than by following prescriptive rules, (c) while those goals are set out in legislation, it is for the system providers and operators to develop what they consider to be appropriate methods to achieve those goals.

acceptable. It is then for the regulator to accept (or reject) the case. Although a safety case gives operators considerable independence and flexibility in determining how they will respond to hazards, they do not have free reign: thus if an operator proposes to adopt an inadequate standard, a safety case regulator may challenge the operator and require the adoption of a better standard; (3) A competent, independent and properly resourced regulator with the requisite level of expertise and who can engage in meaningful scrutiny. The regulator's role is not to ensure that hardware is working, or that documents are up to date, but to *audit against the safety case***,** to ensure that the specified controls are functioning as intended, and this necessitates a sophisticated understanding of accident causation and prevention; (4) Employee participation, both in the development of safety cases, and with whom the regulatory officials carrying out site audits must consult; and (5) A general legal duty of care imposed on the operator to do whatever is reasonably practicable to identify and control all hazards. An operator cannot claim to be in compliance just because it has completed a hazard identification process. It is the general duty of care that raises a safety case regime above a 'tick box' or 'blind compliance' mentality, so that a hazard identification process that is demonstrably inadequate would fail to meet the requisite standard.[17] Applying the underlying logic and structure of the 'safety case' approach to human rights compliance would provide developers with considerable flexibility in seeking to 'make the case' to the regulator to demonstrate that their proposed AI systems can be expected to operate in human rights compliant ways.

4. Traceability, evidence and proof:

In order to facilitate meaningful independent oversight and evaluation, AI systems must be designed and built to secure auditability: this means more than merely securing transparency, but is aimed at ensuring that they can be subject to *meaningful review*, thus providing a concrete evidential trial for securing human accountability over AI systems.[18] Not only is it necessary that systems be constructed to produce evidence that they operate as desired,[19] there must be a legal obligation to do so, requiring that crucial design decisions, the testing/assessment process and the outcome of those processes, and the operation of the system itself, are properly documented and provide a clear evidence trail that can be audited by external experts. Drawing again on the experience of the 'safety case' approach , which entails imposing a legal duty on operators to demonstrate to the regulator that robust and comprehensive systems are in place that reduce safety risks to a level that is 'as low as reasonably practical', we envisage the imposition of a suitably formulated legal duty on AI systems developers, owners and operators to demonstrate that these systems are human rights compliant.

To discharge this legal duty, AI system developers would also be subject to legal duties to prepare, maintain and securely store system design documentation, testing and evaluation reports and the system must be designed to routinely generate operational logs which can be inspected and audited by an independent, suitably competent authority. Taken together, these provide an audit trail through which system designers and developers can demonstrate that they have undertaken human rights 'due diligence' - thereby discharging their legal duty to demonstrate that they have discharged their legal duty to reduce the risk of human rights violations to an acceptable level. These traceability and evidential requirements apply to both

[17] Although the general duty of care is linguistically quite imprecise, its meaning has been elaborated on via case law, through numerous cases in which courts have had to decide whether the duty has been complied with. This case law gives fairly clear guidance as to what the general duty means in particular cases: Hopkins, A (2012) "Explaining Safety Case". Regulatory Institutions Network Working Paper 87. Available via SSRN network;*Safety Science* 49: 110-120; Thomas, M. (2017). *Safety Critical Systems*. Gresham Lectures. London.

[18] Bryson, JJ and Theodorou, A (2019) 'How Society Can Maintain Human-Centric Artificial Intelligence.' In M. Toivonen-Noro, and E. Saari (eds.) *Human-Centered Digitalization and Services,* Springer 12-13.

[19] Kroll J et al (2016) 'Accountable algorithms.' *U. Pa. L. Rev.* 165: 633.

the design and development phase (including verification and validation requirements), and the operation and implementation of systems (logging and black box recording of system operations). Taken together, these obligations are intended to ensure that robust and systematic transparency mechanisms are put in place, the aim of which is not complete comprehension, but to provide sufficient information to ensure that human accountability for AI systems can be maintained.[20]

[20] Bryson and Theodorou n 55 at 14.