SUBMISSION TO THE DEPARTMENT OF INDUSTRY, SCIENCE AND RESOURCES

ENSURING SAFE AND RESPONSIBLE AI IN AUSTRALIA

This submission responds to the Department of Industry, Science and Resources' Discussion Paper, *Safe and Responsible AI in Australia*, published in June 2023. The Discussion Paper seeks feedback on possible governance and regulatory responses to ensure "AI is used safely and responsibly". Drawing on our expertise as interdisciplinary researchers with track records in the development, deployment, and regulation of artificial intelligence (AI) and automated decision-making systems, we have focused on how the Australian Government can proactively manage AI to best ensure that it serves the needs of all Australians.

RECOMMENDATIONS

1. AI's potential justifies a pro-innovation approach that is not about being rapid or reckless, but about prioritising societally beneficial uses of AI that ensure public trust and confidence through trustworthy behaviour. A key element of trustworthy behaviour is evidence-based and contextual assessment of whether or not, within a suite of potential approaches, AI offers a safe and responsible approach to problem-solving.
2. We recommend the creation of an AI oversight body to ensure the development and deployment of safe and responsible AI in Australia. This governance mechanism would foster trustworthiness; the essential precursor to increasing public trust.
    a. The engagement of external experts across both the technical and societal aspects of AI is necessary to ensure independent scrutiny and meaningful civic accountability.
    b. Reporting and auditing requirements and investigative and enforcement powers are necessary to provide a robust evidence base and secure safe and responsible conduct.
3. We recommend the Government investigate proactive forms of governance, such as licensing systems, to demonstrate that AI systems comply with certain standards or have certain measures in place to reduce risks.
4. The uncertainties of AI necessitate a regulatory response that is iterative and responsive to new evidence. A risk-based approach can fulfil this need, but the evidence used to gauge risk must be rigorous, objectively assessed according to known standards, and cognisant of the inability to have a complete understanding of the full risk profile of a given AI system.

WHO WE ARE – UWA TECH & POLICY LAB

The UWA Tech & Policy Lab is an interdisciplinary research centre focused on civic accountability in the tech ecosystem. Based at UWA Law School under the leadership of Associate Professor Julia Powles and Professor Jacqueline Alderson, the Lab has expertise in technology law and governance, biomechanics and bioengineering, data analytics and machine learning, and augmented/virtual/extended reality technologies. This submission was led by Dr Hannah Smith.

RESPONSE TO Q1: DEFINITIONS

Given the distinctive opportunity presented by Australia's commitment to safe and responsible AI and the need for a clear and knowable regulatory framework, the essential starting point for the Government is to define what is meant by 'safe AI', 'responsible AI', and the conjunction 'safe and responsible AI'. Who has responsibilities, for what, and to whom? What does safety require, for what, by who, and to whom? Without answers to these questions, safe and responsible AI is a slogan, not a regulatory target.[1]

It is important in any definitions to retain nuance. That a particular technology may pose a risk to human life does not negate the safety considerations of more minor malfunctions. Equally, safe and responsible AI must address the human and environmental considerations associated with the development and deployment of AI technologies, from the labour conditions of workers engaged in data labelling, filtering, and moderating, to the costly ecological impacts of energy and water demands, to the atrophying of different knowledge systems based on the prioritisation of the logic of optimisation. The potential difficulty of definition does not detract from its necessity. Defining safe and responsible AI provides the Australian Government with an opportunity to lead global discussion on the circumstances of introducing and maintaining AI in society, and how to delineate between what is acceptable and unacceptable.

RESPONSE TO Q2-4: GAPS IN EXISTING APPROACHES

OVERLOOKED RISKS AND POTENTIAL MITIGATIONS

Despite a plethora of activity, such as that of the Digital Transformation Agency, AI Ethics Framework, and National AI Centre, it is unclear how existing Government initiatives promote safe and responsible AI in practice, including within Government itself.

For true system-wide feedback on a coordinated and coherent response to AI, several essential concerns must be addressed that are currently missing from the Discussion Paper. Chief among these are concerns regarding labour conditions, national security, and intellectual property. We find it difficult to envisage any AI being designated as 'safe and responsible' if it neglects the often-appalling working conditions of those vital to the training and monitoring of AI systems,[2] or the risks to national security and IP that attend the drive for vast and centralised data repositories.

The Discussion Paper focuses substantially on the risks of bias and how it can arise during AI development and deployment. Bias is a critical concern, and we recognise that the indelibly human inputs to AI mean it will never be devoid of biases rooted in our hopes, fears, uncertainties, and ignorance. Nevertheless, we recommend the Government adopt a more comprehensive understanding of the different types of AI risk in order to better understand their implications and respond appropriately. An excellent illustration is the risk matrix by Maham and Küspert[3], which presents nine relevant risks arising from AI development, deployment, and from subsequent uses

[1] See further, J. Powles, 'What Does it Take to Be a Leader in 'Responsible AI'', CSIRO Machine Learning and Artificial Intelligence Future Science Platform Annual Conference (MARS 2023), Brisbane, 6 June 2023.

[2] See A Williams, M Micelli, and T Gebru, 'The Exploited Labor Behind Artificial Intelligence' (*Noema, 13 October 2022)* available at < https://www.noemamag.com/the-exploited-labor-behind-artificial-intelligence/>

[3] P Maham and S Küspert, *Governing General Purpose AI: A Comprehensive Map of Unreliability, Misuse and Systemic Risks* (July 2023) available at <https://www.stiftung-nv.de/de/publikation/governing-general-purpose-ai-comprehensive-map-unreliability-misuse-and-systemic-risks>

and misuses.[4] The emphasis on systemic risks is particularly pertinent to public sector uses of AI. To the nine risks specified, we would also add further risks to society, such as the exploitation of labour[5] and the contribution of the development and deployment of AI to the climate crisis.[6]

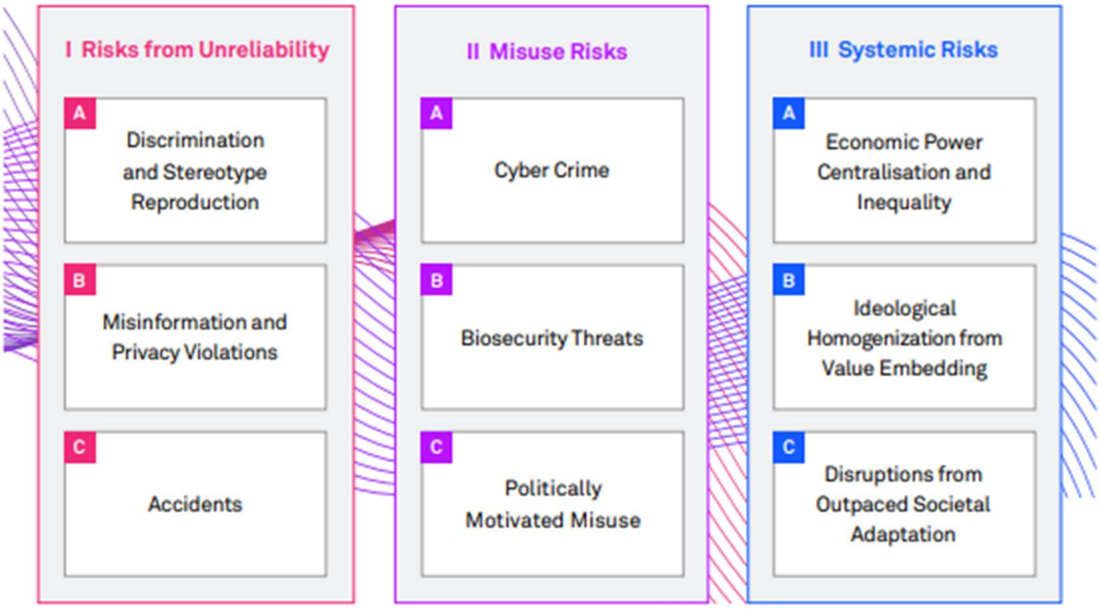
I Risks from Unreliability
A
Discrimination and Stereotype Reproduction
B
Misinformation and Privacy Violations
C
Accidents
II Misuse Risks
A
Cyber Crime
B
Biosecurity Threats
C
Politically Motivated Misuse
III Systemic Risks
A
Economic Power Centralisation and Inequality
B
Ideological Homogenization from Value Embedding
C
Disruptions from Outpaced Societal Adaptation

Figure 1: Risks listed by Maham and Küspert in Governing General Purpose AI: A Comprehensive Map of Unreliability, Misuse and Systemic Risks (July 2023)

A further shortcoming of the Discussion Paper is the limited consideration given to which actors have responsibilities regarding AI, beyond developers and purchasers of AI, or those whose role solely concerns AI from a technical vantage. Many different actors interact with AI, including many within the public sector, and it is critical that their responsibilities are also explicit.

A broader recognition of the spectrum of risks and responsibilities that AI presents will ensure more adequate policy responses and governance mechanisms. Effective policy also requires a clear target, which we recommend to be "promoting societally beneficial uses of AI" – namely, uses that promote human rights, a flourishing democracy, the rule of law,[7] and other human-centred values including those referenced in Australia's AI Ethics Principles,[8] rather than commercial interests. This target demands a graduated and nuanced approach to ensure a just distribution of the benefits and burdens across society. For example, we are particularly concerned that AI deployment risks entrenching inequalities within society that may be masked by *overall*

[4] There is a wealth of literature on the risks of AI and the data that underpins them. See, for example, L Weidinger et al, 'Taxonomy of Risks posed by Language Models' (Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), New York, 2022) available at < https://doi.org/10.1145/3531146.3533088 >.

[5] Fairwork, *Fairwork Cloudwork Ratings 2023* (Fairwork, July 2023) available at < https://fair.work/en/fw/publications/fairwork-cloudwork-ratings-2023-work-in-the-planetary-labour-market/>

[6] E Strubell et al, 'Energy and Policy Considerations for Deep Learning in NLP' (Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, 2019) <https://arxiv.org/abs/1906.02243>

[7] This draws upon the discussion in the report issued by the Council of Europe, *A Legal Framework for AI Systems* (2020) available at <https://edoc.coe.int/en/artificial-intelligence/9648-a-legal-framework-for-ai-systems.html>

[8] Department of Industry, Science and Resources, *Australia's AI Ethics Principles* (2019) available at < https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework>

improvements in economic growth or vague assertions relating to efficiency.[9] Much of the Australian Government and consultancy industry's discussion around AI references the creation of jobs as a key benefit,[10] including the Discussion Paper's inclusion of the Next Generation AI and Emerging Tech Graduates programs.[11] However, the benefits of job creation are unlikely to be equitably distributed. For example, McKinsey's report acknowledges that AI will be "disruptive" but that, eventually, the economy will "adjust" and "new jobs will flow",[12] which neglects the seismic impact this shift will have on the Australian labour force. A Google-commissioned report states that "*only* 29 per cent of the automation driven workplace change will involve workers changing jobs. Whilst these workers are at risk of unemployment, it is important to understand that this does not imply all workers at risk will lose their jobs".[13] The promise of jobs to some individuals at some point in time is no comfort to those whose jobs will be put at risk. The Discussion Paper currently excludes labour concerns from its ambit, but a consideration of who benefits and who suffers must inform any policymaking on safe and responsible AI. The Australian Government has a responsibility to ensure the implementation of AI does not deepen societal inequities.

THE POTENTIAL ROLE OF NON-REGULATORY MEASURES

Non-regulatory measures have a necessary but not sufficient role in developing and deploying safe and responsible AI. For example, industry-led initiatives may be quicker to create and implement than regulatory measures, but they have significant limitations, particularly in relation to monitoring and enforcement. Further, a reliance on non-regulatory measures risks the Government delegating important decisions about the future of Australia's AI landscape to industry actors, whose priorities may not always align with the public interest.

Efforts are also already underway to complement regulatory initiatives with international standards set by bodies such as the OECD, IEEE, and the ISO. The Australian Government could benefit from international alignment with existing standards, such as those relating to AI bias,[14] and engaging with current discussions on future standards, such as those relating to AI management.[15]

[9] See, for example D Leslie et al 'Does "AI" stand for augmenting inequality in the era of covid-19 healthcare?' (2021) BMJ 304 concerning the use of AI entrenching and exacerbating health inequalities, Larson et al., "How We Analyzed the COMPAS Recidivism Algorithm", Pro Publica (May 23, 2016), <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm> that demonstrated an algorithm designed to predict recidivism discriminated against black offenders and the discussion by Hacker and Petkova in 'Reining in the Big Promise of Big Data: Transparency, Inequality , Inequality, and New Regulatory Frontiers' (2017) 15 Nw. J. Tech. & Intellectual Property 1, at p9-11

[10] See for example, McKinsey & Company *Australia's Automation Opportunity: Reigniting Productivity and Inclusive Income Growth* (2019) <https://www.mckinsey.com/featured-insights/future-of-work/australias-automation-opportunity-reigniting-productivity-and-inclusive-income-growth> and AlphaBeta, *The Automation Advantage* (2017) <https://www.roboticsacademy.com.au/automation-changing-way-work/>, a report commissioned by Google.

[11] See the Discussion Paper at p38.

[12] McKinsey & Company *Australia's Automation Opportunity: Reigniting Productivity and Inclusive Income Growth* (2019) at p6

[13] AlphaBeta, *The Automation Advantage* (2017) <https://www.roboticsacademy.com.au/automation-changing-way-work/> at p12

[14] See, for example, the IEEE 7000 Standards and Projects, available at <https://standards.ieee.org/initiatives/autonomous-intelligence-systems/standards/> accessed 31 July 2023

[15] See, for example, the draft ISO 42001 on AI Management Systems available at < https://www.iso.org/standard/81230.html> accessed 31 July 2023.

A level of flexibility in any regulatory response across Government is desirable, due to the complex and emerging nature of AI. However, flexibility is not an inherent attribute of voluntary regulations in the way implied by the Discussion Paper, nor is it absent from binding regulation. Both the EU's GDPR and the proposed AI Act demonstrate that legislation, through the use of a principles-based or risk-management approach, can incorporate the flexibility necessary to respond to the challenges posed by AI. Our concerns about the lack of monitoring and enforcement of voluntary standards inform our recommendation that binding regulation and the creation of an independent, expert oversight body must be considered.

AN AI OVERSIGHT BODY TO COORDINATE AI GOVERNANCE

We advocate the creation of an AI oversight body to ensure the development and deployment of safe and responsible AI in Australia. This could be a robust governance mechanism, shaping a productive relationship between the private and public sectors for the benefit of citizens. This would require the engagement of external experts, across both the technical and societal aspects of AI, to ensure independent scrutiny, the creation of reporting and auditing requirements, and the necessary investigative and enforcement powers to provide a robust evidence base and ensure safe and responsible conduct.

A further role for the oversight body would be to evaluate the efficiency and effectiveness of existing Australian Government work. The number of current and planned consultations suggest that the action taken so far has been insufficient, but in what ways is unclear. If the outcomes of initiatives are not appropriately utilised, the Government's efforts may be better directed towards mainstreaming and iterating the outcomes, rather than starting afresh or adding a further layer of regulatory thinking. If there are true "gaps" in current regulatory approaches, then initiatives need to be more transparent as to what these gaps are and the findings of any reviews of existing initiatives. However, without an evaluation, we are only in a position to speculate.

RESPONSE TO Q6-11: TARGET AREAS

REGULATING PUBLIC AND PRIVATE SECTOR USE OF AI

Approaches to regulating AI must acknowledge that, in many circumstances, there is no bright-line distinction between the public and private sectors. For example, the public sector is heavily reliant on development capabilities in the private sector, while public funding provides impetus to the direction of the private sector. This does not prevent tensions arising where the aims of the public and private sector do not align, and Government must proactively ensure commercial aims do not dominate and capture the development of AI. Rather than to focus on distinct approaches, we suggest that the focus should be on holistic efforts to capitalise upon the respective strengths of the public and private sectors to channel innovation towards serving genuine societal needs.

The private sector's role arises from its agility and ability to rapidly respond to developing circumstances. These attributes can help prevent unsafe or irresponsible uses of AI, but the capacity to act quickly should not be conflated with acting recklessly. All actors must base decisions about the development and deployment of AI on robust evidence and must be willing and able to iterate their governance approaches when further evidence arises. A key element of trustworthy behaviour is evidence-based and contextual assessment of whether or not, within a suite of potential approaches, AI offers a safe and responsible approach to problem-solving.

SUPPORTING RESPONSIBLE AI IN GOVERNMENT AGENCIES

The public sector's role is to guide the development of AI to align with public values. This is particularly the case in relation to funding, partnering, and other contractual agreements between the public and private sectors. The Government should explore developing metrics and other standards to provide an evidence base for decision-makers, informed by a range of expert stakeholders. These metrics should consider both the societal and technical aspects of AI and consider the impact on all Australians, including those that may, for various reasons, diverge from the "norm". We draw attention to the recent Request for Comment of the US National Telecommunications and Information Administration (NTIA) on policies to support AI audits, assessments, certifications, and other mechanisms to create earned trust in AI systems.[16] The NTIA is seeking input on topics including what kinds of data access are necessary, the potential ways regulators and other actors can incentivise the assurance of AI systems, and the different approaches that might be needed in different sectors. These topics demonstrate the potential for public bodies to lead the development of safe and responsible AI, rather than following industry-dominated bodies whose aims may not align with the needs of societally beneficial AI.

Responsible AI within Government requires effective communication as to the role and relevance of AI for public sector workers. Knowledge about responsible AI practices should not be siloed within the IT department or by certain individuals with technical or legal expertise, but instead diffused across a range of roles. Information as to responsible AI practices may arise from junior and/or citizen-facing roles who will experience the human impact of AI practices in a way that technical and legal experts may not, making their perspectives vital.[17] The Australian Government should use the AI oversight body as a clear point of contact for public sector workers to share issues and best practice. We also recommend that any graduate schemes focused on AI rotate around different agencies, to ensure a vital grounding in the different needs of agencies.

IMPLEMENTING TRANSPARENCY AND PROACTIVE PERMISSIONING

Transparency is critical at every stage of the AI lifecycle as a mechanism for ensuring the Government can remain accountable to its citizens and deliver on the promise of AI technologies. The use of metrics and standards alongside a commitment to transparency means that safe and responsible uses of AI must incorporate reporting obligations and requirements. We would also recommend the Government investigates the potential for imposing a licensing system or other form of "permissioning" governance.[18] The appeal of such an approach is that it would proactively protect people from potentially severe consequences from the use of AI by demanding evidence that certain security measures or organisational practices are in place to reduce risks. As the

[16] National Telecommunications and Information Administration, 'AI Accountability Policy Request for Comment' (NTIA, 11 April 2023) < https://www.ntia.gov/issues/artificial-intelligence/request-for-comments>

[17] See, for example, the points raised in the Royal Commission into the Robodebt Scheme Report *Royal Commission into the Robodebt Scheme* (2023), p477, available at <https://robodebt.royalcommission.gov.au/publications/report> where the efforts of the Online Compliance team to flag errors were dismissed because their job was not to check the existence of the debt, only the way it was calculated.

[18] As used in perceived "high-risk" industries such as drug-manufacturing, as regulated by the Therapeutic Goods Authority, and banking, whereby the Australian Prudential Regulation Authority required banking businesses to hold an authorised deposit-taking institution licence.

Robodebt scandal teaches, there is only so much a responsive system can do to repair damage the human cost of poorly developed, implemented, and evaluated automated systems.[19]

REGULATING HIGH-RISK AI APPLICATIONS AND TECHNOLOGIES

Regarding particularly high-risk AI, we emphasise that a pro-innovation approach should not be equated with being reckless in development and deployment of innovation, irrespective of the cost. Australia would not be the only government to decide that certain AI technologies and applications are not amenable to public interests in justice and fairness, and it should embrace the opportunity to delineate these categories. As noted in Attachment B of the Discussion Paper, the proposed EU AI Act (2021) bans certain AI types, such as practices that exploit the vulnerabilities of specific vulnerable groups or entail social scoring, due to the unacceptable risk they pose. More broadly, work led by Noelle Martin, Julia Powles, and Jacqueline Alderson in our Lab has identified the unique risks associated with AI systems that replicate high-fidelity human information (e.g., Meta (Facebook)'s hyper-realistic human avatar program), reasoning that the only way to prevent unmitigated consequences to human safety and autonomy is upstream regulation of the development and deployment of such systems.

Australia must learn from painful experience of attempting to integrate AI into the public sector and the harms to individuals who were often already vulnerable. The Report by the Royal Commissioner into the Robodebt Scheme clearly articulated the numerous failings of the attempts to automate the Australian welfare system.[20] These lessons are reinforced internationally. In Spain, the use of machine learning algorithms to detect apparently ineligible recipients of sick leave benefits has been highly criticised for its opacity, inaccuracy, and failure to deliver on its promises of efficiency and effectiveness.[21] A Danish system for detecting fraudulent welfare payments has been criticised for prioritising the nationality of claimants, leading to criticisms that the system is racist and engaging in ethnic profiling.[22] In the Netherlands, the use of a machine learning algorithm to generate risk scores for welfare recipients was halted in 2021 over concerns of bias.[23]

HOW TO INCREASE PUBLIC TRUST IN AI DEPLOYMENT

In our submission, the Government focus on increasing public trust should be redirected to demonstrating the trustworthiness of Government.[24] Public trust is earned through an ongoing, reflexive commitment to demonstrating trustworthiness through being transparent, accountable, and willing to let a range of actors participate in AI development, deployment, and evaluation.

[19] Section 2.6 of the Report provides a moving account of human costs.

[20] Royal Commissioner, *Royal Commission into the Robodebt Scheme* (2023), Section 2.6, available at < https://robodebt.royalcommission.gov.au/publications/report>

[21] PJ Arandia et al, 'Spain's AI Doctor' *Lighthouse Reports* (17 April 2023) available at <https://www.lighthousereports.com/investigation/spains-ai-doctor/>

[22] G Geiger, 'How Denmark's Welfare State Became a Surveillance Nightmare' *WIRED.com* (7 March 2023) available at <https://www.wired.com/story/algorithms-welfare-state-politics/>

[23] M Burgess et al., 'This Algorithm Could Ruin Your Life' *WIRED.com* (6 March 2023) available at https://www.wired.com/story/welfare-algorithms-discrimination/

[24] For discussions on the difference between trust and trustworthiness see M Aitken, S Cunningham-Burley, and C Pagliari, 'Moving from trust to trustworthiness: Experiences of public engagement in the Scottish Health Informatics Programme' (2016) 43(5) Science and Public Policy 713, available at <https://doi.org/10.1093/scipol/scv075> and M Sheehan et al., 'Trust, trustworthiness and sharing patient data for research' (2021) 47(12) Journal of Medical Ethics 26, available at <http://dx.doi.org/10.1136/medethics-2019-106048>

RESPONSE TO Q14-20: RISK-BASED APPROACHES

HOW TO IMPLEMENT A RISK-BASED APPROACH TO AI

The lens of risk for determining appropriate uses of AI has been influential in other jurisdictions, most notably the EU.[25] However, an often-overlooked aspect of risk-based approaches is that they require knowledge of both benefits and harms. In many instances of AI use, such as when using proprietary AI systems that prevent independent validation of both data inputs and algorithms, the necessary information to determine risk is not available. The dynamism of AI systems means risks may also develop and emerge in unexpected and unforeseen ways, making it prudent to develop governance mechanisms capable of regular updates to risk assessments, with functional systems to report and respond to emergent issues. This includes the ability to roll back the deployment of AI systems where the risk profile is substantially different to an initial assessment. Based on these considerations, in our submission many of the elements of the risk-based approach in Attachment C of the Discussion Paper will not ensure safe and responsible AI.

ChatGPT demonstrates a further difficulty; AI technologies can vary in their level of risk depending on their application.[26] Therefore, if a risk-based approach is used, the assessment of an AI system must be contextual. We recommend the following factors be integrated into the risk-based approach: an assessment of who will be impacted and the risk of entrenching existing inequalities; the existence and effectiveness of any safeguards (e.g., notices and explanations are not wrong *per se* but may fail to fulfil their intended purpose if they do not address the situational context of the users of such systems); the mechanism for monitoring societal impacts; the prioritisation of what is demonstrated (e.g., known risks) above what is promised (e.g., forecast benefits); the potential responses where negative impacts are identified; and the loci of liability.

A risk-based approach must also consider the counterfactual; what would happen if the existing solutions that do not rely on the use of AI are retained? What are some of the benefits and risks that arise from that set of alternative circumstances? This knowledge should form part of any assessment as to the appropriateness of using an AI system. This will enable policy-makers and others involved in decisions surrounding the development and adoption of AI to resist the allure of introducing something new for innovation's sake. Both the current *status quo* and the inherent uncertainty of what AI may deliver are integral aspects to consider within a risk-based approach.

MANDATED REGULATIONS VS. VOLUNTARY SCHEMES

The potential consequences of badly deployed AI are too widespread, significant, and irreversible to justify reliance on self-regulation and voluntary action. All stakeholders have duties and responsibilities that a framework must address, and as previously stated in this response, the appropriate way to do so is through a mandatory regulatory framework.

[25] See 9 European 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 available at <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>

[26] The EU grappled with this issue during the drafting of the AI Act and made clear the requirements that "foundational models" or large powerful AI systems, such as ChatGPT, would have to follow. See, B Perrigo and A Gordon, 'E.U. Takes a Step Closer to Passing the World's Most Comprehensive AI Regulation' *Time.com* (14 June 2023) available at <https://time.com/6287136/eu-ai-regulation/>