José-Miguel BELLO Y VILLARINO

Research Fellow
Law School – University of Sydney
ARC Centre of Excellence for Automated Decision-Making and Society

Sydney NSW 2006 Australia
Telephone: +61 2 9351 0352
jose-miguel.bellovillarino@sydney.edu.au

26 July 2023

SUBMISSION TO SUPPORTING RESPONSIBLE AI: DISCUSSION PAPER

I am a research fellow at the Law School at the University of Sydney and the ARC Centre of Excellence for Automated Decision-Making and Society (ADM+S), I am an award-winning scholar in the domain of regulation of AI, addressing it mainly from a comparative perspective. I have worked for national and international governments including in the regulatory domain of technology and I am a Spanish foreign-service official currently on leave. The present submission is made in my personal capacity and develops some of the points raised in the broader ADM+S submission.

The submission covers two questions: Regulatory approaches to the private and public sector and the role of transparency requirements in the regulation of artificial intelligence.

First, it **advocates for a differentiated approach between the private and public sector** in the short term: the private one based in consolidating reactive regulation to respond to harms and the public one based on pre-emptive regulation based on a proactive approach to mitigation of risks. It notes the advantages of this approach as a regulatory tool and suggests some immediate steps in each area.

Second, it provides several arguments for the government to **reconsider the relevance of transparency for the regulation of AI**, as it will create several challenges in terms of monitoring and enforcement. It notes the difference between governing the use of AI with principles and voluntary frameworks, where transparency has an important role, and with binding regulation, where considering enforcement is essential.

It will be a pleasure to provide my expertise to the Ministry into the next stages of the Supporting responsible AI initiative and related deliverables.

Should different approaches apply to public and private sector use of AI technologies? If so, how should the approaches differ?

1 THE CHOICE BETWEEN PROACTIVE AND REACTIVE REGULATION

There are essentially two options for AI regulation in Australia, consolidating reactive regulation or developing pre-emptive (proactive) regulation. In my view, Australia should develop new pre-emptive regulation for the use of AI in the public sector, while, at least at this stage, could consolidate reactive regulation for the use of AI in the private sector.

This submission first distinguishes between both approaches and then analyses the advantages and disadvantages of consolidating reactive regulation for the private sector, while developing new pre-emptive regulation for the use of AI in the public sector.

The approach suggested may not be ideal for regulatory purposes, but it is immediately applicable, reasonable in terms of regulatory costs, should avoid the most damaging risks derived from the use of AI systems and opens opportunities for leadership in the governance of AI at the international level, through partnerships with like-minded countries, such as Canada.

1.1 PRE-EMPTIVE REGULATION

An AI proactive regulation relies on **creating norms specific for relevant AI systems within the desired regulatory scope**. This requires a **definition for AI** and an assessment of **which AI systems are considered relevant**. Normally pre-emptive regulation will focus on the riskier systems.

Such pre-emptive rules mainly deploy their effect at the time of designing and deploying a system. They most commonly focus on mitigating potential risks generated by the system. It is **generally a safer approach** as any system within scope that is legally deployed would have been prechecked against the rules. The archetypal pre-emptive regulation is the EU proposal for an AI Act.

Yet, it requires a much **more complicated and expensive enforcement regime**. The enforcement—which is mainly done before deployment, but could also happen during its use through continuous monitoring—could be done in two ways.

First, a public entity could check that all systems submitted to it before they are deployed are compliant with the AI-systems specific rules and other relevant rules. This is the mechanism used for extremely risky technologies, such as nuclear energy plants. For AI systems (in general or even for those that involve a risk for human safety or other human rights), systematic ex-ante assessments through public systems would be disproportionately costly.

Second, address compliance through certification against published standards [see ADM+S sub-submission on standards, led by Dr. Henry Fraser]. Developers and deployers ensure that AI systems are in conformity with relevant standards before deployment to be granted a certificate attesting it. Certifiers are often third parties (private entities that charge a fee for those services), but could be the developers/deployers themselves (self-certification). Certifiers are in turn monitored and granted the right to certify systems by public entities. Public entities only intervene at this higher level (certifying certifiers) and/or keeping a registry of certifications. In some jurisdictions, such as the EU the latter are called "notified bodies".

In both cases, the development of those ex-ante technical standards could be left to private not-for-profit bodies led by industry initiative or (more infrequently) it could be a public entity with the technical expertise who develops them (such as NIST in the US or the OECD internationally, which has developed more than 400 standards).

1.2 REACTIVE REGULATION

An AI reactive regulation tends to **focus on harms**. It **does not need to be AI-specific**, as it mainly cares about the triggering event (the harm) irrespective of the cause. For example, the harm from the fall of a lift, caused by the AI system running it or the physical safety system failing, would be essentially treated in the same way. The legal response would be based on the same relevant principles of tort or contract law, even if the process to find the cause of the damage and the tests about causality could differ. In this case, it is likely to be much more complicated to assess causality from the AI system to the falling lift than from a physically defective part produced by a concrete company.

To accommodate for these AI-related complications, new reactive regulations could also be AI-specific. For example, there is a proposal for an EU Directive separate from the EU AI Act initiative focused on adapting liability rules in Europe to AI systems.[1]

Reactive regulations rely mainly on punishments, triggered by events where systems have caused harms. They are **normally cheaper to enforce**, as they will only be triggered for concrete cases that make it to courts (or other enforcement bodies) through complaints. However, the **level of actual safety expected should be lower than proactive systems for the same level of desired safety**, as only the cost of some harms would be borne by companies.

Jurisdictions where these approaches dominate have partially **mitigated the risk of infra-compliance through the creation of a system of enhanced punishment not linked to the harms caused in the concrete case**. This extra cost is added to the compensation for the suffered harm. The **quintessential example is the system of punitive damages in the US**. Having this extra cost for non-compliance incentivises developers and deployers to be more risk-adverse as they take it into account the risk of being found responsible for punitive damages when designing and deploying the systems. Generally, this would work well in relation to physical damages, but not with diffused or low entity harms, where the cost of litigating and coordinating action between complainants would greatly exceed the value of the harm done to each individual person or entity.

Reactive regulations could also be used in a pre-emptive manner in some instances through injunctions (i.e., court orders to stop using the system or restricting the way it is used). However, this still requires a clear knowledge of the potentially affected person or group of people of (i) that the AI system has been actually deployed and (ii) a sufficient understanding of its risks. This is unlikely to happen in many cases.

[1] Proposal of 28 September 2022 for an EU Directive on adapting non contractual civil liability rules to artificial intelligence https://commission.europa.eu/system/files/2022-09/1_1_197605_prop_dir_ai_en.pdf

2 PRE-EMPTIVE AND REACTIVE REGULATION CAN BE COMBINED FOR DIFFERENT PURPOSES

In a 2022 article from Bello y Villarino and Vijeyarasa,[2] we argued that the unavoidable starting point to regulate AI is for the regulator to understand how their existing legal systems would apply to risks and harms caused by AI.

Despite the often-heard claims that AI is not different from other technologies, there is not enough evidence to affirm that (from a reactive point of view) the extant legal regime is adequate to address the particular characteristics of AI systems. In Australia this is particularly relevant in the liability and tort context as there is no federal cause of action for violations of human rights.

On the other hand, we believe that there is a strong case to pass pre-emptive AI-specific legislation for the use of automation in the public sector. This will have significant benefits from a social, economic and industrial point of view.

2.1 REACTIVE REGULATION AS THE STARTING POINT FOR THE PRIVATE SECTOR

There is a strong case to be reactive in the regulation of AI systems deployed in the private sector. First, a solid argument is derived from an interpretation of the classic Collingridge dilemma. At an early stage of the development and deployment of a technology, its impacts cannot be easily predicted. Waiting for the moment we understand the technology may mean that its use is entrenched and bad practices are too difficult to fight. This requires finding a sweet spot for regulation, which, in this case, may require international coordination.

Secondly, strict pre-emptive regulations may hinder innovation or excessively veer it in some directions (e.g., towards less regulated sectors or uses of AI where legal constraints are easier to meet). Yet, having some clear rules eliminating higher degrees of uncertainty will instead favour innovation. Therefore, some level of AI-specific regulation is desirable to promote efficient and safer uses of AI. In this context, if a developer can anticipate a range of harms that its AI system can create and estimate into its model the cost of compensating those harm, it can make sound decisions about whether it is worth developing and deploying the system or not. If there is too much uncertainty about

[2] Bello y Villarino, José-Miguel and Ramona Vijeyarasa, 'International Human Rights, Artificial Intelligence, and the Challenge for the Pondering State: Time to Regulate?' (2022) 40(1) *Nordic Journal of Human Rights* 194.

compensation, only those most willing to embrace risk (or be insolvent in case of harm) are likely to develop their products.

Third, some pre-emptive regulation of the private sector may require international coordination that cannot be achieved in the short term. Specifically, the lack of a significant market in Australia in global terms and the lack of an extensive AI industry also in global terms, may make an Australia-specific regulation of the private sector irrelevant or too selective. For example, some broad scale systems relying on large foundation models with many uses would be too costly to develop to Australian-specific rules. If strict pre-emptive rules are in place, it is likely that those systems would just not be used in Australia, instead of developing a new Australian-compliant one.

There are, however, several options available to the legislator. The first one is to adapt sector-specific legislation to AI systems. Obvious examples are automated mining or self-driving vehicles. Yet, a more **urgent one is to pass horizontal legislation necessary to make reactive regulation for AI a real option, namely in the liability context**. This is likely to require some degree of regulatory direction about:

- Causality
- Distribution of responsibility in the AI supply-chain
- Burden of proof
- Compensation for diffused harms
- Representation of common interests in civil actions
- Urgent injunctions

2.2 Pre-emptive regulation as the starting point for public-sector uses of AI systems

Reactive regulation, however, will not be desirable in the public sector for several reasons. First, the use of AI systems by the public sector must be held to a higher standard considering the scale of the risk in aggregated terms and the different objectives of public action (collective wellbeing and social welfare) and the private sector (commonly, individual profit). Second, the compensation for harms would be borne by the same public sector with no direct consequence for the individuals deciding to deploy those systems. Third, effective or practical uses of AI (or, more generally "automated" systems) in the public sector have to be assessed against alternatives, and they involve policy decisions. For example, a system that may cause some harm to some groups but benefit the majority of the population and increase efficiency levels for the public sector, could be mitigated

for that group through other policies. Fourth, strategically, Australia could learn from the Canadian experience in this domain, focusing first on regulating the public domain. This would be a form of scale-up of the AI Assurance Framework currently applied in NSW (see further below).

If Australia chooses this option, we could present ourselves alongside Canada as a global alternative model for other countries. Our approach would be distinct from the European Union, the United States and China. This Canadian-Australian coordinated approach can even serve as an example for other upper-middle powers such as the United Kingdom, Brazil, Japan or India.

Furthermore, it would place Australia in a privileged position to influence and lead the extension of the work on a international treaty on the governance of AI currently developed within the Council of Europe, but open to non-member states.[3]

2.2.1 REQUIREMENTS

A pre-emptive regulatory system for the public sector will/could require:

1. Learning from the New South Wales experience for AI (NSW AI Assurance Framework)[4] and the Canadian Treasury Board Directive for ADM (ADM Directive)[5] as the two main regulatory systems that can be extrapolated to the Australian federal level. Particularly, the capacity to bypass the NSW framework (if a system is described as "non-AI) may favour opting for the "automated" terminology instead.
2. The development of specific rules adapted to the systems developed within the public sector and those procured from private providers.
3. Adequate testing and certification systems executed or supervised by public officials. Only systems compliant with whatever legal principles established ex-ante for them can be allowed to be deployed by the public sector.
4. Proper risk analysis, that is, the entity and likelihood of actual and representational harms should be considered against benefits for the public good generated by the system.
5. Specific monitoring rules that are adapted to the risk levels of the system.

---

[3] Framework Convention on Artificial Intelligence, Human Rights, Democracy, and the Rule of Law (7 July 2023)

4 'NSW AI Assurance Framework' <https://www.digital.nsw.gov.au/policy/artificial-intelligence/nsw-ai-assurance-framework>

5 Directive on Automated Decision-Making 2019 < https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592>.

6. An analysis of non-AI alternatives. The principles to compare AI to non-AI systems could be incorporated into the regulation, in order create guidelines for accepting the use of an AI system (e.g., digital twins for cities for policy design do not have an alternative which does not use AI, but are extremely risky if that is the main input for policy design of, for example, a traffic system).

2.2.2 ADVANTAGES

The main **advantage of this regulatory approach** is that beyond its limited scope for intervention, its impact can be felt beyond the public sector. There is a compelling case in terms of its expected **externalities over the practices in the private sector**.

2.2.2.1 Impact on public-sector rules on the broader society

1. If the Australian private sector knows that there is a market where they can sell Australian-specific systems compliant with public-sector rules, this is likely to **promote a domestic industry**. This, in turn, is likely to favour that the **practices incorporated in the design and deployment of systems for the public sector**, based on those public welfare considerations, **trickle down into systems designed for the private sector**.
2. Furthermore, because of the scalability and repurposing capacity of these systems**, AI systems developed for the public sector according to more stringent conditions will carry those attributes to their uses in the private sector.** For example, an AI system designed to manage human resources in the Australian public sector; that has been tested for potential discriminatory effects; that has a low carbon footprint; that is capable of producing coherent explanations for its decisions; or can provide levels of confidence for its decisions for potential human supervision, if then sold to the private sector will still have all of those attributes.

2.2.2.2 Fulfilling the obligations of the State

As noted in the submission by the ARC Centre of Excellence for Automated Decision Making and Society (ADM+S) there are powerful legal reasons to proceed with this approach:

1. There is an overlap of the interests of the regulator and the regulated entity. The **State will essentially be regulating itself**, which means that it should be possible to regularly adapt the rules to changing needs and evolving technologies, while remaining committed to the same bigger aim: social welfare. For example, Canada

recently modified its Treasury Board Directive on ADM in 2023 to adapt it to foundation models and generative AI.[6]

2. The Commonwealth has an obligation under international law to protect and promote human rights and must promote the well-being and rights of *all* residents and citizens. This should be done paying attention to societal inequality and social justice and, therefore, introducing positive discrimination objectives into AI systems may be justified and even legally necessary. This would also justify sperate AI rules. In any case, to fulfill these obligations, legislation is necessary to guarantee that government action does not undermine these rights. Public-sector specific regulation must be drafted with these human rights commitments in mind.
3. Having a civil service deploying those systems means that monitoring and enforcement will be possible without judicial intervention. The pre-emptive action will display all its force through supervised implementation. This effect can also be magnified and the legislation refined through procurement policies. Experience with previous systems will access legal documents (terms of reference and specifications for new systems) without a need to modify legislation, if contracting is properly managed.
4. As the scope of government action is very wide, this approach makes possible some degree of regulatory experimentation. Carved-out regimes in certain sectors of public activity could be established, creating 'de facto' regulatory sandboxes for testing regulatory approaches.

[6] Automated Decision-Making, Directive on [2023-04-24]

Question 9: Transparency.

Consultation Question 9 refers to the importance of transparency across the AI lifecycle and invites stakeholders to share your thoughts. I reproduce below a commentary currently at edition stage for an international publication.

The idea of transparency as a requirement for a trustable AI has been thrown around with limited consideration to its meaning from a regulatory point of view (Felzmann et al. 2020; Ehsan et al. 2021; Walmsley 2021). Writing about transparency in an academic paper and using the concept meaningfully from a regulatory point of view are two very different issues.

Yet, instead of focusing our efforts in making transparency operational for regulatory purposes, we all keep using the term as an anchor with the thaumaturgist's power of solving many of the problems in the operation of AI systems.

Whatever we signify with the term transparency (Larsson and Heintz 2020 offer a review of its different discipline-specific conceptions), the idea is appealing for its simplistic and intuitive character: if humans can see through a system, they could understand how it operates—although this requires a leap of faith—and, hence, the system will be trusted more (Ehsan et al. 2021).

More transparency—normally understood as more information and explanations about how the system works and making explicit underlying choices and limitations of the system—would be better than less transparency.

In a simple disjunctive between more and less information, heuristics tells us that more is better—even if we know that that is not always the case (Sunstein 2020).

Generally, offering information does not matter much if there is no incentive for the person receiving it to (i) process that information, (ii) make sense of it, and (iii) act accordingly.

However, when transparency is discussed in policy or legal terms—and it becomes a mandate, for example, ordering an entity involved in the lifecycle of an AI system to ensure the transparency of the system to be authorised to place it in the market—things get messier.

Here, we are not looking at rules of thumb (heuristics) but actual mandates that carry different legal consequences if it is fulfil or violated. In this context **the choice is not**

between making AI systems more or less transparent, but whether this is a plausible and efficient approach to achieve or promote a separate social objective.

Transparency, *per se*, is not a superior value for society than opacity. The dog trained to sniff drugs at the airport is not a transparent system, but we (society) rely on its outputs (bark / no bark) and give a value to it: worth conducting a search or not. If we could wire the dog's brain and adequately analyse its behaviour in terms of triggers of activity, we may better understand the reasons of, for example, some false positives. In exchange, **it would be extremely unpractical, invasive and costly (at least at this stage of development) to wire the brains of each dog.**

The policy or legal choice is then not between full transparency or no transparency at all about how the dog's brain works. It is about how much we care about false positives in relation to the use of the system (deciding to conduct a search or not), compared to the cost of solving that problem through imposed transparency—i.e., the cost of wiring the dog and maintaining that system against the cost/benefit of other options: trusting the dog, installing an entirely different ("transparent") system such as a spectrometry machine, or even not having any system at all and just conducting purely random checks, etc.

Then - **is a legal mandate of transparency for AI systems a good regulatory option or are there better approaches to the problem?**

If nothing else is available, at least transparency will do some good. Public intervention would be limited to punish those that do not comply with it, requiring minimum state-driven action.

Yet, imposing transparency is just a regulatory choice, which needs to be assessed on its merits: How costly would it be for developers? What other problems could it generate? Can it be easily implemented? How efficient will it be to achieve other objectives? And then compared to other options.

These are s not easy questions and, perhaps, transparency is even part of the solution, **but one cannot automatically equate concepts overexploited in ethical frameworks such as transparency or explainability, that, at best, are self-imposed by industry, and take them as the best regulatory choices available to the State**.

For example, it is possible that concepts such as transparency and explainability are not properly suited for regulation as they would require a level of technical expertise which could be acquired by the developer (the reason for their usefulness in self-impose

codes of conduct or ethical framework), but not by the professional user of a system (the bank using the AI system to allocate credits) and, definitely not, by the final "sufferer" of the AI system (the person applying for a credit).

Like with the sniffer dog, what matters is the output of the system and how well it meets our needs and respects our values (e.g., to know that the dog is unlikely to discriminate according to ethnicity, gender or wealth). Regulatory focus, should then start from the outputs of AI systems and not on very complex maps about how the dog's brain works.

REFERENCES

Bello y Villarino, José-Miguel, and Ramona Vijeyarasa. 2022. "International Human Rights, Artificial Intelligence, and the Challenge for the Pondering State: Time to Regulate?" *Nordic Journal of Human Rights* 40(1): 194–215.

Brandeis, Louis. 1913. "What Publicity Can Do." *Harper's Weekly*: 10–13.

Douglas, William O. "Louis Brandeis: Dangerous because incorruptible". *The New York Times.* July 5, 1964

Ehsan, Upol et al. 2021. "Expanding Explainability: Towards Social Transparency in AI Systems." In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA: Association for Computing Machinery, 1–19. https://doi.org/10.1145/3411764.3445188 (October 19, 2022).

Felzmann, Heike, Eduard Fosch-Villaronga, Christoph Lutz, and Aurelia Tamò-Larrieux. 2020. "Towards Transparency by Design for Artificial Intelligence." *Science and Engineering Ethics* 26(6): 3333–61.

Fraser, Henry L., and Jose-Miguel Bello y Villarino. 2021. *Where Residual Risks Reside: A Comparative Approach to Art 9(4) of the European Union's Proposed AI Regulation*. Rochester, NY: Social Science Research Network. SSRN Scholarly Paper. https://papers.ssrn.com/abstract=3960461 (May 10, 2022).

Larsson, Stefan, and Fredrik Heintz. 2020. "Transparency in Artificial Intelligence." *Internet Policy Review* 9(2). https://policyreview.info/concepts/transparency-artificial-intelligence (October 20, 2022).

Loi, Michele, and Matthias Spielkamp. 2021. "Towards Accountability in the Use of Artificial Intelligence for Public Administrations." In *ArXiv:2105.01434 [Cs]*, http://arxiv.org/abs/2105.01434 (July 30, 2021).

Reed, Chris. 2018. "How Should We Regulate Artificial Intelligence?" *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376(2128): 20170360.

Sunstein, Cass R. 2020. *Too Much Information: Understanding What You Don't Want to Know*. Illustrated edition. Cambridge, Massachusetts London, England: The MIT Press.

Walmsley, Joel. 2021. "Artificial Intelligence and the Value of Transparency." *AI & SOCIETY* 36(2): 585–95.