October 4, 2024

Department of Industry, Science and Resources
Australian Government

On behalf of the Center for Data Innovation, we are pleased to submit this response to the Department of Industry, Science and Resources' proposal paper, "Introducing mandatory guardrails for AI in high-risk settings."[1]

The Center for Data Innovation studies the intersection of data, technology, and public policy. Its mission is to formulate and promote pragmatic public policies designed to maximize the benefits of data-driven innovation in the public and private sectors. It educates policymakers and the public about the opportunities and challenges associated with data, as well as technology trends such as open data, artificial intelligence (AI), and the Internet of Things. The Center is part of the Information Technology and Innovation Foundation (ITIF), a nonprofit, nonpartisan think tank.

EXECUTIVE SUMMARY

AI holds vast potential to benefit Australia's economy and society, but realizing this potential requires a thoughtful and balanced regulatory approach. Australia's approach to guardrails for AI would be improved if policymakers:

1. Define sector-specific policy objectives before implementing AI guardrails;
2. Ensure regulation is grounded in evidence-backed risks, not driven by public fears;
3. Avoid myopic AI-focused regulations that overlook broader, non-AI risks.

DEFINE SECTOR-SPECIFIC POLICY OBJECTIVES BEFORE IMPLEMENTING AI GUARDRAILS

The driving force behind the proposal for mandatory AI guardrails is evidence that public trust in AI in Australia is low. Building trust is a laudable goal but that doesn't mean more regulation is the solution. Once a baseline level of regulation is in place, adding more rules does not necessarily increase consumer trust or adoption.[2] Instead, trust tends to rise as individuals gain more experience and familiarity with new technologies over time.

[1] "Introducing mandatory guardrails for AI in high-risk settings: proposals paper," Department of Industry, Science and Resources consultation hub, September 4, 2024, https://consult.industry.gov.au/ai-mandatory-guardrails.
[2] Alan McQuinn and Daniel Castro, "Why Stronger Privacy Regulations Do Not Spur Increased Internet Use," (ITIF, July 2018), https://itif.org/publications/2018/07/11/why-stronger-privacy-regulations-do-not-spur-increased-internet-use.

Moreover, trust isn't a one-size-fits-all concept—its meaning varies across different contexts. In criminal justice, a trustworthy system may mean that AI systems are fair and unbiased. In critical infrastructure, trust is about the systems being reliable, safe, and operationally resilient. When it comes to constitutional rights, trust is rooted in transparency and accountability, ensuring that AI does not infringe on personal freedoms or privacy.

The problem with the proposed guardrails is that they impose blanket solutions with broad, undefined policy goals across these diverse contexts. For example, the guardrail requiring organizations to test AI models and systems to evaluate performance and monitor them once deployed seems sensible on its face. However, without clearly defined policy objectives, it's unclear what specific outcomes these tests are meant to achieve in each sector. In criminal justice, testing could focus on identifying and mitigating discriminatory outcomes. In critical infrastructure, it could assess system reliability and safety under various conditions. In constitutional rights contexts, testing could evaluate how AI decisions can be explained and audited. The crucial term here is “could”—these are complex questions that the government needs to answer before organizations can test.

Without policymakers specifying these objectives, organizations might conduct generic tests that don't address the real concerns in each context. Moreover, adopting broad measures without clear goals can impose unnecessary costs and stifle innovation, as organizations struggle to comply with requirements that may not be relevant or effective in their specific context.

Australia should prioritize defining clear, sector-specific policy objectives for AI deployment. Only after it establishes these objectives should measures—such as specific testing and monitoring requirements—be considered, where necessary, to support these goals. This approach will ensure that the measures are targeted, effective, and directly address the trust concerns in each context.

ENSURE REGULATION IS GROUNDED IN EVIDENCE-BACKED RISKS

While guardrails that address specific risks may be necessary in some contexts, it's crucial that these regulations are based on evidence-backed risks, not simply public fears. The challenge arises when regulation is driven by fear alone—fears that may not accurately reflect the true risks posed by AI systems.

Historically, new technologies have followed a predictable pattern of fear and acceptance, often referred to as the "Tech Panic Cycle."[3] This cycle begins with optimism, followed by rising panic

[3] Patrick Grady and Daniel Castro, “Tech Panics, Generative AI, and the Need for Regulatory Caution,” (Center for Data Innovation, May 2023), https://datainnovation.org/2023/05/tech-panics-generative-ai-and-regulatory-caution.

fueled by media hype, before eventually subsiding as the public grows familiar with the technology. Generative AI is currently in the "rising panic" phase, where exaggerated fears about its potential harm are at their peak.

The danger in regulating too early or too broadly is that it imposes constraints based on hypothetical risks rather than actual, evidence-based concerns. This rush to act, particularly under pressure from public fear, can have unintended consequences. Overly restrictive measures—like those the discussion paper proposes borrowing from frameworks such as the EU's AI Act—have already led to significant compliance costs, as seen in Europe, without necessarily improving trust or safety.[4] In addition to these compliance costs, regulatory burdens imposed by the EU AI Act have caused major technology companies to delay or withhold their latest AI models from the European market, resulting in European businesses missing out on the benefits of cutting-edge AI technology.[5] Australia risks adopting the same burdensome path without fully assessing its own needs and risks.

AVOID MYOPIC AI-FOCUSED REGULATIONS THAT OVERLOOK BROADER, NON-AI RISKS

The discussion paper implies that AI systems, due to their unique characteristics—autonomy, scale, opacity, and others—may require a distinct regulatory framework. It argues that without such measures, AI could pose unprecedented risks, from bias to economic inequality and catastrophic threats. The core assumption in the paper is that AI technologies are the primary source of risk and ignore existing systemic problems in public policy and decision-making, as well as risks from non-AI technology.

Australia's own Robodebt disaster serves to illustrate why. The Robodebt scheme was an automated system the Australian government introduced in 2016 to recover welfare payments. The system matched income data to determine if people owed money and then sent out debt notices. Over 700,000 notices were issued—470,000 of which were later found to be incorrect.

The fundamental problem with Robodebt was not the algorithmic system's inherent capabilities but the lack of supervision and poor policy design behind its deployment. The automated debt collection process was implemented without proper safeguards, legal review, or accountability measures, leading to devastating errors and public distrust. In addition to poor delivery and a negligent appeals process, the system was unlawful because it failed several legal principles, including "innocent until

[4] Mikołaj Barczentewicz and Benjamin Mueller, "More Than Meets The AI: The Hidden Costs of a European Software Law," (Center for Data Innovation, December 2021), https://www2.datainnovation.org/2021-more-than-meets-the-ai.pdf.
[5] Ayesha Bhatti, "EU Competitiveness Hinges on Digital Adoption Not Digital Regulation," (Center for Data Innovation, September 2024), https://datainnovation.org/2024/09/eu-competitiveness-hinges-on-digital-adoption-not-digital-regulation.

proven guilty."[6] It also used a debt-averaging method the government in Australia has since outlawed.[7]

This was not primarily a technological failure but a "massive failure of public administration," as Australia's Federal Court Justice Bernard Murphy put it.[8] By focusing on AI-specific guardrails, the paper risks creating a false sense of security and cementing an approach that neglects real issues: the oversight and decision-making structures behind their deployment. AI is a tool, not a decision-maker, and regulatory efforts should focus on ensuring the responsible use of that tool.

Indeed, instead of focusing on imposing AI-specific guardrails, Australia's regulatory framework should prioritize mechanisms for accountability, transparency, and traceability in decision-making processes, particularly within the public sector. Policymakers should ensure that systems like Robodebt are governed by robust human oversight, legal checks, and clear lines of accountability, rather than scapegoating AI.

[6] Richard Glenn, "Centrelink's automated debt raising and recovery system: A report about the Department of Human Services' online compliance intervention system for debt raising and recovery," (Canberra: Commonwealth Ombudsman, 2017), https://www.ombudsman.gov.au/__data/assets/pdf_file/0022/43528/Report-Centrelinks-automated-debt-raising-and-recovery-system-April2017.pdf.
[7] Jordan Hayne and Matthew Doran, "Government to pay back $721m in Robodebt, all debts to be waived," ABC News (2020, https://www.abc.net.au/news/2020-05-29/federal-government-refundrobodebt-scheme-repay-debts/12299410.
[8] Rebecca Turner, "Robodebt condemned as a 'shameful chapter' in withering assessment by federal court judge," ABC News (2021), https://www.abc.net.au/news/2021-06-11/robodebt-condemned-byfederal-court-judge-as-shameful-chapter/100207674.