Actuaries Institute.

25 July 2023

Department of Industry, Science and Resources
Technology Strategy Branch
GPO Box 2013
Canberra ACT 2600

Email: digitaleconomy@industry.gov.au

Dear Sir/Madam

Response to Department of Industry, Science and Resources – Supporting Responsible AI Discussion Paper

The Actuaries Institute ('the Institute') welcomes the opportunity to provide responses to the Department of Industry, Science and Resources Discussion Paper – Safe and responsible AI in Australia ('Discussion Paper').

The Institute is the peak professional body for actuaries in Australia. Our members have significant involvement in Artificial Intelligence ('AI') in varied sectors of the economy and have contributed to several Government consultations and other work related to the topic of the Discussion Paper. Our contributions include:

- Submission to the Australian Human Rights Commission Discussion Paper 'Human Rights and Technology Discussion Paper', March 2020, noting the Institute's recommendations were referenced several times in the AHRC's final report;
- Institute Members piloting the Australian Government's AI Ethics Principles Framework during 2020-2021, as described in this conference paper;
- The Institute CEO's Paper at the joint ABS/RBA Economic Implications of the Digital Economy Conference in March 2022;
- The Institute's response to the Department of the Prime Minister and Cabinet, Digital Technology Taskforce Issues Paper 'Positioning Australia as a Leader in Digital Economy Regulation – Automated Decision making and AI Regulation' in April 2022;
- The Australian Human Rights Commission and Actuaries Institute joint Guidance Resource 'Artificial intelligence and discrimination in insurance pricing and underwriting', published in December 2022; and
- The Institute's response to the Privacy Act Review Report in March 2023.

Actuaries Institute.

Set out below is our general perspective on what constitutes good regulation for AI, and general recommendations relating to the themes of the Discussion Paper. Further detailed commentary on these recommendations, and additional comments relating to other questions raised in the Discussion Paper covering AI literacy, transparency and the potential for bans, are provided in the Attachment.

Good Regulation for AI

The Institute's principles guiding our response to policy development include a commitment to 'good regulation'. We consider that excessive or unnecessary regulation can obstruct a market from functioning efficiently and can undermine the public interest. Some key elements of good regulation include:

- proportionality – between the regulatory solution and the problem that it intends to solve; and
- appropriate regulatory tools – using self-regulation where possible while acknowledging that prescription can sometimes be appropriate.

The fundamental point made by this principle—and in Institute publications and positions on regulation—is the importance of balance and proportionality. Poorly designed or poorly targeted regulation can create as many problems as it seeks to solve.

It is often suggested that AI and Automated Decision Making ('ADM') are new and somewhat special, and so existing rules do not exist or do not apply. In many situations, this is simply not true. AI/ADM has been present for a long time in many industries, and in many of those areas decisions made by AI/ADM are already subject to broad, principles-based regulation, as are equivalent human decisions.

Given this, the approach considered by the Discussion Paper – to create new regulation applying in a broad manner but only to AI – is likely to cause a range of challenges.

- The approach is likely to prove cumbersome, particularly as technology evolves. The recent debates surrounding the European Union's proposed AI Act regarding the definition of 'AI', and on incorporation of generative AI and foundation models, illustrates this issue.
- Technology-specific regulation may be inconsistent with, or may conflict with, existing principles-based regulation across a wide range of sectors. This will cause challenges for implementation, and then ongoing challenges for both compliance and enforcement.
- AI decisions and equivalent non-AI decisions giving the exact same outcomes for consumers may be subject to different rules. At the extreme, we may find one is allowed where the other is not. This is not only illogical, but it also creates an undesirable loophole which could be exploited.
- Systems which blend a range of human and AI components may be challenging to operate, particularly if regulations for different components strictly vary depending on mechanisms. This will be an ongoing challenge for both compliance and enforcement.

Additionally, there are structural challenges with the 'risk-based' approach proposed, which we outline in detail in the Attachment.

Instead, we should first prioritise creating clarity of the regulation we already have, where it exists. It has become apparent that there are uncertainties in existing regulation which the broad emergence of AI has made clear.

For example, and as we detailed in our response to the Australian Human Rights Commission in 2020, there are well-understood uncertainties surrounding existing law relating to indirect discrimination. These uncertainties are not new. AI and ADM do not create them, rather the recent discussion of AI and ADM has revealed these issues with greater clarity. Rather than contemplating new regulation specific to AI to combat this issue, an appropriate response would be to issue suitable guidance around indirect discrimination. Recent work between the Institute and the Australian Human Rights Commission has started to address this situation for insurance pricing and underwriting,[1] which we note is also referenced in the Discussion Paper. This guidance shows how existing regulation can be clarified for AI practitioners. The Institute would be willing to assist with similar work, drawing on our experience to date with the AHRC and the range of fields in which the profession practices.

In areas where we identify risks associated with AI where no regulation appears to exist, or where regulation is deemed inadequate, this should be addressed. A risk-based approach to designing such regulation is possible – we make suggestions below on what this might look like. Most importantly, in such cases we suggest that regulation should still aim to be technology-neutral, in order that this gap in regulation is properly filled.

In the approach taken and the interventions contemplated, the Discussion Paper primarily seems to concern itself with risks to individual consumers, or groups of them, in narrow situations. While this is important, we note there are broader risks to society associated with AI which may require different treatment. A single risk-tiering model is likely to prove ineffective in managing both categories of risk, since they clearly require different treatment.

Finally, AI creates significant opportunities for society. A risk-based approach to regulation should carefully balance the potential gains from AI with the management of risks from it. The Discussion Paper is relatively silent on how or where this balance ought to be struck, but a risk-based approach requires it.

[1] Guidance Resource: Artificial intelligence and discrimination in insurance pricing and underwriting. Australian Human Rights Commission and Actuaries Institute, December 2022. Available at https://humanrights.gov.au/sites/default/files/guidance_resource_ai_-_2022_v7-2_0.pdf

General recommendations

In consideration of the general principles and perspectives outlined above, we make the following recommendations:

1. Regulation should primarily be outcome focused, rather than technology focused to help ensure it can be enduring/long lasting;

2. Risk-based approaches to AI regulation should:

- be based on a well-defined taxonomy of risks that AI systems may introduce or exacerbate;
- incorporate a well-defined menu of risk-management options that could be imposed by regulation;
- ensure the costs of risk-based regulatory interventions are justified by the risk-reduction created, without obvious gaps or overreach;
- carefully target risk-management interventions to the risks identified for each situation considered, rather than bluntly applying the same interventions across a broad, vaguely defined risk category as proposed in the Discussion Paper;

3. Producing guidance on existing regulation should be prioritised over creating new regulation, in situations where such regulation already exists;

4. A centralised expert body should be created and appropriately funded, to provide assistance to primary regulators in considering AI governance, regulation and guidance.

Each of these recommendations is discussed in detail in the Attachment.

Further feedback and discussion

We welcome any opportunity to discuss this important topic with you in more depth or provide further information. If you would like to do so, please contact Elayne Grace, Chief Executive Officer of the Actuaries Institute, on (02) 9239 6100 or executive@actuaries.asn.au.

Yours sincerely

(Signed) David Whittle

Acting President

Attachment: Detailed discussion of Institute recommendations and other comments

1. Regulation should primarily be outcome-focused, not technology-focused

The content of this section is relevant to questions 1, 2, 5 and 8 posed by the Discussion Paper:

It is our strong view that, wherever possible, regulation should primarily focus on the outcomes for individuals, groups or for the community, rather than the technology (or any other mechanism) that leads to those outcomes. For example, in the context of regulating a loan decisioning system:

- A technology focused approach would require organisations to decide whether the loan decisioning system is AI or not, before determining whether the regulation applies; and
- An outcome focused approach would consider what a 'bad' loan decisioning outcome might look like, and how regulation might serve to protect against that bad outcome regardless of the underlying technology used.

Regulation that is outcome focused creates consistent treatment of similar outcomes however they might occur. We caution against taking the approach of the European Union (EU) (and now contemplated by other jurisdictions) to set specific regulation for AI/ADM, because such technology-focused regulation will result in negative consequences including the following.

- Overlaps and inconsistencies will be created between AI/ADM regulation and other regulation covering the decisions to which AI/ADM contribute. This lack of clarity increases compliance costs and increases the risk of unintentional non-compliance.
- Gaps in regulation may be created if new technologies are not captured by existing definitions, even when they pose similar risks as other regulated technologies. The recent proposed amendments to the proposed EU AI Act in response to generative AI and foundation models is indicative of the definitional challenges which will emerge as technology progresses. This should be particularly concerning given the current pace of change.
- Loopholes will be created which can then be exploited. Bad actors will have a clear incentive to cause bad outcomes via mechanisms which fall outside of the scope of technology-specific regulation.

We note that some outcome-based regulations might be far more significant in the context of AI/ADM than traditional contexts, but this is acceptable if the regulation is reasonable in the context. For example, a right to an explanation of a decision may be deemed appropriate in a particular situation, and this might be substantially more onerous for an AI system than a human one.

In some situations, there may be a need for new regulation considering opportunities created by AI/ADM. For example, facial recognition technology enables surveillance of the public by law enforcement agencies on a scale which was previously impractical. In situations like this where AI/ADM creates something genuinely new, existing regulation may be inadequate, and clarification of existing rules may be insufficient. If new regulation is required to manage novel outcomes created by AI/ADM, wherever possible the regulation should still be written in a technology-neutral manner, focused on outcomes, not made specific to AI/ADM. If it is necessary to create technology focused regulation, it should enhance and complement outcome focused regulation, rather than being a substitute for or conflicting with it.

Specific Comments on Definitions

Noting our position above, we do not consider that it is fruitful to rely on definitions of terms such as 'AI' or 'ADM' when designing regulation. Notwithstanding that, for completeness we identify some flaws in the definitions proposed, which illustrates the problems which will ensue if such definitions are relied upon in regulation.

- The definition of AI appears to be based on the ISO definition of an AI system[2] but has been amended in ways which could be problematic.
  - It adds the word 'predictive', narrowing the definition from that of ISO. This narrowing might hinder its utility if applied in regulation.
  - It adds the phrase 'without explicit programming', which creates confusion and potential loopholes. For example, while it is not always common practice AI models can generally be expressed as a (potentially very complex) mathematical formula. Hence any regulation relying on this definition could be circumvented by having a human program this formula directly into a decisioning system, which would then surely be 'explicit programming'.
  - Further, the Discussion Paper notes that 'AI is unique because it can take actions at a speed and scale that would otherwise be impossible.'[3] We suggest that this is not necessarily true according to the definitions given. If speed and scale are what sets AI apart, then a simple, explicitly programmed system with such capabilities perhaps ought also to be classed as 'AI'.
- The definition of ADM seems to be very broad and could potentially be interpreted to include almost any decision that is based on automatically produced data. This breadth is exacerbated by use of the words 'in any part'. Overly broad definitions may lack utility in imposing regulation, as they may not effectively delineate situations that require intervention from those that do not.
  - For example, under the given definition, webpages use ADM to automatically decide how to render a page based on whether a desktop or mobile device is being used. This instance of ADM is clearly far less important than (for example) ADM affecting the price of a product presented to that same customer as they shop on that website.
- To add clarity for practitioners and the public, it would be beneficial to include examples of things that would fall under critical terms such as 'AI', and equally importantly examples that would not fall under those definitions.

To reiterate the commentary above, the Institute holds the position that regulation should generally be outcome-focused rather than being specific to AI/ADM. However, if it is necessary to create AI/ADM-

[2] 3.1.4 AI system: "Engineered system that generates outputs such as content, forecasts, recommendations or decisions for a given set of human-defined objectives." ISO22989:2022 as linked in Consultation Paper: https://www.iso.org/obp/ui/#iso:std:iso-iec:22989:ed-1:v1:en

[3] On page 3

specific regulation—perhaps to complement some more general regulations—then these terms should be defined with extreme care, and with a mechanism to update the definitions quickly as technology changes or as deficiencies of the definitions become apparent.

2. Risk-based approaches to AI regulation

The content in this section is relevant to questions 2, 5, 6, 14, 15, 17 and 19 posed by the Discussion Paper.

We believe that taking a risk-based approach to AI regulation in the form described in the Discussion Paper will provide neither the required clarity nor the desired outcomes. Instead, to make a risk-based approach work, it must:

- Be based on a well-defined taxonomy of risks that AI systems may introduce or exacerbate;
- Incorporate a well-defined menu of risk-management options that could be imposed by regulation;
- Ensure the costs of risk-based regulatory interventions are justified by the risk-reduction created, without obvious gaps or overreach; and
- Carefully target risk-management interventions to the risks identified for each situation considered, rather than bluntly applying the same interventions across a broad, vaguely defined risk category as proposed in the Discussion Paper;

We outline our reasons for these recommendations in the sections below.

Risk based regulation must be based on a well-defined taxonomy of risks that AI systems may introduce or exacerbate

If it is to be effective, a risk-based approach needs to be grounded in concrete situations rather than at an abstract level as proposed in the Discussion Paper.

Crucial to this is a comprehensive framework or taxonomy of risks – a risk-based approach necessitates a careful understanding of the risks to be managed before risk mitigation is considered. This could take the form of a well-defined list of harms which we consider might be caused by AI systems, for which we might consider that some risk management (in the form of regulation) is warranted.

We do not believe this has yet been adequately completed.

Take, for instance, the 'risk management approach' described in Box 4 of the Discussion Paper. Rather than clearly considering risks worthy of management, it jumps to categorising various forms of AI systems using ambiguous language and definitions, such as 'low risk' which is described as 'minor impacts that are limited, reversible or brief'.

This description invites various questions about the nature and scope of the word 'impact'. What impacts should we be concerned with - threats to life or health, financial loss, misinformation, risks to democracy, environmental impacts, existential threats, or something else? The range of potential harms brought about by AI applications, from personal to societal levels, necessitates diverse forms of risk management and regulation. We must start by being clear about what risks we are discussing, if we are to manage them effectively.

We think that the risks of AI systems can be grouped into two broad categories:

- Risks manifesting in individual harms.

  For individuals, poor outcomes from AI could include matters as diverse as personal injury, financial loss, being misled, being treated unfairly, or discrimination. Often such matters are already considered by regulation, which might already apply or be adaptable to AI.

- Risks that manifest at a broader societal scale.

  These include challenges such as the impact of widespread misinformation or disinformation on our collective knowledge, impacts to democracy, impacts on specific segments of the workforce, or even existential threats. These broad-scale risks clearly require a different set of risk-management strategies and regulatory responses.

Given the vastly different nature and scales of these categories of risk, a single 'risk framework' is likely to prove ineffective in managing them. Even within each category, we suggest a single risk-tiering system such-as that proposed will struggle to be effective, as we outline below.

Existing international proposals only partially outline the risks in question. Many appear to ignore the second category entirely. Some examples in the Discussion Paper are more mature – for example, the Canadian Directive sets out five general themes within its Appendix B[4]. However, to our knowledge none have created a sufficiently robust and detailed risk taxonomy that would allow for proper risk management to occur. Creation and agreement of this taxonomy should be the first step.

Risk-based regulation must incorporate a well-defined menu of risk-management options that could be imposed by regulation

Once a robust risk taxonomy is in place, we can then consider appropriate risk management strategies and options for those risks.

This will be most effective if applied in narrower contexts, rather than in an abstract, high-level manner - as we discuss in the section below. The approach of the Discussion Paper, perhaps influenced by the wider global discussion around AI regulation which we believe has prematurely shifted to 'what to do', risks leading to inconsistent and ineffective application.

Attachment C of the Discussion Paper represents an early version of what is required for a menu of options, but much more depth and breadth is required to make this effective. To improve this and make it fit for purpose, it should:

- Be a complete and exhaustive list of the interventions proposed, explained in detail;
- Consider the specific risks from the risk taxonomy that each intervention aims to address;

[4] See https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592

- Identify the potential costs and limitations of the intervention; and
- Ensure all risks from the taxonomy have appropriate intervention(s) that might be considered.

We should ensure the costs of risk-based regulation are justified by the risk-reduction created, without obvious gaps or overreach

We suggest utilising the risk taxonomy and menu of interventions directly when considering risk management activities like regulation or standards, in a specific context.

Instead, a blunt, aggregate approach is proposed by Box 4 which we believe will be ineffective. The interventions are imposed in a blanket manner to all things falling into a certain 'risk-level'. There is no discussion of the need for, nor effectiveness of, those interventions across each of the diverse AI contexts for which they are proposed.

This approach creates two poor outcomes:

- Insufficient or ineffective intervention.

  There will be situations where AI systems are classified into a tier, but if asked to contemplate the specifics, we would choose to impose stronger interventions due to the risks in question of that particular AI system falling outside of our collective risk appetite, even after intervention.

- Waste.

  There will be interventions imposed on AI systems which serve no purpose, because the risks they serve to manage are not present for that particular AI system. Similarly, but less severe, some interventions will impose excessive costs for only modest benefits, in that specific context.

To illustrate, risks to life or health due to an AI system error should be listed in a risk taxonomy and clearly worthy of management. Box 4 suggests that 'Use of AI in safety-related car components and in self-driving cars to make real-time decisions' is higher risk than 'Use of AI-enabled chatbots to direct citizens to essential or emergency services', warranting a range of more intensive and expensive interventions. We assume this is because of risks to life or health, though there may be other risks as well. But is this classification – and intervention – appropriate? Does it represent an appropriate trade-off between risk and reward?

- First, regulation and standards already exist to manage risks in vehicles. Perhaps no further intervention is required. Imposing interventions that are not required is wasteful and does not represent good risk management.
  - Before imposing additional standards, we suggest an appropriate action might be to ask the vehicle safety standards authority to consider whether its standards appropriately respond to the risks identified, and to augment them if required.
  - We should also consider whether interventions from the menu are viable. For example, humans might not be able to be 'put into the loop' of an autonomous braking system, which exists to operate without human input. But the proposed risk-tiering system seems to require this.
- On the other hand, for the chatbot it might be less obvious whether an existing regulator or standards body exists. In this case, we could carefully consider whether the interventions proposed by the risk-tiering in Box 4:

- Reduce the risks to tolerable levels, and
- Achieve this at an appropriate cost, without unreasonably negating the potential benefits of the AI system.

- Overall, we expect an analysis of this form might conclude two things:
  - Arbitrarily classifying a chatbot as 'lower risk' than an autonomous vehicle, with less onerous interventions applied, may be inappropriate
  - Some of the interventions proposed by the aggregate categories may not represent a good cost/benefit – or risk/reward – trade-off, for the particular situation at hand.

To solve some of these issues, below we make an alternative proposal to aggregate risk-tiering.

Risk-based regulation must be carefully targeted to the risks identified for each situation considered, rather than bluntly applying the same interventions across a broad, vaguely defined risk category as proposed in the Discussion Paper.

There are several reasons why a blunt, aggregate, 'tiered' approach to risk-based regulation of AI of the form proposed in the Discussion Paper is inappropriate.

- Reason 1: Inconsistency and ambiguity in classification terms.

  Terms like 'low', 'high', 'significant', 'material' and 'difficult' are not merely used as helpful descriptors but rather are used to *define* risk levels. These terms are subjective and open to interpretation, leading to inconsistencies in application if they are used as definitions. While it is common in risk management to use such terms as descriptors in general communication, it is good practice to ensure that the actual underlying definitions are more precise, to ensure consistent application.

- Reason 2: Oversimplification of a multidimensional problem.

  Risk categories usually simplify a complex, multidimensional problem into a single dimension, causing a lack of clarity unless carefully defined. For instance, an AI system with a very high impact (Box 4 – 'high risk') but easily reversible outcome (Box 4 – 'low risk') isn't obviously positioned in the proposed classification system. Again, while simplifications of this form are common practice in risk management for communication, their use as definitions of the risk categories is likely to cause challenges and inconsistencies.

- Reason 3: Difficulty in categorising situations with multiple impacts.

  AI use cases often involve multiple forms of risk, making a single risk classification challenging to apply, and perhaps ineffective at managing risks. For example, an AI system used in hiring (Box 4 – 'medium risk') could involve various risks, like unfairness, discrimination, or privacy concerns. Some of these may be considered more significant than others in some situations, requiring more intensive risk management. Some of these risks might not be present at all, in other situations. Grouping all risks together at the AI-system level presumes similarity of significance of all risks imposed by an AI system. This will lead to poor outcomes, both insufficient intervention and waste are to be expected.

- Reason 4: Lack of justification for classifications of example use-cases.

  Without detailed explanations for why certain use cases are classified as they are, professionals, such as actuaries, might struggle to confidently classify AI systems. Comprehensive guidelines are needed to generalize use case classifications effectively.

- Reason 5: Insufficient justification for interventions.

  Interventions listed as risk-reduction measures are often not well justified. Their effectiveness at reducing risk will vary depending on the specific situation, and better alternatives might exist. As we note above, it's critical to thoroughly assess and justify each intervention in the context of its application, so that we are managing risk to the extent we require – not too much causing waste, and not too little causing gaps.

- Reason 6: Risks manifesting at a broader societal scale require different interventions.

  As we outlined above, certain risks associated with AI operate at a broader societal scale, while others operate at the level of individuals. The interventions required to manage such risks will likely be very different to those aiming to prevent individual harms. Use of a single risk tiering model does not allow such variation in risk-treatment.

In contrast, we propose the following high-level process that is more focused and risk-based than the current AI regulatory proposals:

1. Formulate a risk taxonomy that focuses on risks imposed by AI systems.
2. Formulate a menu of interventions that might serve to manage those risks.
3. For a specific AI system under consideration, identify relevant risks from the taxonomy.
4. For the risks identified, examine whether existing regulation can manage the risk appropriately.
    a. If the answer is yes, ensure the regulation is suitable, understood by AI developers and operators, and reduces the risk to acceptable levels at an appropriate cost.
    b. If not, consider whether the risk needs intervention from the menu. If so, contemplate creating new regulations or assigning an existing regulator to oversee this additional risk management intervention in this context.

This process results in greater focus on gaps and areas of concern compared to the broad-brush approach suggested by the Discussion Paper. It also allows us to intervene more strongly and change our level of intervention in areas deemed necessary, without having to adapt an entire framework and then impact other AI systems – it will ultimately be more nimble. To apply it would require some centralised expertise, particularly in directing the activities of Step 4.

An additional benefit of the approach is that it can adapt easily to different levels of risk tolerance. For example, it is common for there to be less tolerance for certain risks to individuals when delivering government services compared to equivalent private market services, notably as the monopoly nature of government service does not have the natural protections of a competitive marketplace (if an individual is unhappy with services, they can go elsewhere). Our framework easily accommodates such variation in risk tolerances.

3. Prioritisation of guidance on existing regulation

The content in this section is relevant to questions 2 and 4 posed by the Discussion Paper.

We consider that the immediate regulatory action required to allow greater safety and adoption of AI/ADM is to clarify the operation of existing regulation. This follows naturally from the discussion above. For many situations, an analysis of risk as we propose above will conclude that existing regulation should be able to manage the risk suitably, if properly understood and applied.

The reason for this is that Australia often adopts a principles-based approach to regulation, particularly in areas where there might be heightened risks to individuals. This has the advantage of being able to cover a wide range of industries and situations. However, it also means that its application to specific applications is often unclear, including but not limited to situations related to the use of AI/ADM.

As an example, recent debate about the use of facial recognition technology by retailers is at least in part a debate about how to interpret aspects of Australia's Privacy Act. This debate has arisen due to a lack of clear guidance about how the Privacy Act should be interpreted in this context. Where guidance is lacking, and uncertainty exists, it should be unsurprising that some might take a more liberal interpretation of what regulation might allow than others. We note that such uncertainty is often avoidable—in this situation guidance on facial recognition systems could have been issued substantially ahead of any implementation within retail settings.

Additionally, a lack of clarity within existing regulation can be a barrier to innovation—particularly in sectors like financial services where the costs of non-compliance can be high. Unclear regulation is not good regulation, and this carries costs. A lack of certainty over the law can—and does—result in the abandoning of projects that may be of value to both the community and to industry, for fear of breaching that unclear regulation.

While case law and regulatory guides often exist to provide some clarity over the interpretation of high-level regulatory principles, these are often backwards looking, written after problems have already emerged, or are written with traditional, human-centric decisioning systems in mind. Hence many existing regulations are likely to be unclear in the context of AI/ADM.

By nature, automated systems involve data, numbers, and mathematics. This includes systems involving unstructured data which humans may not intuitively recognise as being mathematical, but which an AI/ADM system must encode into a mathematical form to utilise. This essential role of mathematics in AI/ADM means a level of precision is imposed on the AI/ADM decision process which human decision processes may not traditionally contain. This means that high level 'principles-based' regulation, written in words, may not always be precise enough to authoritatively guide the design of such systems.

It is essential that guidance which is intended to be applied in a mathematical context such as AI/ADM avoids such uncertainty. This does not necessarily mean that guidance needs to itself contain mathematics (though this may be appropriate in some situations), but at the least the guidance given should not lead to ambiguity when translated into the mathematics of an AI/ADM system. Use of greater

precision in guidance may have the added benefit of providing greater clarity on the expected outcomes of human decision systems.

For example, the concept of fairness is embedded in many principles-based regulations around the world today. An example in Australia is the requirement under the *Corporations Act 2001 (Cth)* for financial services companies 'to do all things necessary to ensure that the financial services covered by the licence are provided efficiently, honestly and fairly'[5]. But what precisely is intended by a requirement of 'fairness'? Recent academic studies of 'fairness' in particular instances of AI/ADM have yielded important insights, most notably the incompatibility of intuitively reasonable mathematical definitions of 'fairness'[6]. This means a general requirement to act fairly is not a specific enough instruction to encode into an AI/ADM system. If we must mathematically define what terms like 'fairness' mean for our decisions, as we must for AI/ADM, we require far more granular, specific guidance as to the correct interpretation of such words in particular situations.

Therefore, any regulatory guidance must be authoritative and should be specific and detailed enough to guide the design and operation of AI/ADM. We suggest a starting point should be situations where AI/ADM is likely to be introduced or is already present, and where the decision being made can potentially cause harm to consumers. This guidance must allow us to confidently translate the words in regulation into the mathematical instructions required for AI/ADM to be constructed and operated. Simply: the law must be made clear.

The Commonwealth can act to create regulatory clarity around AI/ADM by instructing all relevant regulators under their authority to undertake reviews of existing regulations in the context of AI/ADM, and to issue guidance as needed to create that clarity. This is aligned to Step 4(a) in our proposed risk management methodology above. Regulators could also seek to conduct such activity without Commonwealth instruction—including regulators operating at other levels of jurisdiction such as at State level.

We suggest that industry and professional bodies will be willing and able to assist in identifying specific areas where greater clarity is needed and will be able to make suggestions as to what clarity might mean in those contexts, in response to any further consultations. An example of such a collaborative approach is the project jointly undertaken by the Institute and the Australian Human Rights Commission

[5] s 912A (1)(a)

[6] For an early example of this vast literature on fairness in AI systems, see Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:1609.05807.

to produce guidance on AI and discrimination in insurance pricing.[7] We suggest the level of detail and practical illustrations contained in this guidance highlights what is still lacking in many other contexts. Even just focusing on discrimination law alone, similar guidance could be developed for AI in dozens of other high-stakes contexts.

4. Centralised expert body

The content in this section is relevant to questions 3, 4, 5 and 7 posed by the Discussion Paper.

Consistent with, and to help support the recommendations elsewhere in this submission, the Institute recommends the creation of a centralised expert body for AI with three primary responsibilities.

1. To conduct risk analysis according to the framework we identified above, importantly including directing risk management activity to other agencies of Government (i.e. steps 4(a) and (b) of our framework). This activity should be centrally conducted in order for it to be consistent and to leverage scarce expertise most effectively.

2. To provide AI expertise to primary regulators to assist them in drafting guidance for AI/ADM that interpret existing regulation, in line with our recommendations above. A centralised approach is likely to find efficiencies across the various domains encountered and may be more cost-effective than asking each primary regulator to attract and retain staff with expertise in AI regulation and interpretation.

3. To conduct broad 'horizon scanning' to ensure AI regulation and the risk analysis supporting it stays up to date. This will allow more proactive response to emerging issues than we have witnessed to date – essential given the current pace of change in AI.

5. Other Comments

The content in this section is relevant to questions 9, 10 and 11 posed by the Discussion Paper.

Enhanced AI literacy

We believe it is essential to increase AI literacy across all segments of the population, including both the general public and professionals that use AI every day. AI should not be a mystery. A set of online learning modules could be curated to explain AI in simple terms and debunk common myths, serving

[7] Guidance Resource: Artificial intelligence and discrimination in insurance pricing and underwriting. Australian Human Rights Commission and Actuaries Institute, December 2022. Available at https://humanrights.gov.au/sites/default/files/guidance_resource_ai_-_2022_v7-2_0.pdf

both school students and the general public. Engaging domain experts and good teachers can help shift public attitudes to AI from a fear-based attitude to one of curiosity and acceptance.

Public trust in AI deployment can be increased by using live AI deployment examples to illustrate the mechanisms that are in place to protect individuals' data, to ensure appropriate decisions are being made, and to manage risks appropriately. For example, an infographic might be developed to describe how algorithms are used to recommend items for your shopping cart when buying groceries online, and the multiple controls in place within such a system to protect consumer data and privacy.

Transparency

Consistent with our other recommendations, we should consider mechanisms for creating transparency as potential risk management interventions, to be assessed and justified in specific contexts. This should be considered within the menu of options available.

Transparency to consumers should only be required where it genuinely adds value – particularly where the form of transparency is mandated. For example, mandatory privacy notice and disclosure regimes such as those required under the EU's General Data Protection Regulation (GDPR) are often seen as an irritant rather than something that consumers actually value. We believe a similar form of 'pop-up'-style notification for AI systems would have similar issues.

If transparency of a decision is of value, then it will usually be of value regardless of the technology used to make that decision. For example, we might decide that individuals should have the right to an explanation about how their bank loan decision was made. If so, this right should apply whether their bank loan decision was based on an AI/ADM or if a human bank manager made that decision based on their expertise.

In terms of what should be made transparent, consideration should be given to a range of factors such as data inputs to a model (acquisition, collection, storage, maintenance and application), human intervention in the decision-making process, and findings from regular self or independent reviews. Different forms of transparency will serve to manage different risks. The aims should be risk-reduction, and the framework we outlined above should be able to accommodate such considerations.

Banning certain applications

Following our previous argument that regulation should be outcome focused, bans on high-risk applications should also be outcome rather than technology focused.

Banning an outcome only when AI was used to create it creates perverse incentives, where individuals can legally commit bad outcomes by using mechanisms other than those defined as being 'AI'. This also risks being unable to adapt to unforeseen mechanisms to achieve those bad outcomes, particularly as technologies evolve.

That said, we do not discount the possibility that there may be a place for specific prohibitions of AI or ADM, but we suggest these should be exceptions, not the rule.

We also observe that from a risk-management perspective, a ban is not the most extreme intervention one could take to manage a risk. More costly measures than bans can be imposed. For example, we might proactively monitor compliance with a ban, or to make the action itself more difficult for an individual. An analogy can be drawn from road safety: banning the use of mobile phones while driving was insufficient to reduce the risk sufficiently, so this has been supplemented in recent years by road safety cameras for enforcement, public awareness campaigns, and (in some cases) options within mobile devices to restrict usage while driving.