Supporting Responsible AI: discussion paper

Prepared by : Dr Michael G. Kollo, CEO Evolved Reasoning

This paper has been prepared in response for a consultation on how the Australian Government can mitigate any potential risks of AI and support safe and responsible AI practices.

This is significant topic, and having attended the in-person session at the Sydney CSIRO offices, I thought it would be more pragmatic and useful to provide direct feedback from points that were raised in that session, as well as others that maybe were not covered.

Firstly, I will limit my comments to those areas where I feel the government has a direct role to play through regulation. Standards bodies, both general and industry-specific, will increasingly create a multitude of ideas-based standards, that companies and industries will adopt based on their business-models, affiliations, and leadership. This is certainly the direction in the US regulatory landscape, as well as Europe for medium and low level risk systems.

Broadly speaking, I think the role of government is to promote industry development and economic growth while staying within the existing laws of consumer protection and data privacy.

AI is both a tool, built from data and algorithmic components, and an enablement of certain activities (both good and bad) in the digital space. I think its useful to separate the two for the purpose of this consultation.

As a tool, algorithms are built to solve a problem (a forecast, or a decision) generally in two ways:

1. For company specific use cases. Examples tend to be profiling customers for products, recommendation system, forecasting repair and maintenance for specific manufacturing components and even identification systems. These are non-transferable models usually, and are specific to the operations and datasets of that company. Their reach is also contained to that of the company’s clients and use cases. There is more limited scope for damage here, unless the company is large and has internal scale. (see notes on scale later)
2. ‘Common systems’ that are shared by business use cases, like language and health outcomes. These are built by using big cross-societal datasets, for broad usage cases, and are sold broadly and globally. Examples are language models, and healthcare models, that are generated through enormous investment and intended scale. These are transferable models, and they can be sold and re-sold to a variety of companies embedded in to products. An example can be an open sourced language model developed in the US used to train an industry specific language model in South Africa, that is then used within a product to teach healthcare to customers in Eastern Europe. Their transferability makes it difficult to identify these models for bias in the place that they are created or transferred, but rather they should be identified in the place that they are used.

This is a huge topic, I am sure that you can appreciate that by the number of submissions, and mapping the entire AI landscape is not my intention here.

However, I would like to point to two specific components that I believe are the central ‘nexus’ points of dangerous AI, and that are related to the topic of regulation specifically.

1. The intention to monitor individuals and/or denial of service or product.

These are the use cases (not characteristics) that companies embark on as a natural course of their business. For example, CVs are today screened by off-shore companies to ensure they are appropriate before proceeding to the hiring manager. Those that are not allowed to proceed are examples of ‘denial of service’ in that the process disqualifies them. If done by humans, we expect that companies provide proper guidance for people that are reviewing CVs to ensure that the decision is unbiased as to what kinds of skills and characteristics they are looking for.

If done through an algorithm, we would expect the same bias free operation of the algorithm because its ‘decision to deny’ further progress is a significant source of potential discrimination and economic impact in our society. The real issue emerges that the decision process is considered to be ‘outsourced’ to an algorithm, and that business is now no longer able to represent or understand the drivers of the decisions. The ‘ethical testing’ of the algorithm therefore needs to align with the business values, and ultimately the education that is required for the business to understand the algorithm itself.

The point of the regulation is therefore to ensure that any services that are using algorithms are consistent with the business’s code of conduct and business ethics. In this sense, it’s a recognition that an algorithm is an extension of a business’s operations and their values, rather than something removed from it.

.

1. The utilised scale of an AI algorithm as a condition of its perceived danger.

Intelligent algorithms, like AI, are perceived to be impactful because of the enormous scale that they can create. The scale is in decision making: a single algorithm can make decisions for hundreds of thousands, or millions of credit applicants, much more than a single human could. In a human-decision world, individual humans can be biased, but a large collection of humans making decisions (like the credit officers of a bank) will result in a normally distributed set of biases, with a small (hopefully) set of decisions living on the left hand side of the distribution, and accounting for ‘bad’ or ‘biased’ decisions. This assumes that these errors in judgement happen randomly, and they are not systematically trained into the human decision makers. It also means that any bad outcomes that do happen are not expected to happen again: asking for a credit loan from different humans should lead to a ‘bias free’ outcome, as the individual errors do not repeat.

With AI, a single decision algorithm with a small error or bias will repeat again and again in the same way (for prescriptive algorithms). This is equivalent to handing a million credit decisions to a single human; something that our democratic values would be against.

The challenge here isn’t the fact that there is some bias in the decision making of algorithms, that will always be the case. The problem here is one of scale that allows a single algorithm to reach so far across our society. As an example, we could regulate all algorithms that are making decisions over 50,000 individuals a year. This would allow smaller algorithms to operate through principles, and the larger algorithms to operate through stricter regulations in line with their societal reach and impact.

1. The limits of Explainability of AI

Earlier generations of AI systems could generate ‘fundamental’ explanations of core drivers of decisions. If using a cohort of variable, these systems could point to specific variables that are were more or less important than others. These are more simple and linear in nature, and usually fit into human logic and intuition. For example: the primary drivers of credit decisions are someone’s asset base, and their current salary.

The next generation of AI models, neural networks, became increasingly difficult to explain, because they were highly contextual and non-linear. Now, the explanation started to be a bit trickier: For person A, the drivers are primarily around their assets and liabilities because they are self-employed, whereas for person B, the drivers are primarily their salary (because they are employed by a big bank) for example. This means that two decision are not directly comparable in drivers, though both are ‘fair’ within their own cohorts.

The current Generative AI models, like image recognition models, are much harder to explain from first principles. No single token or pixel is enough to give an explanation, and even if it were, it would be too far removed form human understanding. The testing of these systems will need to be ‘top down’ through simulation and experienced use. Eg we tested this system on 50,000 applicants, and we found that it was primarily driven by X, or Y. This requires a different kind of framework to the above systems, and should be considered if an explanation is required or mandated.

Thank you for reading our submission. Happy to expand on any of the above point. I think a broad mental framework is required, but can be borrowed from multiple sources, and then retro-fitted to Australia’s unique industry and economy.

My primary concern is that in our efforts to protect consumers and society, we under-invest and therefore under-utilise critical AI tools, that lease us more exposed (economically) and therefore in a worst position than any individual bias could have.

Dr Michael G. Kollo

CEO Evolved Reasoning

[michaelkollo@evolvedreasoning.ai](mailto:michaelkollo@evolvedreasoning.ai)