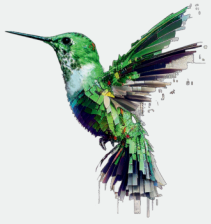
# MANDATORY GUARDRAILS

Submission to Australian Government

STIRLING & ROSE



october 2024

STIRLING & ROSE

Director
Technology Strategy Branch
Department of Industry, Science and Resources
Industry House
Canberra ACT 2601

4 October 2024

Dear Department of Industry, Science and Resources,

**Consultation on introducing mandatory guardrails for AI in high-risk settings**

Stirling & Rose (**S&R**) welcomes the Department's consultation on the proposal paper for introducing mandatory guardrails for AI in high-risk settings (**Consultation**). S&R continues to support the Department's commitment to Safe and Responsible AI.

S&R is a boutique emerging technology law firm and global legal policy institute with deep experience in the subject matter of artificial intelligence (AI), digital assets, smart legal contracts and autonomous organisations (including their sub-sets of AI organisations and decentralised autonomous organisations). We advise organisations seeking to use AI responsibly, major investors (hedge funds, VC funds, and lenders) and entrepreneurial organisations operating at the vanguard of emerging technology on major real-world application and enforcement of all the above. We have done so at an international level as the executive team of a major international law firm since the inception of the space. Our partners have backgrounds in international regulatory law and disputes, insolvency and personal property law, asset and structured financing, ICT contracts, and machine learning. We collaborate with the Gradient Institute, Australia's pre-eminent responsible AI institute to offer legal acumen and technical expertise to our clients in comprehensive AI Strategy Services.

We have been speakers at the AI Regulators Symposium held 21 May 2024 at the University of Technology (**UTS**), Sydney, the Generative AI Summit held at UTS 15 August 2024, and the UTS Data and Digital Week held 23-27 September 2024. We have also been invited to speak at the 27th ASIC Annual Forum to be held in Sydney on 14-15 November 2024. S&R facilitated submission writing workshops in Sydney (27 October 2024) and Perth (25 October 2024) for organisations seeking to participate in the Consultation. We are invited speakers on AI at events around the world, including Singapore, the Netherlands, Austria, Indonesia and France.

We thank the Department for its work in this Consultation.

Yours sincerely,

**STIRLING & ROSE**

# 1 Executive Summary – Prepare Now for the AI Future

## 1.1 The future – AI in everything

Stirling & Rose urges the Government to prioritise robust public discussion on AI regulation and Lawful AI Critical Infrastructure[1] which foreshadows the capabilities, market/societal implications and risks of AI as it will be in 10 - 20 years, rather than as it is now or in the near term.

Recent discussions on how to regulate AI, including the discussions occurring as a result of the current Consultation, are predominantly focused on the current state of AI. Just as a good AFL player anticipates the future moves on the field and positions themselves ahead of the play, we urge Australian policymakers and regulators to prepare for where AI is heading, including moving to develop Lawful AI Critical Infrastructure.

The current AI tools and platforms are controllable, identifiable and alienable as an AI tool or platform – for example ChatGPT is an AI platform. AI as an alienable "other" will not be the case for long. In our thought leadership, advisory and policy work, our experience strongly suggests that to "play the movie forward" is a future state ***where AI is embedded in everything – including us, our institutions, our consumer goods – every sector and every industry – from our fridge to our cars, our money and our contracts.*** This might take five years, this might take 15 years, it is however, inevitable. In this postulated future, regulation of AI as a technology is likely to apply to regulate everything – every product, every service. We also postulate that Lawful AI Critical Infrastructure will be key to respecting the rule of law in an AI-integrated future.

Good law is technology agnostic. Law "parametrised" by the current shape of a specific technology, particularly a rapidly transforming technology such as AI, will be quickly found to be not fit for purpose. Any AI regulation must take into account the coming inability to alienate an AI tool from an underlying asset, entity, or even biological unit.

## 1.2 Responsible AI Problem

To cast forward, Australia should be preparing for a future where increasingly capable, agentic AIs require diminishing human control and oversight over time and are ultimately capable of autonomously operating enterprises which deliver economic value and in which humans invest as shareholders. These AIs will be capable of autonomous evolution, self-replication and creating or "digitally birthing" and nurturing new, more specialised AIs optimised for specific tasks or goals.

This raises the responsible AI problem, that all existing legislation is predicated on regulatory non-compliance being ultimately sheeted back to a responsible human person. By comparison, AI is not responsive to the human-centric sanctions underpinning traditional regulatory structures such as shame, disqualification, state-sanctioned violence and imprisonment.

[1] Lawful AI Critical Infrastructure refers to a future digital system which enables authentication, regulation and governance of AI through smart legal contracts

We postulate new measures will be required to combat the responsible AI problem, such as the granting of legal personality to AIs through a new legal entity, the AI Organisation (**AIO**). However, the granting of such legal status should be conditioned upon satisfaction of certain requirements e.g. registration with appropriate agencies, unique identifier as an AIO, evidence of de minimis functionality, effective sanctions for aberrant behaviour, economic reserves, adjustments for intent.

Going forward, another mitigating tool to also help address the responsible AI problem is the smart legal contract[2]. Smart legal contracts hosted on secure (likely critical) infrastructure can act as an AIO governance tool with greater ability to moderate the complexity of risk allocation, rules of engagement, obligations and remedies between differing members/shareholders and regulatory bodies in respect of the AIO and its individual assets.

The increasing capability and multifaceted nature of AI demands an adaptable, future-focused policy response and regulatory framework.

The future of AI in Australia hinges on our collective resolve to forge a path that is both progressive and conscientious, reflecting a commitment to civil society, progress and ethical stewardship in the age of intelligence machines.

### 1.3 Responsible Data underpins Responsible AI

Responsible and authenticated data is a precondition to responsible AI. Data is the indispensable resource underpinning an AI-driven economy. Accurate, fair, lawful AI needs vast amounts of high-quality data, with trusted provenance and clear legal attribution. Without data, AI cannot be developed. Without fresh, current data, existing AI applications fall into the error and obsolescence of model drift. Without contextually relevant, representative data, AI struggles to generalise from training data to real-world settings, resulting in errors of accuracy, failures of fairness and reinforcement of harmful biases and discrimination. Consideration of the legal requirements and frameworks for data is time-critical as data available for AI training becomes increasingly scarce.

Discussions on AI regulation predominantly focus upon the EU AI Act. However, we consider that the European Data Act is an important, though frequently overlooked, touchstone in advancing AI regulation. Regulating and standard setting for AI is a complex and dynamic endeavour. By comparison, the understanding of data is mature, although data value is not yet fully realised.

The recognition of data as a strategic asset should be a cornerstone of AI policy in Australia. Lawful data collection, control, authentication and access are core requirements in the regulatory shepherding of responsible AI outcomes. Regulatory planning and development of infrastructure around public and private data rights is foundational to ensuring that lawful collection and access to data is protected and monopoly-style aggregations of data are kept in check.

---

[2] A smart legal contract is a legally binding agreement that is digital in nature and contains some self-executing or self-enforcing elements.

### 1.4 Lawful AI Critical Infrastructure

Our Australian economy is increasingly digital and real-time. AI will accelerate this. Solutions to manage the digital nature and speed of an economy where AI exists at mass scale will be necessary for legal persons to continue to exercise legal rights and access legal remedies. We consider that Australia, like all sovereign nations, will need to invest in smart legal contract infrastructure to support the collection of lawful, authentic, quality data and the assetisation of the data (such as contracts) as second-generation digital assets.

This infrastructure will need to be developed with an eye to data as sovereign wealth and the ability to govern and effectively sanction increasingly capable and autonomous AI over the coming years.

We strongly recommend that Australia consider the creation and resourcing of a Lawful AI Critical Infrastructure Advisory Committee that works closely with or is a committee governed by the ACCC. We recommend Australia establish smart legal contract infrastructure to support the authentication, digitisation and assetisation of lawful data and legal agreements as second-generation digital assets.

### 1.5 Cautious and Prudent Domain Specific Approach

In the immediate term, Stirling & Rose recommends a cautious and prudent approach to AI regulation focused on leveraging existing laws and structures and making necessary changes to address concrete (rather than speculative) aspects of harm or negative impact. We consider that Australia should observe, evaluate and participate in global review of regulatory and similar initiatives (e.g. standard setting) implementation in other jurisdictions to ascertain whether such approaches might be suitable for adoption in Australia. At this juncture, we consider there will be substantive innovation, policy and legal debt associated with Australia implementing a broad, economy-wide approach in this highly complex and rapidly changing space.

### 1.6 Australia's place in global AI regulation

In the global context of AI regulation, Australia's position having regard to its size and GDP, is one of a participant rather than a trendsetter. Given that Australian entities are or anticipate exporting their AI products or services, alignment or harmonisation with international regulatory frameworks and limiting additional compliance burdens as far as reasonably possible is important to ensure Australian organisations can operate effectively and competitively in a global AI market.

### 1.7 Definition of AI

Against all of these complexities, there is also the persistent issue that the ontological delineation of AI remains a subject of both profound scholarly contention and practical contention as demonstrated in the submission writing workshops facilitated by Stirling & Rose in Sydney and Perth in September 2024.[3]

[3] We expect the terms "AI System", "AI Model: and "Agentic AI" to continually metamorphosis in response to technological advancements and the burgeoning acceptance of erstwhile AI frontiers as part of everyday life.

## 2 Recommendations

We recommend the Government coordinate and work with relevant state and federal agency stakeholders to:

(a) **Implement a domain specific approach to AI Regulation with optionality to consider framework legislation.** Review all Federal and State laws to understand how the adoption of AI and autonomous algorithms will impact the application, interpretation and enforcement of those laws in the short, medium and long term. This review should be undertaken with a domain specific regulatory lens, with optionality to explore framework legislation informed by the outcomes of the review, where it may reduce regulatory burden for organisations operating over multiple sectors and harmonise certain regulatory components between Commonwealth and potentially State laws and between different Acts of Parliament;

(b) **Use existing mechanisms.** Identify and leverage existing governance mechanisms and legal protections already in place that can mitigate the potential risks and harms of AI in a technologically neutral way. There are existing regulatory powers in the *Privacy Act 1988* (Cth), *Competition and Consumer Act 2010* (Cth), *Copyright Act 1968* (Cth), *Corporations Law 2001* (Cth) (including directors' duties and accessorial liability), prudential standards and Discrimination Acts[4] capable of responding in varying degrees to the issues of data scraping, data matching, harmful algorithmic discrimination and lack of explainability; and

(c) **Encourage Government Departments to voluntarily implement the proposed Guardrails.** This enables the mandatory guardrails to be tested in an implementation scenario to gather data on their cost, practicability and effectiveness to further inform regulatory considerations.

(d) **Monitor and Assess Regulatory Outcomes in other jurisdictions.** The point of this is for Australia to learn from other jurisdictions and gauge how the different approaches work/don't work, as well as how they need to interact with global standards.

(e) **Participate in International Forums.** Continue to take an active role in representing Australia in international policy forums and standard-setting contexts.

(f) **Intensify AI Education Efforts.** Accelerate AI education to prepare the workforce for future jobs, maintain global competitiveness and raise awareness on what actions may be required to use AI responsibly. Educating Australian citizens will enable them to engage meaningfully in public discussions, shaping policies and laws that address concerns, fostering inclusive growth and advancing national interests.

(g) **Establish a Lawful AI Critical Infrastructure Advisory Board.** Establish a Lawful

[4] i.e. *Age Discrimination Act 2004* (Cth), *Disability Discrimination Act 1992* (Cth), *Racial Discrimination Act 1975* (Cth) and the *Sex Discrimination Act 1984* (Cth).

AI Critical Infrastructure Advisory Board to consider and report back on the critical infrastructure required to govern and regulate AI Organisations, authenticate lawful data and realise the value of Australian data as sovereign wealth and support the digitisation and assetisation of legal agreements as second-generation digital assets. We support the ALC Safe and Responsible AI in Australia submission at para 134:

*"Growth in the use of AI in the operation of infrastructure is likely to be significant in the coming years. This is an area that requires particularly wide-reaching safeguards, especially in relation to critical infrastructure. Current legislation in relation to critical infrastructure does not, in the Law Council's view, sufficiently address the operation of AI".*

## 3 AI in High-Risk Settings

### 3.1 General

S&R supports a technology-neutral approach to AI regulation.

We expect what is considered high-risk AI to continually metamorphosise in response to technological advancement, increasing AI accuracy and the burgeoning acceptance of erstwhile AI frontiers as part of everyday life. We consider the principles and list-based definition in the Proposal Paper to be helpful touchstones in implementing AI responsibly rather than to found regulatory capture of AI. Nevertheless, we offer the following insights on the two options proposed to identify AI subject to regulatory capture as high risk.

### 3.2 Proposed Principles

The proposed principles provide significant flexibility in determining whether or not an AI use is a high-risk application. While flexibility can accommodate dynamic changes, it may also present the drawback of insufficient clarity and certainty for organisations attempting to determine whether they fall within the scope of the proposed mandatory guardrails. Our experience, including in facilitating workshops where participants were asked to determine whether an AI was high-risk based on the proposed principles, is that application of the principles is a challenging task when applied to real-world cases.

A large, well-resourced organisation may be better placed to determine whether it must comply with the mandatory guardrails or to navigate around regulatory capture by applying the principles in a way that aligns with their interests. By contrast, small and medium enterprises are unlikely to be able to apply the necessary resources to this task. For example, application of the proposed principles requires an understanding of fundamental human rights and related considerations. While understanding of these elements are essential for safe and responsible application of AI in high-risk cases, organisations operating in low-risk environments may lack the necessary knowledge to accurately assess and confirm their own low-risk status. The lack of resources within small and medium enterprises, particularly where the AI use is at the borderline of high-risk classification, resulting in a chilling of AI use and innovation within this segment. The result is that potentially beneficial AI innovations are not pursued or implemented.

### 3.3 List Based Definition

A list-based definition is advantageous for entities that clearly do not fall within its scope, as it provides confidence that they are excluded from regulatory capture. Our experience, including in facilitating workshops where participants were asked to determine whether an AI was high-risk based on the list, reported that the list was 'easier' to apply, but also noted that the list-based approach would likely define many more AI's as high-risk, than the proposed principles. There was also a view, which we support, that low-risk applications may be caught by the use cases listed but should not be subject to the guardrails.[5]

**Example**

A University Lecturer uses Cogniti, an AI agent/chatbot creating tool for educators, to create a Socratic tutor to chat with students about their assignment. The lecturer uses the chats between students and the chatbot to evaluate formative learning outcomes. Under the list-based definition, this is likely a high-risk use case "AI system used in …evaluating learning outcomes or monitoring student behaviour". The lecturer does not have the resources to discharge the obligations in the 10 proposed guardrails. As a result, the chatbot is never deployed.

Regulatory capture should be avoided where it does not address meaningful risks.

### 3.4 Recommendations

We recommend a domain specific approach to AI regulation. If such a targeted, domain specific approach is adopted, a broad definition of high-risk AI is less relevant or unlikely to be necessary. Further, were it required, the definition could be specific to the domain in question, crafted following consultation with the relevant domain experts.

## 4 General-Purpose AI (GPAI)

### 4.1 General

S&R supports a technology-neutral approach to AI regulation. Nevertheless, we offer the following insights on the regulation of GPAI.

### 4.2 GPAI Regulation in Australia may be premature

We consider it may be inopportune for Australia to act to regulate GPAI in Australia because:

(a) As the Proposal Paper notes, "most highly capable GPAI models are not currently developed domestically"[6];

(b) Regulation of GPAI in large and influential jurisdictions such as the EU will have a flow-on benefit to Australian users who may choose deploying an EU AI Act compliance GPAI in preference to a non-compliant GPAI;

---

5 We note that the EU AI Act uses a list to identify high-risk AI, but also includes "*mechanisms to ensure that low-risk AI uses are not inadvertently captured*" - Proposals Paper page 26 and footnote 64.

6 Proposals Paper page 29.

(c) Australian GPAI regulation may prevent law-abiding organisations and individuals from gaining access to certain AI functionalities or the most capable AI models. In this regard, we note that both Meta and Apple have announced that certain functionalities will not be available in the EU because of the challenges in complying with EU regulations.[7]

(d) The proposed definition of GPAI[8] may pose significant challenges due to its broad scope. The definition may capture many AI models, even those designed for narrow tasks, can be adapted for various purposes – for example by the technique of transfer learning. This creates a scenario where a wide range of AI technologies, including those which pose minimal risk, would be subjected to regulatory capture. Additionally, the rapid pace of AI development and the increasing techniques to use AI systems for multifarious applications means that the lines will increasingly blur between specific and general-purpose AIs.

Having said this, there are good questions that should be asked around Australian citizen and Australian organisational data being used to train GPAI without opportunity for those Australian "creators" to be provided fair value for their content (regardless of what might be the *de minimis* nature of that fair value) and likely in breach of Australian copyright laws. We understand that the Attorney General's Copyright and AI Reference Group (**CAIRG**) is currently considering this issue.

## 5 Proposed Guardrails

### 5.1 General

S&R supports a technology-neutral approach to AI regulation. Nevertheless, we offer the following insights on the proposed guardrails.

### 5.2 The devil is in the detail – EU AI Act insights

The guardrails represent a valuable set of touchstones for safe and responsible AI, however, the true challenge lies in the finer details of their implementation. The EU AI Act may be instructive as an example of what those finer details might be.

Under the EU AI Act, high-risk AI systems require a swathe of governance actions e.g. risk management system (Article 9) and data governance (Article 10). These requirements are described at a high level, with use of relative terms e.g. "*appropriate and targeted* risk management measures" (Article 9) "*appropriate* for the intended purpose of the high-risk system" "*appropriate* measures" "*relevant* data gaps" (Article 10) (italics added). This ambiguity complicates both the compliance process for organisations and the ability of regulators to monitor and enforce compliance. The EU AI Act requirements anticipate standards to bring granularity to requirements. However, unlike other areas of regulation, AI lacks mature industry standards, and it may provide

[7] See Reuters, September 25 2024, '*Meta will not immediately join EU's AI Pact ahead of new law'* and Bloomberg, 22 June 2024 '*Apple Won't Roll Out AI Tech In EU Market Over Regulatory Concerns*'.

[8] The definition of GPAI on page 28 Proposals Paper: "*An AI model that is capable of being used, or capable of being adopted for use, for a variety of purposes, both for direct use as well as for integration in other systems*"

difficult for standards, particularly technical standards, to keep pace with AI advancement. Notably, the key AI standards such as ISO/IEC 42001:2023 AI Management System and NIST AI Risk Management Framework are for risk management processes rather than technical standards.

Further efforts are needed to develop comprehensive guidance, establish clear standards and provide additional resources to support the effective implementation of the guardrails.

### 5.3 Guardrail Trade-offs

The guardrails will need to be balanced against one another, where implementing one aspect may negatively impact another. For example:

(a) *Meaningful human oversight* v *AI Model accuracy:* To enable human control or intervention in an AI system to achieve meaningful human oversight will require important trade-off decisions such as between human intervention and system accuracy (testing results); and between human intervention and privacy (i.e. what personal data can the human see in order to intervene).

> **Example**
>
> An AI diagnostic support tool in a breast ultrasound machine outputs a percentage likelihood of breast cancer. The AI prediction model is not the most accurate model available but has been selected by the developer because it provides explainability to doctors. Here accuracy has been traded-off in favour of explainability.

(b) *Transparency across AI supply chain v AI System security:* Transparency may inadvertently provide opportunities for malicious actors to compromise the security of an AI system. The level of transparency will need to be assessed against the guardrail requiring protection of AI systems.

### 5.4 The "many hands" problem

We note that requirements of the proposed mandatory guardrails differ depending upon whether the entity is an AI developer or an AI deployer or both.

The line delineating developer from deployer is unclear and likely to become more blurred as it becomes increasingly easy for deployers (even those without computer science training or the ability to code) to customise AI models and change AI outputs through techniques such as retrieval augmented generation and system prompts and to create agentic agents.[9]

> **Example**
>
> A University Lecturer takes Cogniti, an AI agent/chatbot creating tool for educators, developed by the University of Sydney, and adds extensive system prompts to guide its behaviour in a specific manner. At what point does the University Lecturer become a developer (rather than just a deployer) of that AI system?

[9] For example, by using Microsoft Copilot Studio to create AI agents.

AI systems are frequently the result of a multi-tiered structure where multiple AI models serve as building blocks for other more sophisticated or specialised AI models. For each individual AI model in the chain, responsibility is further diffused among those providing and procuring the training data and those developing and fine-tuning the AI model. Further changes to model performance may be made using techniques such as retrieval augmented generation and system prompts. Finally, end-users of an AI system can potentially manipulate the system to deliver adverse outcomes using techniques such as prompt injection and other forms of adversarial attack or use the system for a purpose which was never intended. In the submission writing workshops convened by Stirling and Rose in Sydney and Perth, developers expressed concern regarding the degree of developer responsibility in a complex AI supply chain network where deployment use cases and end-user behaviours may be "many hands" removed from the developer's original intent.

### 5.5 End Users

We note that there are no proposed obligations on End Users. Given the scope for adversarial attacks and misuse by End Users, we consider that there may be appropriate cases (e.g. non-consumer) where Developers and Deployers should not be held accountable where the use case diverges materially from use cases intended and reasonably foreseeable by the Developer or Deployer or where there has been a material or malicious breach of the terms of use.

### 5.6 Informing Users of AI decisions, interactions and content

Many users of AI experience AI not as AI, but as products and services in which AI may well be invisible (and non-alienable). Ubiquitous AI integration is likely to become inextricably intertwined with human and system operations thus blurring the demarcation between human and AI interaction. We foreshadow that in the future, it may well be that individuals assume they are dealing with AI decisions, interactions and content in some form or degree unless informed that they are dealing solely with a human decision, interaction or content.

There is currently something of a groundswell around the notion that users have a right to know, or alternatively should be informed when they are dealing with AI. This position may well be an interim step, but ultimately it is at odds with the hybrid future non-alienable AI trajectory discussed above in the Executive Summary. In the future, the volume of decisions, interactions and content that are #humancreated will likely constitute the minority compared with those of AI.

Further, there is practical complexity in determining what degree of AI involvement would mandate such transparency.

**Example**

My friend uses an AI filter to enhance our appearance in photos. This raises the question of whether such content requires an AI content watermark.

To make this point another way, if AI systems (or hybrid AI systems) perform more accurately and more efficiently than a system reliant on human oversight/intervention –

the correct expression of a director or legal duties may be to disclose whether it is a human (not an AI system) that has been involved in, for example, decision making. An instructive analogue for this evolution is the use of Technology Assisted Review (**TAR**) in the discovery process of litigation. Provision of legal services relying on human-only discovery, particularly where there are high volumes of documents to review, and the failure to use TAR, may be a breach of legal duties, given that TAR is faster, more efficient and more accurate in legal discovery[10]

Accordingly, we offer an alternative perspective, that transparency be required for #humancreated decisions, interactions and content. In this counterargument, watermarking is limited to when the decision, interaction or content is #humancreated and made transparent by a watermark which identifies:

(i) the decision, interaction or content as #humancreated;

(ii) the authenticating legal entity; and

(iii) the digital infrastructure verification details (e.g. such as SHA-256 hash).


The #humancreated watermark requires Lawful AI Critical Infrastructure authentication technologies to support digital signatures and time stamping, in order to verify the origin and integrity of #humancreated.

### 5.7 Deep fakes

We call out for particular attention to the burgeoning threat of deep fakes to individual privacy, democratic processes, market integrity, security and societal trust and emphasise that those seeking to deceive will not comply with laws requiring watermarking. The risks and harms of deep fakes are not confined by geographical boundaries making the challenge of regulating deep fakes a global concern. The

[10] *McConnell Dowell Constructors (Aus) Pty Ltd v Santam Ltd & Ors* [2016] VSC 734

interconnected nature of today's digital world means that a deep fake created in one country can easily affect individuals, organisations or political processes in another.

This makes the development of Lawful AI Critical Infrastructure an imperative that extends beyond Australia to global application.

In the interim, education on the importance of maintaining healthy scepticism, critically evaluating information and authentication through multiple trusted sources serves as the most effective safeguard.

### 5.8 Government as Exemplar

We also recommend that the mandatory guardrails for AI in high-risk settings be tested in a soft law approach by implementation in certain Government Departments where decision-making has high impact on individuals or communities. This practical implementation experience would enable Government to ascertain and document the degree of complexity (for example, in determining whether an AI application is high-risk), resourcing, financial impact and potential impact of implementation. Regulatory Framework

### 5.9 Prudent Domain Specific Approach

Based on current knowledge, we consider that a technology-neutral approach coupled with an impact-focused orientation is preferred to direct, economy-wide regulation of AI as a technology. In governing AI, outcomes should be uniformly treated irrespective of whether those outcomes are the result of human, AI or human + AI action.

*An act or omission which is deemed unlawful when committed by a legal person should be equally unlawful if performed or failed to be performed by AI.*

This principle holds particular relevance in contexts requiring a degree of mental element or 'mens rea'. While AI as currently conceived, lacks the consciousness and subjective intent traditionally associated with a mental element, the development and deployment of AI is nevertheless, currently driven by the intentional acts of legal persons. This necessitates the extension of legal norms and sanctions to AI-committed acts. For example, within certain laws, this may be done by changing the definition of legal "person" or "agent" to include AI.

At this time, we prefer an approach to AI regulation which:

(a) Is technology neutral (not specifically directed at AI) i.e. domain specific approach;

(b) Identifies and addresses issues and gaps in regulation;

(c) Builds on existing laws and structures;

(d) Harmonises state, federal and international requirements as far as possible to minimise compliance costs on Australian organisations, i.e. consider framework regulation; and

(e) Appropriately balances protection of individuals from harm, appropriate redress and preservation of civil society with the advantages and benefits AI innovation can offer.

### 5.10 Systematic review of legislation through AI-centric lens

We recommend the Australian government undertake a systemic review of all existing legislation through an AI-centric lens to identify areas where existing laws may fall short in addressing the harms and risks posed by AI.

## 6 Governing Increasing Capable AI Systems of the Future

### 6.1 Responsible AI Problem – AI Organisations (AIOs)

We foreshadow the potential for AIOs to have capabilities closely mimicking centralised human governance such as board directors[11] and CEOs, but with superior access to data for real-time decision-making. We note that there is a continuum in the degree of autonomy in AIOs, with governance delivered (at least in the near term) via a combination of AI and human decision-making. The pathway to fully autonomous AIOs will likely involve an ongoing expansion of the matters handled autonomously by AI agents, that is, without human intervention. We emphasise the importance of legal accountability for AIs and the need for legal personality to protect internal and external stakeholders (e.g. consumers) interacting with the AIO. We also note that an AIO may be "digitally birthed" i.e. be the progeny of another AIO (including many generations removed from a human founder, creator or developer). The case for legal responsibility only to be traced to humans, cannot be sustained where agentic capabilities of AI continue to expand.

### 6.2 AI Organisations (AIOs) as legal persons

Granting legal personality to an AIO should not be unconditional. We propose a list of bare requirements to recognise that an AIO is a separate entity that may functionally operate with complete independence separate from its founders/developers. These conditions include registration with appropriate agencies and a unique identifier as an AIO, evidence of de minimis functionality, sanctions for aberrant behaviour, economic reserves, adjustments for intent and protections for minority members/shareholders. The idea is to provide a balance between recognising the separate functionality of AIO as an organisation and ensuring that humans (e.g., directors and officers) are held accountable for their actions and decisions related to AI which they deploy in Corporations.

While we support personal liability being attached to any humans that direct and manage the AI, there are practical and legal challenges with relying on strict human responsibility alone. With respect to an AIO, the universe of potentially liable persons are the founders, developers, creators, or company which provided the resources for the development of the AIO. In the case of civil or criminal actions which require mental intent (such as fraud) it may be difficult or even inappropriate to attribute the requisite intent to a particular individual. Such a cause of action against a human actor may not be sufficient recourse for an injured party.

Further, a responsible human may be too far removed from an actual decision or action of the AIO to attribute fault to the human. The realistic possibility of an AIO alone being responsible for making an aberrant decision or action which founders, creators, or

[11] Please see our article on "Synthetic Directors" on the Stirling & Rose website.

developers are not reasonably able to foresee, (particularly when the founder, creator and developer is another AIO rather than a legal person) draws a crucial distinction between an AIO and a corporation or any other traditional non-human legal personality. A corporation, partnership or unincorporated association ultimately depends on a human in the loop to make a decision, whereas the AIO may be able to independently make decisions for itself with limited or no human direction. Technically, this may be achievable without the AI being sentient or having the ability to 'think' as we currently understand 'thinking'.

This untethering of traditional organisation into an AIO is also possible, for example where a responsible "human in the loop" (director) dies, yet the organisation continues to operate according successfully to pre-established agentic instructions.

For these reasons, there is merit in potentially granting legal personality to protect internal and external stakeholders interacting with the AIO. To iterate:

(a) The increasing capability of AI will mean that greater management of an organisation for day-to-day functions by AI becomes feasible and widespread;

(b) An AIO presents potential improvements over a human-led organisation. The capacity to review more relevant information quicker, undertake more efficient decision-making, and to be less prone to environmental biases (e.g. improper influence and fraud) than human counterparts means that there may be productivity gains and other benefits from AIOs. Individuals and investors may prefer to be shareholders in an AI-led organisation than shareholders in a human-led corporation; and

(c) From the perspective of third-party protection, if the above two propositions are accepted, then the AIO itself should have legal personality to be subject to legal sanction not only for the benefit of injured parties but also to empower regulators to deter aberrant behaviour. This provides an additional avenue of relief to the extent that action against individuals is not appropriate.

However, we have two key caveats in relation to conferring legal personality to AIOs.

First, such legal personality is not unconditional, and we set out details of conditions that lawmakers may wish to impose on an AIO in exchange for conferring legal personality.

Secondly, recognising legal personality of AIOs should not abrogate personal responsibility of human actors to the extent they are involved in decision-making and management. Granting legal personality to an AIO expands the recourse available to external and internal stakeholders to both the AIO and/or human actors (as appropriate), rather than a liability shielding mechanism for any one responsible person.

The liabilities of responsible human actors and the AIO should be on a non-mutually exclusive continuum that reflects the involvement of the human actors and the AIO in the actual decision-making and management of the AIO. There is also a quibble that it may be incorrect to assume that developers are only ever "human persons".

### 6.3 Suggested Conditions for Granting Legal Personality to AIOs

A suggested list of conditions may include the following:

(a) **Registration:** All AIOs should be registered with the appropriate local and international agencies and should have their own unique identification (such as a registration number or digital fingerprint). We would propose that there should be a dedicated governmental authority to regulate AIOs and monitor the register.

(b) **Identification as an AIO:** All AIOs should have a signifier in their names such that they are readily identifiable as an AIO in all interactions. Much the same way companies are identified by "Ltd", "Inc", "LLC" or other signifiers of their corporate status, AIOs should also be identifiable as an AIO – for example ACME AIO Ltd.

(c) **Functional AIO test:** An AIO must be able to provide reasonable evidence of its capability to discharge de minimus functionality, i.e. those things necessary to discharge the AIO's legal obligations. For example, paying annual AIO registration fees. This evidence would need to be provided as a condition to registration of the AIO, but also possibly on a periodic and ongoing basis to confirm the continued functionality of the AIO in what is a dynamic and rapidly evolving arena where consumer harm in interacting with an AIO (or at least with an early generation AIO) may be elevated.

(d) **AIO sanctions:** Economic risk allocation does not solve the role of "skin in the game" whereby personal liability and the potential for adverse action against an individual ensures alignment of interests and actions. Even with limited recourse vehicles, there is still a responsible person (in the case of a company, directors still owe director duties to the company, and in certain circumstances to, or are personally responsible for harm inflicted on, third parties e.g., misleading and deceptive conduct).

(e) **Suspension or termination for aberrant behaviour:** The applicable governmental authority regulating AIOs should be conferred with powers to suspend the operation or terminate an AIO and exercise rights of forfeiture against AIO assets. While it is impossible to imprison an AIO, effective sanctions may include suspension or termination of the AIOs ability to access data which is a critical input for effective AIO operations or suspension or termination of all active operation of an AIO. To prevent AIOs "phoenixing" a sanctioned AIO, there would have to be a means of identifying the code of a phoenixed AIO from the underlying code of one that has been sanctioned.

We reiterate that even if AIOs had legal personality and could be separately held liable, it remains prudent to hold individuals accountable: (a) to the extent that the private citizens are engaged in directing the AIO; (b) to the extent that it is difficult to properly censure an AIO to ensure that there is 'skin in the game' and there is insufficient deterrence from bad behaviour.

(f) **Economic reserves:** It may be feasible to improve the economic risk of AIOs to counterparties by imposing some form of credit support. This could be in the form of actual asset backing (i.e. collateralisation) to meet potential claims or guarantees provided by individuals or legal entities of substance or insurance (such insurance to be for a sufficient level of cover for the AIOs activities and incurred obligations and with no material exclusions).

A counterargument to this is that without some form of collateralisation, an AIO represents a similar risk profile to external counterparties as corporate or other limited liability vehicles which have minimal asset backing e.g., shell companies. In which case, it is a matter of caveat emptor for parties that fail to do their due diligence (assuming that the AIO is registered and identified as an AIO pursuant to conditions 1 and 2 above.).

(g) **Intent of an AIO:** As referred above, causes of actions and offences which presuppose mental intent will need to be adjusted for AIOs, given there may not be a governing 'mind' in the legally understood sense. This may involve abolishing the mental intent from such causes of actions and offences to the extent an AIO is involved or relying on a completely objective (and possibly expert determined) test to determine what the AIO should have done. Corporations attribute intent to their directors and officers. We do not think this presumption should automatically apply to AIOs, but it should be available as a continuation of our liability continuum comment above, to the extent that a founder or other human actor is a shadow director or is otherwise involved in the material decision-making.

(h) **Minority AIO member/shareholder protections:** Bare minimum protections may be granted to minority AIO shareholders (e.g., no compulsory transfer or extinguishment of minority shares). This is in conjunction with the evidence required to establish a functional AIO may comprise the base governance protections to minority member/shareholders to avoid potential fraud on the minority.

(i) **Lawful AI Critical Infrastructure:** is required to authenticate, regulate and govern AIOs.

Going forward, a smart legal contract supported by Lawful AI Critical Infrastructure can act as an AIO governance tool with greater ability to moderate the complexity of risk allocation, rules of engagement, obligations and remedies between differing members/shareholders and regulatory bodies in respect of the AIO and its individual assets.

For example, a network or constitution and some form of member/shareholder/unitholders AIO agreement, connected to the third-party legal agreements drafted and constructed as smart legal contracts (with natural language and coded instructions) could do the work of automating the performance of an AIO or a hybrid AIO with human governance rights and obligations.

If the Department would like further information on this final point, we would be delighted to provide additional written or verbal comments.

## 7 Conclusion

The increasing capability and multifaceted nature of AI demands an adaptable and future-focused policy response, regulatory framework and Lawful AI Critical Infrastructure.

The recommended technology-neutral, impact-focused approach ensures that

regulatory measures are not confined to the limitations of current AI technologies but are adaptive to the evolving landscape.

Finally, we postulate a future where increasingly capable AI autonomously operates enterprises and in response, set out a rationale and indicative requirements for extending legal personality to AI Organisations (AIOs), including the key requirement for Lawful AI Critical Infrastructure in the age of intelligent machines.