Tech Council

TCA RESPONSE TO QUESTIONS ON NOTICE
ADOPTING AI INQUIRY

Question 1:

Senator SHOEBRIDGE: We could adopt the EU model on AI safety and AI transparency. That would be one option if we wanted to pick a global example.

Mr Black: As I said, there are aspects of the EU model on transparency that I think are worthy of further consideration in Australia.

Senator SHOEBRIDGE: Could you respond on notice about what elements of that you think are positive and what elements of that you think create unnecessary challenges.

Mr Black: Absolutely, I'd be verry happy to.

Senator SHOEBRIDGE: Going to deep fakes—

CHAIR: Before you do, could you include that on the US model and the Canadian model in particular as well—and Singapore.

The Tech Council of Australia (TCA) supports transparency in AI systems to ensure trustworthy, safe, and responsible AI use. This fosters confidence and supports responsible AI adoption in Australia.

Overleaf, we summarise the different approaches to AI transparency in the EU, US, Canada and Singapore. Importantly, we have observed that there is no 'Brussels effect' on AI policy – that is, countries around the world are not coalescing around the EU model and have proceeded to take different approaches to AI governance. We provide a comment on the TCA position on each of these jurisdictions, their approaches to transparency, and offer points that Australia can learn from.

TCA note on transparency

Transparency for AI systems is a key principle at the highest levels of international governance and for industry when it comes to responsible AI adoption. This includes the UN General Assembly's recently adopted Resolution A/78/L.49[1], the OECD's AI Principle including Transparency and explainability[2], as well as statements from Governments around the world in the recent Seoul and Bletchley Declarations (signed by over 27 and 28 countries and the EU, respectively).

In Australia, transparency is embedded in our National AI Ethics Principles and is being considered in the context of the Government's work on the voluntary AI safety standard, as well as mandatory guardrails for AI systems in high-risk settings.

The general principle of "transparency" for AI however, can be taken to mean different things to different people. It could mean *seeing through* a system (to understand a system's overall performance, informational aspects around the materiality of the hardware, development, or maintenance requirements, and user notices and disclosures) or *seeing into* a system (which requires expert knowledge and techniques to understand the nature of the data, connections, algorithms and computations that generate a system's behaviour including its techniques and logic).

This latter type of transparency is especially difficult to achieve due to the complexity of large-scale AI models and algorithms, especially ones deploying deep learning techniques. These types of models have millions and billions of parameters and can rely on multi-layered neural networks to

[1] UNGA Resolution A/78/L.49: Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development: (j) for transparency and reporting requirements applicable with international, legal, and subnational legal frameworks and (k) promoting transparency, predictability, reliability and understandability throughout the life cycle of AI systems, p6.

[2] OECD, Recommendation of the Council on Artificial Intelligence, Principles for responsible stewardship of trustworthy AI, Principle 1.3: AI Actors should commit to transparency and responsible disclosure regarding AI systems. To this end, they should provide meaningful information, appropriate to the context, and consistent with the state of art: (i) to foster a general understanding of AI systems, including their capabilities and limitations, (ii) to make stakeholders aware of their interactions with AI systems, including in the workplace, (iii) where feasible and useful, to provide plain and easy-to-understand information on the sources of data/input, factors, processes and/or logic that led to the prediction, content, recommendation or decision, to enable those affected by an AI system to understand the output, and, (iv) to provide information that enables those adversely affected by an AI system to challenge its output.

generate outputs. While these methods mean that it is challenging to fully understand how specific outputs are generated, it is also this same characteristic that makes AI capable of discovering complex patterns and correlations within vast amounts of data that humans may overlook, process, or find difficult to detect. This discovery potential fuels scientific advancements and innovations to enable breakthroughs in various fields.

The UK Central Digital and Data Office (CDD) and Responsible Technology Adoption Unit (RTA) offers two helpful classifications related to explanations to provide transparency that may be useful for Australian policymakers, these are:

(i) *process*-based explanations which give information on the governance of an AI system across its design and deployment; these are primarily aimed at demonstrating that an AI system has been developed with good governance processes and best practices (e.g. taking adequate testing and assurance measures across system production and deployment); and,

(ii) *outcome*-based explanations, which tell you what happened in the case of a particular AI-generated outcome (e.g. clarifying the input data used for a particular model or an easily understandable user explanation).

The field of developing AI model transparency is also rapidly advancing, marked by continuous research and development with more innovative technical solutions and tools being developed frequently. Illustrative of this progress are the results of Stanford University's second *Foundation Model Transparency Index,* recently released in April 2024. The index revealed 'substantial improvements' since October 2023, underscoring important progress in the development of transparency measures for AI in just six months.

TCA also highlights the continued development of international technical standards for AI systems. These are helpful as they serve as common global frameworks for AI governance that encourage interoperability and oversight at a technical systems level. In particular, ISO/IEC 42001:2023, which focuses on procedures for AI management systems, contains actions to address transparency as well as explainability for responsible AI use and deployment.[3] Moreover, ISO/IEC is currently in the process of drafting a standard on the taxonomy of information elements to assist with AI transparency (ISO/IEC DIS 12792). The US' NIST AI Risk Management Framework is another example of a widely adopted standard used by industry that addresses transparency throughout the AI lifecycle, from design decisions to post-deployment.

Further pitfalls to consider on transparency measures:

- **Notices, disclosures, and capacity to receive:** How information is conveyed to another individual or entity will inevitably differ depending on the context. Different actors will require varying levels of detail in the information they receive. The use-case affects what type of explanation required. Several factors will affect what information an individual expects or finds useful.
- **Technical records for transparency:** Technical records will need to be kept in such a fashion that they are understandable. This may become especially challenging as systems become even more advanced and integrated. Conversely, scrutinising authorities reviewing records or documentation may require a degree of specialised technical knowledge. The disclosure of technical records also needs to balance the commercial confidentiality and intellectual property rights underpinning an AI system with the appropriate handling procedures and protections.

The TCA encourages an outcomes-based and principles-based approach to transparency that is coherent with international standards, including ISO/IEC 42001:2023 and the US NIST AI Risk Management Framework. Best practice governance and technical solutions are moving swiftly and it is important that the Australian Government work with industry and researchers to ensure our requirements do not quickly become outdated.

[3] In a similar way that ISO/IEC 27001:2002 has emerged as the global gold standard for cybersecurity risk management, we anticipate that ISO/IEC 42001:2023 will be a leading standard for AI in a similar way.

European Union

The EU's approach is more centrally coordinated than other jurisdictions in the world and contains binding rules for application. The EU's need to revise the EU AI Act after the release of ChatGPT's new model in March 2023 highlights the inflexibility of rigid legal frameworks designed for rapidly evolving technologies. The TCA does not support the adoption of an EU AI-style Act for Australia, instead recommending that we utilise and build on our existing framework of technology-neutral and sector-specific laws and standards that are governed by expert regulators.

The EU Act regulates different types of AI systems based on their risk levels (minimal/low risk, limited risk, high-risk and unacceptable risk). For those in the high-risk category, the EU places transparency obligations requiring that:

- providers of such systems undertake registration in a public EU database prior to deployment (Article 51);
- developers provide information to deployers with instructions for use that is accessible and comprehensible to users, containing details including human oversight measures, data testing, characteristics of an AI system including its capabilities and limitations (Article 13); and,
- for certain AI systems and General Purpose AI (GPAI) models, additional obligations apply for systems that generate synthetic content relating to labelling and deepfakes (Article 52).[4]

A key aspect of the EU Act is the inclusion of roles and responsibilities of actors across the AI value chain. The obligations for these assessments are, however, likely to entail considerable costs and compliance burdens, particularly for SMEs seeking to employ AI systems.

The EU takes a 'hard law' approach with detailed prescriptive requirements. While this has the benefit of enabling some regulatory certainty, this approach also comes with several downfalls. The stringent regulations may struggle to keep pace with the rapid evolution of AI technologies, as has already been demonstrated with the release of GPT-4 and with changes in best-practice AI governance. Burdensome requirements risk stifling innovation in the sector, especially for smaller industry players, limiting their ability to compete with larger companies. We believe this will lead to lower investment and technology development and adoption, potentially putting the region at a competitive disadvantage. The Act will also prove challenging to amend if implementation issues are found to arise.

Notably, the EU also has overlapping requirements relating to transparency and AI as well as ADM systems extant in the European regime. This includes provisions in the GDPR, which deem that algorithmic systems are prohibited from making significant decisions affecting legal rights without human supervision. The GDPR already also guarantees an individual's right to 'meaningful information about the logic' of algorithmic systems ('the right to explanation'). The Digital Services Act (DSA), passed in November 2022, also considers AI and creates new transparency requirements that require independent audits, enabling independent research on large platforms; it also requires platforms to explain content for content recommendations and offer users alternative recommender systems. Similarly, the Digital Markets Act enables the EU Commission to conduct inspections of 'gatekeeper' data and AI systems. While the GDPR, DMA and DSA are not primarily about AI, they also provide overlapping transparency requirements that apply to AI systems.

Lessons for Australia
• Establishes a risk-based approach that enables regulatory efforts to be focused on use cases presenting the highest risk; this includes transparency requirements on developers to provide information to deployers that can help them better understand the characteristics, limitations and development process associated with the models they are refining or deploying.

[4] EU AI Act: (i) a notice requirement to ensure AI systems are designed and developed in a way that natural persons are informed that they are interacting with an AI system; (ii) providers of AI systems generating audio, image, video, or text content to ensure outputs are labelled in a machine readable format and detectable as AI generated or manipulated; and, (iii) deployers generating image, video, or audio content constituting a 'deepfake' to disclose content has been AI generated or manipulated.

- Leverages technical standards (via CEN-CENELEC) that are aligned to international standards (ISO/IEC) to underpin the Act's practical application.
- Creates duplicative or overlapping regulatory obligations on transparency across different legal regimes, which adds additional costs for businesses that Australia should avoid.

United States

The US' voluntary approach has resulted in meaningful industry commitments and steps towards responsible AI deployment and innovation at the highest levels of Government, most recently through President Biden's Executive Order. The US has invested significant Government funding and resources into driving responsible AI research and development, increasing the federal government's capacity to use and manage AI, as well as enabling opportunities for private-sector collaboration to drive responsible adoption.

There are a number of policies within the US approach relating to AI transparency. This includes the US Executive Order, the AI Bill of Rights, as well as the work of the National Institute of Standards and Technology (NIST) on AI models.

The US Executive Order instructs government agencies to undertake transparency measures related to AI, including:

- The Secretary of Commerce to identify techniques to authenticate content and label synthetic content, such as through watermarking (4.5);
- The Secretary of Labour to develop and publish principles and best practices for employers that could be used to mitigate AI's potential harms to employees, including transparency principles for the use of AI systems in workplaces (s6);
- Independent regulators to conduct due diligence and monitor any third-party AI services they use; as well as clarify requirements and expectations related to AI transparency (s8);
- The Office of Management and Budget to publish (March 2024) for agencies to strengthen the use of AI in government, including steps to watermark or label outputs from Generative AI, as well as issue instructions to Government agency use cases; and,
- Requiring companies developing dual-use foundation models to provide the Federal US Government with continuous reporting and records on the activities of the model and results of red-teaming activities (s4.2).

The US AI Bill of Rights, published in October 2022, sets out the Government's five principles which include "safe and effective systems" and "notice and explanation". The Bill of Rights is supported by a technical companion that sets out practices to guide the design, use and deployment of AI systems setting out actions on transparency to provide plain language documentation and descriptions of system functions; notice that such systems are in use; be notified of significant functionality changes; explanations for decisions by ADM systems; and enable meaningful explanations that are calibrated to the level of risk based on context.

The US has also mobilised the National Institute of Standards and Technology agency to develop resources that enable organisations to manage the risks of AI. Transparency is highlighted as a characteristic of trustworthy AI within the NIST AI Risk Management Framework (AI RMF) and through the NIST AI RMF Playbook which provides practical steps organisations can take on transparency.

Lessons for Australia

- A whole-of-government approach to uplifting capacity within Federal departments and agencies, enabled through executive coordination and leadership at the highest levels.
- Significant support and federal funding into enabling AI research and development which is expected to contribute to the development of new technologies, techniques and methods that mitigate AI risks.

- Strong focus on encouraging industry to undertake watermarking and content authenticity to improve the transparency of AI-generated content;
- Approach is underpinned by a strong understanding of the need to maintain global AI leadership and strategic/economic competitive advantage;
- A non-prescriptive approach that acknowledges evolving best practices for AI governance, assurance and testing;
- The development of iterative standards that enable flexibility to evolve as technology advances;
- Regulatory approach that enables and empowers existing regulators to increase their own regulatory capacity and apply their contextual domain expertise.

Canada

Canada's approach to AI regulation is intended to focus on AI in high-risk settings, but in some respects is broader and less defined than the EU's approach. The delegated regulation and boundless scope of AIDA (contained within Bill C-27), lack of distinction between high and low-risk contexts, and extent of discretion provided to the relevant Minister has raised concerns amongst industry.

Canada's proposed Bill-C27, the *Digital Charter Implementation Act*, includes a proposed Artificial Intelligence and Data Act (AIDA). If passed, AIDA will establish a series of obligations for AI design and development for high-impact AI systems. Recent amendments to the AIDA introduce additional provisions for transparency for high-impact and general-purpose AI (GPAI) systems. These provisions include obligations to:

- inform the public of AI system use where a person could believe they are interacting with another human being, disclosure to be provided when they are communicating with an AI system. This is subject to several exceptions including where a system is a consumer product (s6(1)) or where GPAI systems produces outputs such as text, images, or videos, measures to be taken so that members of the public can identify this as an AI output (s7(1));
- provide plain language descriptions of AI systems when they are deployed for the first time, including its capabilities and risks of harm or biases, which must be published and publicly available (s7(1)(f), s8(1)(a), and ss10);
- where a GPAI or high-impact system has caused harm, the operator must provide a report to the Commissioner in accordance with AIDA regulations (s8.2(1) and s11).

The Canadian approach leans on delegated and subordinate regulation to furnish details for specific rules and obligations. The Minister of Innovation, Science, and Industry is empowered with considerable authority to enforce compliance, including conducting audits and imposing penalties for violations. However, this leaves considerable uncertainty around what AIDA may foreseeably govern, which may include any and all possible harms. Given that AI is a cross-sectoral technology, this broad discretion also risks encroaching into other existing regulatory domains. This model poses challenges for predictability and clarity, which may leave industry uncertain on how they should adapt their business operations to comply with the law. There are also significant concerns that AIDA lacked broad public consultation in a meaningful way.

Lessons for Australia

- Undertake appropriate public and stakeholder consultation before proceeding with legislation.
- Be wary of broad Ministerial discretion to enact rules and regulations, without establishing clear scope, thresholds and boundaries.

TechCouncil

Singapore

Singapore has taken an approach to AI governance that leads with the development of practical tools and techniques for AI assurance and public-private partnerships. Singapore's National AI Strategy, first articulated in 2019, encompasses a governance framework for AI that encourages the development of toolkits offering testing, reporting guidelines, and industry-derived use cases. Singapore is capitalising on the opportunity gap to enable testbeds and a supportive environment to develop responsible AI tools that may be adopted worldwide. In global discussions, Singapore has served as a leading reference point for innovative governance and as an exemplar of how Government can work with leading industry actors to co-design and develop best practice tools for responsible AI adoption.

Like the US, Singapore has coupled a largely voluntary governance model with ongoing infrastructure investments to support the country's digital economy which is critical for long-term competitiveness. The US and Singapore Governments also recently published a joint statement on Shared Principles and Collaboration on AI[5]. This underscores Singapore's need and commitment to collaborate with leading international partners while signalling that Singapore's digital economy is both open and interoperable.

In striking a balance between allowing innovation in AI to develop, and safeguarding public interest in AI safety and governance, the Singapore government has developed various governance frameworks and tools that guide AI deployment and promote the responsible use of AI, including:

- A Model AI Governance Framework (2019, updated in 2020) that provides detailed guidance to private sector organisations on addressing key ethical and governance issues when deploying AI solutions. It details measures to enhance transparency in AI models such as user notices and disclosures, and recommended practices for explainability and transparency including model documentation.
- A draft Model AI Governance Framework for Generative AI (2024), which expands on the 2020 framework to address new issues that emerging from Generative AI. Amongst the nine dimensions articulated in the framework, it includes a statement on Content Provenance and articulates the importance of "transparency about where content comes from as useful signals for end-users".
- AI Verify, an AI governance testing framework and toolkit designed to help organisations validate the performance of their AI systems against AI ethics principles through standardised tests. AI Verify was developed by the IMDA in consultation with private sector organisations
- The AI Verify Foundation (AIVF), a not-for-profit foundation to crowd in expertise from private sector organisations including Adobe, Amazon, Google, IBM, Microsoft, and the global open-source community to develop AI testing frameworks, standards and best practices.

Lessons for Australia

- Development of practical tools and assurance mechanisms that operationalise transparency principles, developed with industry buy-in and enabled by public-private partnerships;
- Partnership and collaboration with leading countries like the US to secure strategic and economic competitive advantage to ensure transparency as well as other AI governance mechanisms are aligned;
- Significant support and federal funding into enabling critical AI infrastructure and broader investment into local innovation ecosystem to enhance development on all vectors of responsible AI innovation including transparency;
- Agile and responsive government development of frameworks and guidance to inform businesses, including the work of regulatory guidance on transparency principles.

[5] US Department of Commerce, Fact Sheet: U.S. Singapore Shared Principles and Collaboration on Artificial Intelligence (June 2024).

Question 2:

Senator SHOEBRIDGE: I could be wrong, but I think a number of your members attended the Munich Security Conference, where there was broad agreement by a significant part of the global tech industry to fight back against what was described as deceptive artificial intelligence election content. The statements were positive, but I'm yet to see any allocation of internal resources within the industry to actually enable it to happen and any outcomes from it. We're in this year of elections, 2024, and last time I checked we're about halfway through it, and I'm yet to see any actual clear activity from the sector—no compulsory watermarking, no taking down of content. None of what was promised in the Munich Security Conference seems to have been delivered. Is it a lack of interest, a lack of resourcing? What is the barrier?

Mr Black: As I said, the companies are investing quite significantly into the global Content Authenticity Initiative, the purpose of which is to combat misinformation and deepfakes by verifying the provenance and authenticity of images, of video content et cetera so people can tell if it's been altered or produced by AI. There is substantial investment happening in that space, and we are seeing a lot of companies already taking action. I am happy to take on notice, specifically, the question on the Munich AI commitments, consult with those companies and come back to you with a fulsome response on the action that's been taken, but I know that there has been significant investment.

At the Munich Security Conference on February 16, 2024, twenty-five leading technology companies agreed to combat harmful AI-generated content specifically in elections and voting contexts. Known as the Tech Accord, this initiative involves voluntary commitments by each signatory throughout 2024 to develop and implement technologies aimed at countering deceptive AI-generated content.

The TCA is acutely aware that deepfakes present a critical challenge due to their potential misuse by malicious actors to undermine democracy, mislead the public, harm and harass individuals or groups of individuals, and erode trust in authentic and credible information sources. We support efforts to counter the role of deceptive AI-generated election content.

TCA members who are signatories of the Tech Accord include Adobe, Amazon, Google, IBM, LinkedIn, and Microsoft. As TCA members invest in more capable models, they are equally deeply committed to enhancing AI responsibility. This includes equipping users with tools to detect synthetically generated content, and empowering them with contextual information to help prevent the spread of misinformation. Overleaf is a summary of their efforts to fulfill their commitments under the Accord.

The Tech Accord is a set of voluntary commitments that complement other existing industry efforts. Several TCA members are also signatories to the White House Voluntary AI Commitments, which calls on companies to develop robust mechanisms, including provenance and/or watermarking systems for audio and visual content, as well as the recent Seoul Declaration and Frontier AI Safety Commitments signed in May 2024.[6]

We value the Senate's interest in the Tech Accord and welcome this opportunity to share the progress that has been made. While significant strides have been made towards meeting the Accord's goals, we acknowledge and recognise there is still much more work ahead.

Adobe	Adobe deeply considers the impact it has on users, consumers, and society at large. Adobe is encouraged by the Accord's collective commitment to developing and implementing technological solutions like provenance, watermarking, and classifiers to label and provide metadata about AI-generated content. Adobe is supportive of the Accord's commitment to foster public awareness through education campaigns on the risks of deepfakes. Adobe continues to advocate in all relevant countries for mandatory labelling of all online election campaign material made with AI. • **Content Authenticity Initiative:** Adobe' has spearheaded industry initiatives to fight AI-generated information, founding the Content Authenticity Initiative (CAI) in partnership with the New York Times in 2019. Adobe continues to invest in this

[6] This includes: internal and external red-teaming of frontier AI models and systems for severe and novel threats; to work toward information sharing; to invest in cybersecurity and insider threat safeguards to protect proprietary and unreleased model weights; to incentivize third-party discovery and reporting of issues and vulnerabilities; to develop and deploy mechanisms that enable users to understand if audio or visual content is AI-generated; to publicly report model or system capabilities, limitations, and domains of appropriate and inappropriate use; to prioritize research on societal risks posed by frontier AI models and systems; and to develop and deploy frontier AI models and systems to help address the world's greatest challenges.

	community, which has now grown significantly to over 3000 diverse global organisations that are committed to finding ways to include provenance standards in their applications. • **C2PA Content Credentials Standard:** The CAI supports the Coalition for Content Provenance and Authenticity (C2PA) and the development of an open technical Content Credentials standard. C2PA helps address the prevalence of misleading information online through the development of technical standards for certifying the source and history of media content. It is an important tool that provides metadata about the provenance of an image, and is seeing broad adoption across many companies that are part of the Munich Accord. • **Research on misinformation in Australia**: As part of a global study about consumer experience of misinformation and the impact of generative AI, Adobe recently published the findings of its *Future of Trust Study* for Australia and New Zealand, surveying over 1000 consumers with insights on the need to protect election integrity. • **Responsible and inclusive datasets:** Adobe's generative AI models in Firefly are responsibly trained on Adobe Stock images, openly licensed content, and public domain content where copyright has expired. It does not train Firefly models on customer content. • **Responsible Innovation and generative AI:** Adobe is guided by AI Ethics Principles including accountability, responsibility and transparency. Adobe has developed standardised processes from design, development and deployment including training, testing, AI-impact assessments, diverse human oversight, and feedback mechanisms to remediate potentially biased outputs.
Amazon	Amazon is committed to secure, safe, and fair democratic processes, which is why Amazon has joined other technology companies in a pledge to work together to identify, counter and prevent deceptive AI content to protect the integrity of elections. • **Amazon Titan:** Amazon has focused on building tools for its AI products and services that could be at higher risk for creation of content that could be used for election disinformation. Amazon Titan's foundational models are built to detect harmful content, reject inappropriate content in the user input, and filter the model's outputs containing inappropriate content (such as hate speech, profanity, and violence). Amazon's Titan Image Generator is a generative AI foundation model that customers can use to quickly create and refine realistic, studio-quality images. Importantly, its image models are designed for the purpose of creative content generation, and not for the depiction or information retrieval of real-life people or events. • **CP2A standard:** Amazon is quickly working to implement the Coalition for Content Provenance and Authenticity (C2PA) standard in its generative AI models. C2PA helps address the prevalence of misleading information online through the development of technical standards for certifying the source and history of media content. C2PA is an important tool that provides metadata about the provenance of an image, and is seeing broad adoption across many companies that are part of the Munich Accord. • **Customer and content policies:** Amazon is committed to developing safe, fair, and accurate AI and ML services and providing customers with tools and guidance to assist them in building and using AI and ML applications responsibly. There are content policies across its services that mitigate the risk of disinformation and other harms to its customers. Amazon's Responsible AI Policy prohibits the use of AI tools to depict a person's voice or likeness without their consent or other appropriate rights, including unauthorised impersonation. Amazon policies prohibit fraudulent

	activity and Amazon prohibits impersonation. Amazon has a clear Acceptable Use Policy that governs the use of its services. When Amazon is made aware of anyone who is not in compliance with this policy, it investigates and takes appropriate action. • **Content detection:** Across its services, Amazon uses machine learning and other technology to identify content that violates its policies, regardless of whether it is AI-generated or not. All of Amazon's services offer mechanisms enabling users to flag content that may be fraudulent or otherwise violate its terms and conditions. Amazon's trust and safety teams review potentially fraudulent content against its policies, and they remove or disable content that violates its policies when they are made aware of it, including AI-generated election-related disinformation. • **Collaboration on best practices:** Amazon also leverages a range of third-party organisations to help them coordinate across industry, civil society, and government on issues related to election disinformation and other areas of AI safety. Amazon is a member of the Frontier Model Forum, an industry non-profit working to drive forward industry best practices around AI safety. Amazon is also a participant in the Partnership on AI, which is leading multistakeholder research on deceptive AI content. Amazon is also the private sector partner of the Global Challenge to Build Trust in the Age of Generative AI, a flagship project of the G7 to fight misinformation, in cooperation with the UN, the Organisation for Economic Co-operation and Development (OECD), and the Global Partnership on AI. Amazon is a member of the National Institute of Standards and Technology (NIST) AI Safety Institute Consortium, in which it participates in the working group developing safety guidelines and standards around synthetic media safety.
Google	Google is committed to supporting democratic processes by surfacing high-quality information to voters, safeguarding platforms from abuse, and equipping campaigns with the best-in-class security tools and training. Google is also leveraging AI models to augment their abuse-fighting efforts. Google has introduced new tools, tests, and disclosures to help people navigate content that is AI-generated. Since signing the Munich Accord, Google has developed features for: • **Ads disclosures**: Google's expanded political content policies now require advertisers to disclose when their election ads include synthetic content that inauthentically depicts real or realistic-looking people or events. This disclosure must be clear and conspicuous, and must be placed in a location where it is likely to be noticed by users. This policy will apply to image, video, and audio content. (note: Google's ads policies already prohibit the use of manipulated media to mislead people, like deep fakes or doctored content). • **Content labels:** YouTube's misinformation policies prohibit technically manipulated content that misleads users and could pose a serious risk of egregious harm. In March 2024, Google launched a new tool in YouTube's Creator Studio to enable creators to easily share when the content they're uploading is made with altered or synthetic media, including generative AI. Creators will be required to disclose this content when it's realistic, meaning that a viewer could easily mistake what's being shown with a real person, place, or event. Labels will then appear within the video description information, and if content is related to sensitive topics like health, news, elections, or finance, they will also display a label on the video itself in the player window. More information can be found here. • **A responsible approach to Generative AI products**: Google continues to prioritise testing including on metrics related to misinformation and fairness. Google will soon restrict the types of election-related queries in which Gemini (previously Bard) will return responses to users.

	• **Features providing users with additional context**: Google's About this image in Search feature helps people assess the credibility and context of images found online while Google's double-check feature in Gemini enables people to evaluate responses generated by Gemini Apps. • **Collaboration on best practices:** Alongside the Tech Accord and initiatives such as the White House commitments, Google has joined the Partnership on AI (PAI) Responsible Practices for Synthetic Media: A Framework for Collective Action, as part of the community of experts dedicated to fostering responsible practices in the development, creation, and sharing of media created with generative AI.
IBM	IBM has signed on to the Munich AI Elections Accord and continues to implement the processes in the Munich Accord, as well as the recent commitments from the AI Safety Summit in Seoul. • **IBM's policy on Deep Fakes and use in Elections:** to support the commitment to foster public awareness and media literacy in the Accord, IBM has published a policy on Deep Fakes outlining three key priorities for policymakers to mitigate with regard to protecting elections, creators, and people's privacy. • **Watsonx.governance:** IBM has invested and expanded its watsonx.governance toolkit for generative AI and machine learning. Watsonx.governance enables an audit trail capability to help other enterprises design and fine-tune LLMs and includes features like a foundation model library, tools for data preparation, model development, and monitoring. • **IBM AI 360 tools:** In 2018, IBM was the first to launch a free library of bias mitigation algorithms, the AI Fairness 360 toolkit, AI Explainability 360 incorporates bias mitigation and explainability into its own products as well as those for IBM customers. These are technical open-source tools for developers to create and test trustworthy AI. These features are embedded in watsonx.governance and continue to be strengthened. • **AI Ethics Leader and AI Ethics Board:** IBM's internal governance mechanism reviews technology use cases, promotes best practices, and conducts internal education. The IBM Ethics Board was central in IBM's decision to sunset its general purpose facial recognition and analysis products, considering the risk posed by the technology and the societal debate around its use. • **Detecting AI:** Years before generative AI became a household phrase, IBM Research and Harvard helped develop one of the first AI-text detectors, GLTR. This works by inspecting the visual footprint of automatically generated text enabling a forensic analysis of how likely an automatic system has generated a text. • **AI Impact Assessments (AIAs):** IBM uses and encourages broader industry and government use of AIAs. These are designed to aid internal actors seeking to make decisions within the broader governance established by risk management frameworks, taking into account measures to assess transparency, explainability, robustness, privacy, and bias.
LinkedIn	In May, LinkedIn Adopted the C2PA Content Credentials in support of the Munich AI Elections. A visible "Cr" icon will be labelled on human- or AI-generated content of images and videos that contain C2PA metadata. Clicking on the "Cr" icon will enable users to trace the origin of AI-created media, including the source and history of the content, and whether it was created or edited by AI. This feature is enabled in LinkedIn's feed and will expand to additional surfaces including ads.
Microsoft	Microsoft continues to advance responsible AI practices keeping people and their goals at the centre of the design process, and considering benefits and harms AI systems can have on society.

Microsoft recently published its first annual Responsible AI Transparency Report (May, 2024), taking a step beyond the White House Voluntary Commitments and the Munich Accords. The report shares Microsoft's maturing practices on responsible AI, provides insight into how Microsoft builds applications that use generative AI, and how Microsoft makes decisions and oversees deployment. Since signing the Accord, Microsoft has launched several technical customers can use to develop generative AI applications more responsibly, this includes:

- **Microsoft Elections Communications Hub**: Microsoft has created "Election Communications Hubs" to support democratic governments around the world as they build secure and resilient election processes. This hub provides election authorities with access to Microsoft security and support teams in the days and weeks leading up to their election, allowing them to reach out and get swift support if they run into any major security challenges. These hubs build on existing security programs such as the Azure for Elections offering available to state and local election agencies and their partners in the U.S.
- **Microsoft Content Integrity Check Tool:** Launched in March, Content Integrity Check helps political candidates, campaigners, and elections organisations to maintain greater control over their content and likeness. This includes (i) a tool to certify content by adding Content credentials and (ii) tools to enable the public to check if a piece of digital content has Content Credentials and get facts about digital content and its origins.
- **Provenance Capabilities:** Microsoft continues to build provenance capabilities into our products, including in DALL-E 3 series models hosted in Azure OpenAI Service.
- **Prompt Shield:** Prompt shield is a jailbreak risk detection model called prompt shield. Prompt shield recognises four different kinds of user prompt injection attacks (UPIAs): changes to application rules, embedded conversations for model confusion, role play, and encoding attacks. It is now expanded to include protections against indirect prompt injection, where generative application processes malicious information not authorised by developers or users.
- **Safety evaluations in Azure AI Studio:** Released in March, Safety Evaluations works to help organisations evaluate AI-generated outputs for content risks including hate, violence, sexual, and self-harm content, as well as content that may cause unfairness. This builds on Microsoft's existing Responsible AI Toolbox (RAI Toolbox), which brings together assessment tools like model interpretability, error analysis, data explorations, and causal analysis through different stages of model debugging and decision-making.
- **Risks & safety monitoring in Azure OpenAI:** Released in March, Microsoft's new feature helps to detect the misuse of generative applications through real-time harmful content detection and mitigation, with insights into content filter performance on actual customer traffic and identifying users who may be abusing a generative AI application. The user detection feature analyses trends in user behaviour and flags content to generate reports for customers to decide whether to take further action in Azure AI Studio through an abuse report where suspected.
- **PyRIT (Python Risk Identification Tool for Generative AI):** Released in February, PyRIT is an open-source red teaming accelerator, enabling security professionals and machine learning engineers to proactively find risks in their generative AI applications. Since its release on GitHub, PyRIT has received 1,100 stars and copied more than 200 times by developers for use.
- **Continued improvement on tools for information integrity:** Microsoft regularly undertakes third-party evaluations on its products and improves features and functionality to improve prompt blocking, content filters, and integrates safety

	system mitigations. Microsoft has launched 30 responsible AI tools that include more than 100 features to support responsible AI deployment, building on Microsoft's Responsible AI Standard and the NIST Risk Management Framework. • **Transparency notes:** Microsoft has also published 33 Transparency notes since 2019, to provide customers with detailed information about their platform services like Azure OpenAI Service. • **True Media Research Partnership:** In April, Microsoft also partnered with AI researcher, Oren Etzioni and his new non-profit, True Media. True Media provides governments, civil society and journalists with access to free tools that enable them to check whether an image or video was AI-generated and/or manipulated. Microsoft's contribution includes providing True Media with access to Microsoft classifiers, tools, personnel, and data. These contributions will enable True Media to train AI detection models, share relevant data, evaluate and refine new detection models as well as provide feedback on quality and classification methodologies.

Question 3:

Senator SHOEBRIDGE: Can you name a significant player that has put in place mandatory watermarking for their generative AI produced material?

Mr Black: I can come back again with a full list. I know Adobe, for example, through its Firefly platform, has been a leader in the watermarking space, to identify content that has been AI generated or modified. I'm happy to come back and provide a more fulsome answer to that question.

TCA supports the use of watermarking as a method for identifying AI-generated content. Watermarking allows for information to be embedded directly into content, even when an image undergoes some modifications. AI labelling and watermarking helps users know when an AI system is being used or where content has been generated, modified or informed by AI.

It is essential to recognise that watermarking is not a silver bullet or panacea for misinformation or deepfakes. It should be one tool, within a broader set of tools, that will enable greater transparency for AI-generated or AI-modified outputs. In adopting watermarking, it will also be important to counter unintended consequences from false positives or misuse, like when false watermarks are added by bad actors. It will be important to combine multiple methods, including the use of authentication and verification technologies, while also uplifting public capacity for critical digital literacy and awareness.

Below is a table of TCA members who have put in place mandatory watermarking for generative AI produced material.

Adobe	Adobe's text-to-image generative AI tool, Adobe Firefly, includes default content credentials that are automatically attached. The watermark is applied to downloads or exports of AI-generated content. Content Credentials are tamper-evident metadata that can be applied to creator assets at export or download. They increase transparency around the origins and history of the assets they are applied to. Adobe customers can use Firefly-generated content with confidence for commercial purposes because Adobe uses a multi-layered, continuous review and moderation approach to block and remove content that violates Adobe's policies. Adobe also offers customers IP indemnification for Firefly-generated content.
Amazon	All images generated by Amazon's recently announced Titan Image Generator will contain an invisible watermark that can be validated by an API. Built-in watermarking is designed to help reduce the spread of deceptive content and disinformation by providing mechanisms to identify AI-generated images. Amazon's terms of service prohibit the alteration or removal of the watermark. By making the image validation API publicly available, Amazon's aim is to make it easier for organisations to quickly assess whether an image has been generated or augmented by Titan Image Generator. Amazon is among the first model providers to widely release built-in invisible watermarks that are integrated into the image outputs and designed to be resistant to alteration. Amazon has also implemented a mix of training data and input filters to help prevent the generation of images of public figures, to further mitigate the risk that Titan Image Generator might be misused for deepfakes or other deceptive content. Amazon engages in rigorous testing, assessments, and improvement of its services. Amazon has also implemented important safeguards into its generative AI offerings, including (i) combatting risks around synthetic content by offering built-in watermarking with Titan Image Generator, (ii) applying filters on user inputs and model outputs for Titan models to reduce the likelihood and spread of harmful content; and (iii) automated abuse detection in Amazon Bedrock to detect and block potentially harmful content. Amazon makes its watermark validation tool widely available through Amazon Bedrock, which helps confirm whether a given image was generated by Titan Image Generator. Amazon's Bedrock Guardrails also help customers implement safeguards and responsible AI policies by providing content filters with configurable thresholds.
Google	In February, Google announced that all AI-generated content from Google's latest image model, Imagen 2, will be marked by SynthID, a state-of-the-art watermarking

	technology tool developed by Google DeepMind in refined in partnership with Google Research. SynthID embeds digital watermarks directly into the pixels of AI-generated images, audio, text or video in Google's consumer products. It can also scan content to detect digital watermarking to identify if an image, or part of an image, was generated by Google's AI tools through the 'About this image' feature in Search or Chrome. Google continues to invest in new techniques to improve the safety and privacy protections of its models. For more information click here.
LinkedIn	As above, LinkedIn has adopted the C2PA Content Credentials as a mandatory measure for AI-generated content of images or videos that contain C2PA metadata.
Microsoft	Content Credentials are automatically added to all images created with Microsoft's most popular consumer-facing AI image generation tools, including Bing Image Creator, Microsoft Designer, Copilot, as well as in Microsoft's enterprise API image generation tools via Azure OpenAI. Content Credentials are cryptographically signed and sealed as part of the image file's metadata; this information is tamper-evident and can be examined with tools such as Content Authenticity Initiative's open source Verify Tool. Microsoft is also now piloting a tool called "Content Integrity Certify" that allows users to add content credentials to their own authentic content in an easy-to-use tool. Microsoft has additionally released a public tool called "Content Integrity Check" that allows any member of the public or the media to check for the existence of content credentials and see provenance details.