Safe and responsible AI in Australia

"Now there is no question in my mind that artificial intelligence needs to be regulated. It is too important not to. The only question is how to approach it."

Sundar Pichai, Chief Executive Officer and Executive Director, Google and Alphabet
January 20, 2020

Executive Summary

Artificial Intelligence (AI) has the potential to unlock significant economic, scientific and social progress for nations like Australia and for the wider global community.

It is already deepening our scientific understanding of the world around us, which will enable dramatic progress in human health and wellbeing, and let us better tackle pressing global challenges like climate change. It underpins the democratisation of powerful creative tools that will profoundly increase human productivity, and support an economic transformation that may affect our societies as deeply as electrification or the creation of smartphones and global communications networks.

As with any powerful emerging technology, AI also presents challenges and must be developed with the utmost care. Industry, researchers, stakeholders in civil society, and governments must work together to ensure AI applications are trustworthy, and live up to AI's promised societal benefit while mitigating risk.

To this end, Google supports a policy agenda oriented around three pillars: unlocking opportunity, promoting responsibility, and enhancing global security.

It is our position that regulation of AI can be consistent with innovation. Pro-innovation policies include proportionate, risk-based measures to address the threat of harms. They can ensure Australians are able to access information about when an AI system is being used to make a decision about them and find plain-English descriptions about how those AI systems work, thereby increasing public trust and accountability. Such policies can encourage scientific excellence and high standards for representation and privacy, and consistent best practices for dataset testing and documentation. And they can support the creation of legal frameworks that encourage the domestic innovation and international interoperability that will secure Australia's place in the global AI ecosystem.

This submission provides detailed policy recommendations that aim to support the Australian Government's ambitions to create a risk-based regulatory framework, enable transparency and accountability, and ensure that Australia maximises its local capability and secures a significant place in the global AI ecosystem.

Google Australia

Every day, Google helps millions of Australians and Australian businesses to harness the benefits of technology to communicate, collaborate and find the information they need. For more than 20 years, Google Australia's employees have provided research and development services for Google's innovative products, helping improve the internet for the benefit of billions of users around the world, including millions of users in Australia.

Australian businesses gain $47.1 billion worth of economic value each year through Google Ads, AdSense, Play, Ad Grants, Search, Maps, and Cloud, and Australians as individuals gain $19.5 billion worth of annual consumer benefit. In the 2022 financial year, we invested over AU$1 billion dollars in our Australian operations and employed more than 2,000 people, a large portion of whom work in our engineering division.

Those engineers work on a diverse range of products including Google Photos and new technologies for internet users in Australia and around the world. In fact, an Australian startup invented the technology that became Google Maps. Google Australia's engineers also work on maintaining Google's global infrastructure and core systems.

In 2021, Google announced the Digital Future Initiative (DFI), a $1 billion investment in Australia over five years focused on infrastructure, a new AI-focused research centre and additional research partnerships, which it was independently estimated would provide a $1.3 billion boost to Australia's GDP and support 6,500 additional jobs across the economy.

This investment in national capability has since led to the launch of our first ever local AI research hub, partnerships in Quantum Computing research with Australian universities, and a growing number of AI-driven initiatives designed to help address Australia's greatest needs including the Blue Ocean Carbon project with CSIRO and the Department of Foreign Affairs and Trade (DFAT) to help solve climate challenges, a digital Career Certificate program to upskill Australians in critical digital skills, and an initiative with Cochlear and Australian audiology researchers to develop the next generation of AI-powered hearing technology.

The DFI will continue to deliver world-class digital technology, skills and partnerships for Australian science and industry over the coming years, supported by our investments in local infrastructure that enable Australian institutions to access the world's most advanced hardware and digital systems that Google can provide and that Australia's legal framework enables, including AI.

Why AI is important for Australia

AI is a technology of global structural importance, and therefore vital to Australia's national interest, economic development, trading relationships and supply chains. McKinsey's 2018 analysis of evidence relating to AI implementation suggested that AI could deliver an additional global economic output of about $13 trillion by 2030, boosting global GDP by about 1.2% a year. Within the spectrum of AI models, research on generative AI by McKinsey in 2023 indicates that generative AI alone could have a productivity dividend worth trillions to the global economy, estimating it could add the equivalent of $2.6 to $4.4 trillion annually to global GDP.

We have witnessed a dramatic rise in the importance of technology investment to Australia's economic capacity and to the standard of living of its citizens. Australia's technology ecosystem has grown rapidly, with its domestic technology industry scaling rapidly and traditional industries investing in their own digital transformation, supported by global communications and computational infrastructure and platforms. Today, technology is Australia's 3rd biggest industry, behind mining and banking.

Research conducted by the Technology Council of Australia (TCA) shows that Australia is on track to have 1.2 million tech workers by 2030, with the workforce standing at 935,000 as of February 2023. This is the product of strong growth with an 8% increase in tech jobs – double the growth in all other jobs – in the past year. There are now more software and application programmers in Australia than plumbers, hairdressers or secondary school teachers.

Shaping the future of hearing technology from Australia

Google's Hearing Initiative

There are 1.5 billion people globally who have some form of hearing loss. For many decades, Australia has led the way in building more accessible hearing technology. The cochlear implant was developed here and has gone on to become the gold standard for hearing clinical protocols, diagnostics and treatments for people living with severe to profound sensorineural hearing loss.

As part of our Digital Future Initiative, Google's AI experts are working with Australian leaders in the field of hearing to explore new possibilities and AI solutions for hearing healthcare. This collaboration involves five organisations across healthcare service delivery, research and technology sectors; including Cochlear, Macquarie University Hearing, National Acoustic Laboratories (NAL), NextSense and The Shepherd Centre.

Together, we'll be focused on new applications of AI and machine learning to develop listening and communications technologies, overcome current challenges – and pave the way for more customised and effective hearing healthcare.

Barriers to Australian Innovation

We applaud the Australian Government's interest in ensuring Australia does not miss out on the next wave of innovation and value creation enabled by AI, as we are already seeing the impact of a lack of allowances under Australian copyright legislation for AI research and open source innovation more broadly.

Legal uncertainty creating barriers to investment and talent loss
The legal uncertainty for certain AI research and computation activities created by the current Australian copyright framework is already impeding our ability to build Google's AI research capacity and investment in Australia, compared to other more innovation-friendly legal environments fostered by nations like the United States, with its fair use protections, or Singapore which updated its copyright law in 2021 to include exceptions to copyright for the purpose of computational data analysis, including training machine learning systems. In our APAC region, countries such as Singapore and Japan have established policies designed to create legal certainty and enabling IP frameworks to attract investment as regional AI innovation hubs.

We have entered a period when general institutional knowledge about AI systems is improving, and across our society there is renewed excitement about the democratisation of technology being driven by public access to new language and image-based generative AI systems supported by foundational models. This excitement is broadening interest in AI research and commercialisation. In the context of a fast-scaling Australian technology industry, there is already an emergent risk to Australia's ability to secure innovation talent (both local and from overseas) and investment from overseas.

As industry's understanding of the legal risks of developing these kinds of technologies in Australia grows, particularly related to the use of data from the open web, we can foresee a loss of local talent and investment that would otherwise have scaled from domestic industries, creating Australian jobs, investment and IP..

This is the core risk to Australian talent, investment and local development capability we would encourage the Government to address in its technology governance framework.

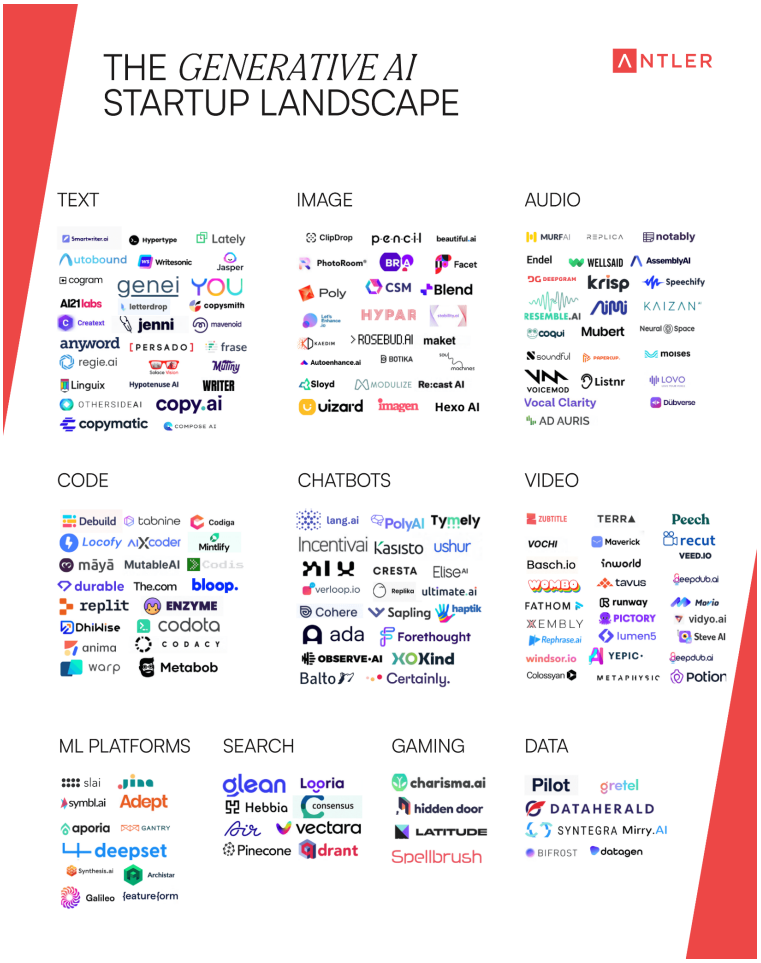
Data governance
We also see risks emerging in data governance. Just as imports support critical elements of Australia's economy in other domains, from transport and logistics to construction and consumer retail, Australia's technology ecosystem relies on global computational resources and technology infrastructure investment to support local capability and development of systems that serve Australian interests.

As innovation in computational capability at the device, infrastructure and network level continues at pace, and due to the fact many of these innovations are initially deployed in offshore computation clusters, it remains in Australia's core interest that institutions are able to rely on global computation to support research, development and productisation of AI systems and services designed to serve the interests of Australians.

Policies that restrict Australian organisations from training Australian AI models using offshore computational resources, or from using certain kinds of data for model training for input into global systems (for example in areas like cybersecurity protection or medical systems) will raise costs for Australian AI innovators and may even prevent global AI systems from being tailored to Australian needs and circumstances. Any policy that lessens Australian technologists' ability to customise AI applications for Australians could have a negative impact on the Australian public, including hampering our ability to reduce bias in AI applications across communities in Australia.

An Increasingly Competitive Ecosystem

The AI market is open and dynamic, with rapidly lowering barriers in both upstream building and downstream deployment and lower economies of scale than many other industries, leading to a particularly strong dynamism in application development. There is an increasing number of new entrants and more established app developers releasing a continuous stream of apps and tools for a wide variety of use cases. The following diagram provides a snapshot of a non-exhaustive sample of apps already available to consumers and businesses in the field of generative AI alone.



Source: Antler Gen AI Report

Pro-Innovation Regulation

Google aims to support the Government's work on safe and responsible AI by providing recommendations for a robust regulatory framework. Effective regulating of this technology is vital to underpin public confidence and spur innovation. This will provide a framework for fairness, safety and security to underpin Australian AI innovation and support Australia's success in the global technology and services export ecosystems. Our recommendations are focused on development of a proportionate, risk-based framework.

A Proportionate, Risk-Based Framework

Google supports a risk-based approach to AI regulation, as proposed in Box 4 of the Discussion Paper. We support the Government's focus on context-specific risks of AI, targeted at the right use cases, and differentiated requirements and obligations depending on the assessed risk-level. We also support the proposed approach of allowing AI to be used in high-risk settings where the risks and costs are justified and can be explained. Where AI legislation is proposed, it is important that such legislation be developed in close consultation with the broader Australian tech ecosystem, including industry, academics and civil society.

(i) Risk/impact assessments

Risk/impact assessments are essential to ensure that potential risks are mitigated while preserving a pro-innovation regulatory environment. It is our long-held view that a risk-based framework must take into account the likelihood of harm alongside the severity of harm, as well as considering *the cost of not using AI* in terms of forgone benefits (ie. the cost of lost opportunities).

We recommend that Government consider the following additional factors:

- **Severity and likelihood of harm.** Conventional approaches to assessing risk take into account the severity of harm compared against the likelihood of its occurrence. Severity can be sorted into categories such as "catastrophic", "major", "moderate", "minor" and "negligible"; and the probability of an adverse effect similarly as "very likely", "likely", "possible", "unlikely" and "very unlikely". Adding consideration of probability of harm to the Government's draft risk management approach is recommended because it allows for various combinations of severity / likelihood to qualify as high-risk (e.g. not only "major / likely" but also "catastrophic / very unlikely", "minor / very likely"). This may also be important to inform prioritisation of assessment workstreams in a resource-constrained environment. Clearly reflecting that severity and likelihood are both important in managing AI risk will encourage organisations developing and deploying these technologies to mitigate the severity of harm while simultaneously reducing its likelihood.

- **Guidance on when risk classification changes**. Beyond the three levels of risk (low, medium, high) as reflected in Box 4, we recommend that any regulation also include guidance on specific thresholds when the risk classification of a given AI application changes (for example, reaching a larger number of people or being used economically

critical processes), and reflect that the goal is to mitigate the severity of harm while also reducing its likelihood.

- **How the attributes of a particular AI system impact its overall risk and benefit.** Specific design features and operational constraints and mitigations may reduce or increase overall risk. Factors that could affect the overall calculation of risk include:
    - Consistency across user groups and / or operating environments;
    - Reversibility of errors made by the system and the degree and reliability of human control;
    - The existence of continuous learning safeguards (for example, a limit on deviation from a predicted baseline model); and
    - The presence of and effectiveness of organisational governance within the organisations developing and deploying an AI system.
- **How the overall risk of an AI system compares to existing alternatives.** An assessment should acknowledge the opportunity costs of not using AI in a specific situation or of intentionally developing AI without particular capabilities. The risks and benefits of AI systems should be weighed against existing (non-AI) approaches, including human judgment. If an imperfect AI system is shown to perform better than the status quo at a crucial life-saving task, for example, it may be irresponsible not to use the AI system. Where the alternative of not using AI poses a greater risk than using AI, AI deployment should be supported, given the net benefit to society.
- **Treatment of R&D or early stage products.** In the early stages of development there will often not be a clear view as to the ultimate shape of a product (indeed it may not even be clear what is technically feasible), and thus it is not possible to thoroughly assess risks or what consultations may be necessary to address such risks until a later stage. It is therefore important that confidential piloting of an AI application be allowed prior to any risk assessment, within the bounds set by existing regulation. If such pre-assessment testing is not permitted, it may result in organisations being unable to accurately assess risks and therefore having to take an unduly precautionary stance in terms of the necessary requirements and investment, which would hinder innovation.
- **Treatment of products which receive significant updates.** Carrying out a new risk assessment should be required when there has been a significant change to the functionality of the product that is likely to materially alter its performance in testing or safety disclosures. However, generic over-the-air updates (OTAs) such as security patches, bug fixes, or simple operational improvements after placing a product on the market should not trigger a renewed risk assessment. Potential determinants of whether a modification should spark a new assessment could include a significant alteration in the training data or model or a change in external factors related to model efficacy benchmarking (e.g. if medical authorities altered the test required for a specific diagnosis).
- **Clarity on who is responsible for risk/impact assessments**. We recommend that the responsibility for risk/impact assessments lie with the deployers of AI tools, who are best positioned to undertake risk assessments for a specific use because only they know the

context of that use. A "deployer" can be defined as "an entity that puts into service an AI system developed by another entity without substantial modification." While developers of open source or off-the-shelf, multipurpose AI component systems can and should provide general information about their systems (including guidance on operating boundaries and model behavior in the context of unforeseen inputs or user entries), they should not be held responsible for conducting deployment risk assessments or validation, as they are unlikely to be well positioned to verify a system's end uses. A practical approach for the Government could be to provide procedural "due diligence" guidance but assign responsibility for conducting and documenting risk assurance exercises to the front-line organisation deploying an AI application, with the resulting documentation available to regulators should public deployment cause harms. Post-deployment, if concerns arose that an application had been misclassified, remedial action could be taken.

- **Impact assessment for high-risk AI systems.** We support the Government's suggestion for impact assessments to be peer reviewed by external experts. We also recommend regulatory disclosure of capacity and risk assessments, with protection for trade secrets and controlled technologies as appropriate.

(ii) Notices

We support the proposal in Box 4 to waive notice requirements for low risk applications and plain language notice requirements for medium risk applications. This will reduce regulatory friction, encourage innovation and lower compliance and business costs.

On the requirement for high-risk applications to publish system explanations, we recommend that such explanations be clearly defined and based on what is practicable for companies to publish. We recommend the following elements be included as part of AI notices and accountability disclosures:

1. **Topline indication of how the AI system works** – Organisations should not be expected to reveal full details about AI models or underlying code, including due to risks to business confidentiality and the potential for adversarial gaming of the system. That said, model deployers should be expected to detail general logic and assumptions that underpin an AI application, particularly if it is designed for use in high-risk settings. It is also good practice to highlight the inputs that are typically the most significant influences on output, and any inputs likely to be deemed sensitive or unexpected. Any inputs that were excluded that might otherwise have been reasonably expected to have been included (e.g., efforts made to exclude gender or race) should also be noted.

2. **Expectations about how an AI system will be used** – When relevant, developers should clarify whether any operational constraints were intended in deploying an AI technology, such as whether the tool was designed to function independently or with a level of human oversight. There is evidence that users interact with AI systems and react to errors differently depending on such assumptions, so this information will help deployers build suitable mental models when using an AI application. While it is impossible to anticipate every possible use of an AI system, developers can state the intended use case for the model or system (e.g. those use cases against which its performance was tested and/or for which it is being marketed).

3. **Known limitations on performance** – It will often be hard to describe in lay terms a model's expected limitations or level of accuracy when operating under changing conditions, but general guidance can still be given. Research has shown that an AI application's performance is best contextualised by presenting it alongside existing human performance statistics, where they exist. Concrete examples of successful and unsuccessful use cases are also helpful, particularly any challenging edge cases or known pitfalls regarding existing non-AI approaches the system has been explicitly designed to overcome.

4. **Where appropriate, additional technical information about AI system performance for expert users and reviewers** such as consumer protection bodies and regulators. This could include information about:

    - How well the AI system performs against industry-standard evaluation datasets measured against key metrics (e.g. bias and fairness);
    - The frequency and cost weighting assigned to different errors (e.g. false negatives or false positives);
    - How the AI system's performance compares to existing human-performance benchmarks; and
    - Whether there may be any conceivable use cases enabling "dangerous capabilities" like cyber-offense, media manipulation, and weapons acquisition.

(iii) <u>Human in the loop/oversight assessments</u>

We support the Government's approach of meaningful human intervention and the recognition that there are circumstances where human in the loop requirements may not be appropriate. Forms of oversight that are common sense in one setting could be harmful and undermine the core value of an AI application in another.

For example, requiring an AI system's output to be reviewed by a person before being actioned may make sense for some applications (e.g. AI systems used for critical, non-time-sensitive medical diagnostics).

But we may overestimate human accuracy, consistency and lack of bias. For some applications this may lead to sluggish output, reduced privacy (if it means more people see sensitive data), or undermined accuracy (if human reviewers lacked the necessary expertise or were shaping output with their own biases). At an extreme, it could even put people at risk, for example, by delaying automated safety overrides.

(iv) <u>Explanation</u>

The Government's current proposal recommends that specific explanations of decision, AI output of application be made to users for medium risk applications, and made available publicly or to experts and regulators for high risk applications. Explainable AI can help build trust and confidence by helping in the identification of harms, empowering users to make informed choices and holding AI providers accountable. However, it's equally important to ensure that the standards imposed for explainability should be workable, and recognise the practical constraints and tradeoffs to ensure they enable innovation.

Explanations can be costly in terms of technical resources. There are also trade-offs with other goals like model accuracy (e.g. if more accurate but harder-to-explain techniques must be foregone). If

every outcome of an AI system were mandated to be fully traceable and supported by a detailed explanation — a far higher standard than any human-based system can meet — it would in practice restrict AI systems to an extremely limited, basic set of techniques (e.g. static decision trees). This outcome would dramatically undermine AI's social and economic benefits.

In addition, tailoring explanations to be meaningful and suit the needs of a range of audiences is difficult and time intensive. The ability to trace back and explain outcomes from AI systems operating at scale on a daily basis will likely differ greatly from the more extensive probing possible during development and upfront testing. While there has been much progress in tools to support developers (such as Google's Explainable AI tool for Cloud AI customers) providing explanations at scale and in real time remains challenging, in part because the detail and scope of what is needed varies significantly by sector and audience, and expectations may evolve as best practices emerge.

(v) Monitoring and documentation

In common with other complex software systems, AI systems are often directly enabled by the work undertaken to develop and document high quality datasets. Data excellence likewise underpins many of the solutions to identifying and addressing challenges such as bias. Internal governance systems that include robust monitoring and documentation of datasets can help support accountability.

For its part, Google has developed template documentation tools known as data and model cards, that are used to simplify and standardise information about an AI model or its underlying dataset(s).

- **Model cards** are short documents accompanying trained machine learning models that typically include information such as the model's intended use case, the data used to train the model, the model's performance on different metrics, any known biases or limitations of the model, and any potential risks or unintended consequences that could arise from its use. Model cards can also include information about the model's training and evaluation processes and how the model can be deployed and integrated into different applications.

- **Data cards** are a dataset documentation framework aimed at increasing transparency across dataset lifecycles. They provide structured summaries of ML datasets with explanations of processes and rationale that shape the data and describe how the data may be used to train or evaluate models. At a minimum, data cards include the following: (a) upstream sources, (b) data collection and annotation methods, (c) training and evaluation methods, (d) intended use, and (e) decisions affecting model performance.

Regarding the Government's proposal to require external audit of internal monitoring and documentation for high-risk applications, we flag that any independent external audit should be aligned to international, industry-accepted criteria to ensure consistency. Further, independent auditors would need to be professionally qualified and entrusted to only certify organisations that meet the appropriate standards. Providers of such services lack accountability absent standards or government guidance. These assessments should be carried out by qualified individuals who are themselves accountable for the quality of their work, in line with the longstanding approaches to privacy, security, financial and other types of audits.

The Government would need to balance such audit requirements against the risk of creating security vulnerabilities, exposing trade secrets and confidential information, or hindering innovation or the development of useful applications. Compounding this challenge is the current lack of

consensus on technical standards and responsible practices for developing and deploying AI, including that for monitoring and documentation.

How Google Manages Risk

Google embeds ethical AI risk management at the core of our products and services (e.g. Pixel devices, Google Bard and Google Shopping) from the design stage. Internally, Google uses an AI risk-assessment framework (AI RAF) alongside a Product Maturity Model Assessment — stewarded by our Responsible Innovation and Responsible AI and Human-Centered Technology teams, which focus on the sociotechnical realisation of our AI practices — to ensure that internal development and deployment practices are consistently aligned with our AI Principles and our broader compliance ecosystem, including the Civil and Human Rights Program. Outputs from these assessments are mapped to a prescriptive maturity model framework, that details clear actions developer teams can take to improve their machine learning (ML) models and advance from one maturity level to the next.

Google's work in responsible innovation

Our approach to responsible AI innovation starts early, before teams plan a new AI application. When a team starts to build a machine learning (ML) model, dataset or product feature, they can attend office hours with experts to ask questions and engage in analyses using responsible AI tools that Google develops, or seek adversarial proactive fairness (ProFair) testing. Pre-launch, a team then can request an AI Principles review.

1. Intake

Any team can request AI Principles advice. Reviewers also consider an ongoing pipeline of new AI research papers, product ideas, and other projects.

2. Analysis

Reviewers analyze the scale and scope of a technology's potential benefits and harms.

3. Adjustment

Reviewers recommend technical evaluations (e.g., checking for unfair bias in ML models).

4. Decision

Reviewers decide whether to pursue or not pursue the AI application under review (e.g., Cloud AI Hub and text-to-speech).

AI Principles reviewers are in place to implement a structured assessment to identify, measure and analyse potential risk of harm. The risk rating focuses on the extent to which people and society may be impacted if solutions did not exist or were to fail. Reviewers also consider a growing body of lessons from thousands of previous AI Principles reviews conducted since 2019.

To date, more than 32,000 employees across Google have engaged in AI Principles training. Given our growing understanding of effective hybrid and remote learning, we continue to expand and modify the courses. For example, this year we adapted our popular four-part Tech Ethics self-study course to a one-part deep dive based on internal feedback. Similarly, we launched the Responsible Innovation Challenge — taken by more than 13,000 employees — as a series of engaging online puzzles, quizzes and games to raise awareness of the AI Principles and measure employees' retention of ethical concepts, such as avoiding unfair bias.

Last year we also launched a new internal program for senior managers and leaders: the Executive AI Principles Ethics Fellowship. This program includes educational workshops and training on the

AI Principles for leaders across multiple product areas and geographies. It's based on our six-month AI Principles Ethics Fellowship, which we launched in 2020 and which has trained a diverse set of 50 employees from across 17 global offices and 15+ Employee Resource Groups to learn about responsible AI and contribute their perspectives to Google's AIP operations. During the fellowship, among other duties, fellows develop fictional, future scenarios for AI ethics challenges, addressing topics such as deep fakes and misinformation.

Their hypothetical scenarios supplement a growing body of responsible innovation case studies that our AI ethics review teams draw upon as references when making decisions. The new Executive AI Principles Ethics Fellowship is tailored to business decision makers' needs. The inaugural cohort consisted of sixteen executives across ten product areas, including Cloud, Devices and Services (hardware), and YouTube.

Google AI Principles

We will assess AI in view of the following objectives. We believe AI should:

1. **Be socially beneficial:** With the likely benefit to people and society substantially exceeding the foreseeable risks and downsides.
2. **Avoid creating or reinforcing unfair bias:** Avoiding unjust impacts on people, particularly those related to sensitive characteristics such as race, ethnicity, gender, nationality, income, sexual orientation, ability and political or religious belief.
3. **Be built and tested for safety:** Designed to be appropriately cautious and in accordance with best practices in AI safety research, including testing in constrained environments and monitoring as appropriate.
4. **Be accountable to people:** Providing appropriate opportunities for feedback, relevant explanations and appeal, and subject to appropriate human direction and control.
5. **Incorporate privacy design principles:** Encouraging architectures with privacy safeguards, and providing appropriate transparency and control over the use of data.
6. **Uphold high standards of scientific excellence:** Technology innovation is rooted in the scientific method and a commitment to open inquiry, intellectual rigor, integrity and collaboration.
7. **Be made available for uses that accord with these principles:** We will work to limit potentially harmful or abusive applications.

In addition to the above objectives, we will not design or deploy AI in the following application areas:

1. Technologies that cause or are likely to cause overall harm. Where there is a material risk of harm, we will proceed only where we believe that the benefits substantially outweigh the risks, and will incorporate appropriate safety constraints.
2. Weapons or other technologies whose principal purpose or implementation is to cause or directly facilitate injury to people.
3. Technologies that gather or use information for surveillance violating internationally accepted norms.
4. Technologies whose purpose contravenes widely accepted principles of international law and human rights.

You can find more information here, including annual updates on Google's AI Governance Operations.

Key Principles

We recommend that the Government also keep the following additional considerations in mind in the design and development of any AI regulation.

(i) Ensure parity in expectations between non-AI and AI systems

Like any system, including the human-based processes AI systems are often built on, AI systems are not perfect. They do, however, offer the potential to dramatically improve on current human-based decision making for certain use cases. The operational benchmark for AI systems should not be perfection, but instead the comparative performance of current processes (if existing) or an available human-powered alternative.

There is a real risk that innovative uses of AI could be precluded by demanding that AI systems meet a standard that far exceeds that required of non-AI approaches. Sometimes this may be deliberate to retain deep human capacity in a particular domain, but more often it is likely to be due to a lack of insight into, or tacit acceptance of, flaws in existing non-AI processes, and people's natural tendency to be more forgiving of mistakes made by a human vs a machine. To help offset this, Google would recommend that there should be parity in terms of expectations between AI and non-AI approaches, unless there is a clear justification as to why it should differ in a particular use case and context.

A key area in which this principle should apply is setting minimum performance standards. A sensible starting point would be to expect AI systems to match or exceed similar accuracy and fairness standards when compared to current approaches. There may however be good reasons to deviate from this. In some situations a lower level of accuracy may be acceptable — such as if an urgent response is needed, the cost of inaction is high, and there are simply not enough qualified people on hand to do the job (e.g. identifying potentially damaged infrastructure after an extreme weather event).

In other contexts the reverse may hold, such as if there are plenty of qualified people happy to do the work and using an AI system might only be justified if it was shown to perform significantly better (e.g. self driving cars that have far fewer accidents than human drivers). Similarly, fairness is a vital consideration. Even if an AI system performs more accurately and reliably across the general population, that may not be sufficient to justify its use if it performs significantly worse than existing approaches for certain subgroups.

A similar principle of comparison can also be applied to expectations of transparency and explainability for AI systems. However in doing so, it is important not to exaggerate the standards met by human-powered systems. There are many settings where an explanation is not required of human decision-makers, and even if an explanation is provided, there is no means of ensuring that it accurately represents the key factors influencing a person's decision (for example, a decision maker may opt to withhold mention of certain factors that influenced their decision, or their decision may be influenced by unconscious bias).

(ii) **Clear delineation of roles and responsibilities between AI developers and deployers**

We recommend there is a clear delineation of roles and responsibilities between AI developers and the organisations deploying particular use cases. In practice, this would apply to many AI applications deployed by third parties using open-source software or general purpose APIs.

The organisation deploying an AI application should be solely responsible for any disclosure and documentation requirements about the AI application because it is best positioned to identify potential uses of a particular application, monitor its performance and mitigate against misuse.

Even in cases where an application is provided by a developer directly to the deployer, and no modifications are made, deployers will often be best positioned to understand downstream use cases and their attendant risks, implement effective risk management strategies, and conduct post-market monitoring and logging, which developers of general use systems are not equipped to do.

For example, a deep fake detector API could be extremely beneficial to small or medium enterprises (SMEs) seeking to combat manipulated media. In itself, this is not inherently high-risk. But such a system might be used in a separate high risk context (for example, as part of a law enforcement operation) without the knowledge of the provider releasing the non-high-risk API, which would constitute a high-risk application of the technology.

(iii) **Consistency of requirements for Public and Private sector AI use cases**

Accountability mechanisms for AI systems should not be different for developers supplying the public sector or private sector. In both cases, developers should be expected to deliver to deployers: documentation, guidelines, and recommended practices to support implementation as they use or build their own systems, as well as guidance regarding acceptable uses of their systems.

Furthermore, as public agencies become deployers, they will be best positioned to understand how they have chosen to deploy the system and the attendant risks. AI tools are not appropriate for every use case, and use-case-specific risks are best assessed by those with the most close knowledge about how a system is used. Deployers should have a comprehensive risk management program in place to assist in evaluating whether the proposed AI tools are fit-for-purpose for their use case.

(iv) **Consideration should be given to regulatory trade-offs**

- **Trade-offs with model accuracy:** Explanations can demand technical resources or cause tradeoffs with other goals like model accuracy (e.g. if transparency requirements preclude the use of more accurate but harder-to-explain techniques). Transparency requirements that limit the functionality of AI would dramatically undermine AI's social and economic benefits. Moreover, transparency requirements need to be carefully designed to ensure they are advancing a legitimate objective.

- **Trade-offs with security:** Overly broad transparency requirements could make it easier for bad actors to spoof, manipulate, or exploit AI models. Fully open dissemination (i.e. open

sourcing) of AI systems without appropriate controls, protocols, and safeguards could lead to the release of potentially harmful AI capabilities. Today, abuse by malicious actors is limited by guardrails (e.g. refusal layers) built by developers to help mitigate against inappropriate uses. Full transparency or access to elements of frontier models may actually pose greater danger to at-risk individuals.

- **Trade-offs with privacy and consumer protection:** The law should provide flexibility to use the technology to protect consumers. For example, an adversary could use AI to create harmful content like scam celebrity endorsements for financial products. While facial recognition technologies can be used to significantly improve precision in detecting malign advertisements and preventing user harm, privacy safeguards in many jurisdictions prevent biometric data processing without express user consent, which is difficult to get at scale. As a result, such laws may effectively prohibit the use of AI tools to detect fraudulent behavior.

Legal Frameworks for AI Innovation

We recommend that the Government implement regulation and policies that help support AI innovation and responsible deployment. Generally speaking, addressing the following areas in a coherent and consistent cross-functional manner will be important to building confidence and enabling investment in Australia's national AI ecosystem:

- Competition safe harbors for open public-private and cross-industry collaboration on AI safety research.
- Proportional privacy laws that protect against the disclosure of private information and enable trusted data flows across national borders.
- A legal framework for the transformative use of data on the open web for training of AI models.
- Copyright systems that enable appropriate and fair use of copyright-protected content, while allowing publishers and content creators choice and control over the presentation of their works to the public.
- Clarity on potential liability for misuse or abuse of both general-purpose and specialised AI systems (including open-source systems, as appropriate) by various participants - researchers and authors, creators, implementers and end users.

(i) A hub-and-spoke approach that supports existing regulatory expertise

We support a hub-and-spoke approach to AI regulation which leans on existing regulatory expertise. At the national level, we recommend an agency with technical expertise like the Commonwealth Scientific and Industrial Research Organisation (CSIRO) is tapped to provide sectoral regulators overseeing AI implementation with technical advice. AI will present unique issues in financial services, health care, and other regulated industries and issue areas that will benefit from the expertise of regulators with experience in those sectors - which works better than a new regulatory agency promulgating and implementing upstream rules that are not adaptable to the diverse contexts in which AI is deployed.

As the Government's paper has identified, AI is already subject to a number of existing regulatory frameworks. We recommend as part of this process that the Government Issue detailed guidance

articulating how existing legislation applies to the use of AI and where improvements need to be made.

The focus on AI regulation should be on specific applications of AI — not the science of AI itself. There is an immense diversity of current and potential AI applications across almost all areas of society — healthcare, financial services, transportation, defence and many others. The use cases and their impact on people and organisations are not the same. Moreover, many "AI issues" are actually issues common to the operation of any existing human systems and other complex software already used by retailers, banks, insurance companies, hospitals and manufactures. Consequently, AI regulation is likely best addressed first through sectoral approaches that leverage existing regulatory expertise in specific domains, rather than one-size-fits-all approaches.

Sectoral experts are typically well-positioned to assess context-specific uses and effects of AI and to determine whether and how best to regulate them, although sometimes additional resources may be required, including technical AI expert capacity. For instance, health-focused agencies are best positioned to evaluate the use of AI in medical devices and energy regulators are best positioned to evaluate the use of AI in energy production and distribution. It will also be useful to have consistency in oversight and the expectations for human and machine actors performing the same task unless there are justifiable grounds for difference.

Sectoral regulators should update existing oversight and enforcement regimes to apply to AI systems, including clarifying how the oversight of existing authorities applies to the use of AI and how to demonstrate compliance of an AI system with existing regulations. For example, risk assessments based on international consensus and multi-stakeholder standards like the ISO 42001 series or the NIST AI RMF can demonstrate compliance of a medical device with FDA rules in the United States. Regulators should reference existing standards and frameworks and how these frameworks can be used to manage risk.

If action is needed, regulators should avoid duplication and speed implementation by expanding established due diligence and regulatory review processes to include AI. When an AI application is not obviously covered by existing regulations, clear guidance should be provided on the due diligence criteria companies should use in their development processes. This would enable robust upfront self-assessment and documentation of any risks and mitigation strategies, and could also enable further scrutiny after deployment.

(ii) International standards for AI

Several efforts are underway to establish internationally recognised standards for AI, including within ISO and IEEE, and industry-driven initiatives such as MLCommons. While these efforts can highlight key areas for attention, it is important they can evolve in line with the rapid developments in underlying AI technologies. Ultimately it is unlikely that a single set of standards will emerge to suit all circumstances: multiple families of standards are more apt. As in similar domains such as cybersecurity, regulators should avoid the temptation to "pick winners" and instead allow flexibility for the optimal standards approach to be chosen for each context.

Regulators should avoid prescribing technology-specific or overly prescriptive standards that cut across different domains of AI use cases - instead allowing for the development of broad standards that can be adapted for specific contexts and use cases.

(iii) International AI policy alignment

As noted in the Government's paper, as a relatively small, open economy, international harmonisation of Australia's governance framework will be important as it ultimately affects Australia's ability to take advantage of AI-enabled systems supplied on a global scale and foster the local growth of AI. Australian policymakers should continue to play an active role in international policy alignment, working with allies and partners to develop common approaches that reflect democratic values. Several steps can be taken, including:

- Supporting participation by Australian experts in international multi-stakeholder standards processes.
- Encouraging multinational adoption of common approaches to AI regulation and governance, supported by a common lexicon based on the work of the OECD.
- establishing effective mechanisms for information and best-practice sharing with allies and partners, as well as between governments and the private sector (for example, sharing information to identify actors engaging in economic espionage and attacks on AI systems). This approach has proven effective in the cybersecurity domain.
- Advocating for trusted data flows across national borders, ensuring that allies and partners do not restrict data flows between each other to ensure high-quality data and cost effective, carbon efficient computation is available for Australian AI modeling and training.
- Promoting copyright systems that enable appropriate and fair use of copyrighted content to enable the training of AI models in Australia on a broad and diverse range of data, while supporting workable opt-outs for entities that prefer their data not to be used in training AI systems.
- Support the creation of a multi-stakeholder partnership to establish joint strategies, consistent best practices and standardisation in AI threat testing, AI bug and bias bounties, and solutions like watermarking, metadata, and other techniques to combat AI-enabled mis/disinformation.

We encourage Australian policymakers to actively participate in the work of the OECD and the Global Partnership on AI, two fora that are emerging as international clearing houses for progress in AI governance, as well as the Global Network Initiative.

Regulatory Approach

This is a complex cross-economy issue, with interlinked digital and physical supply chains, global pools of expertise and research capacity, and rapidly advancing capabilities. There are many trade-offs, and detail is paramount for responsible implementation.

As with other technologies, there are new policy questions that arise with the use of AI, and governments and civil society groups worldwide have a key role to play in the AI governance discussion. No one company, country, or community has all the answers; on the contrary, it's crucial for policy stakeholders worldwide to engage in these conversations.

We appreciate the highly consultative approach taken by the Australian Government, building expertise in Government and a culture of collaboration across Australian and international industry, and we are committed to continuing to help Australia build a fit-for-purpose AI regulatory framework.

ENDS