A Selective Response to *"Safe and Responsible AI in Australia"*

Preamble: At the beginning of the "Safe and Responsible AI in Australia paper", the authors in defining Australia's AI Ethics Principles refer to principle one as "*Human, societal and environmental wellbeing: AI systems should benefit individuals, society and the environment." and principle seven as "Contestability: When an AI system significantly impacts a person, community, group or environment, there should be a timely process to allow people to challenge the use or outcomes of the AI system.".* Later they said, *"The focus of this paper is to identify potential gaps in the existing domestic governance landscape and any possible additional AI governance mechanisms to support the development and adoption of AI.".* At no point in the remaining text is the issue of environmental wellbeing or the impact of AI on the natural environment actually addressed explicitly or indirectly, regarding safety, responsibility and government regulation. My key feedback relates to this issue, as addressed in response 2A below. My remaining responses relate to other questions and issues that I felt were important to address.

Q2. What potential risks from AI are not covered by Australia's existing regulatory approaches? Do you have suggestions for possible regulatory action to mitigate these risks?

A. Risk: The direct and indirect harmful impact of AI on the natural environment, in particular:

- The risk associated with AI across its entire life cycle.
- The impact of each aspect of AI at different stages in the AI life cycle, on the overall state of the natural environment, ecosystems, ecosystem services and processes, waterways & oceans (hydrological systems), air, land, soil, microbiome health and diversity, plant and animal health and biodiversity.
- The immediate and cumulative harm of AI on the natural world.
- The general public and corporate and government impact of AI development, use and disposal on the natural world.
- The uptake of AI technologies across all sectors of life and its potential short term impact on Australia's natural environment with a current population of 25 million and its long term future impact on Australia's natural environment with a much larger population.
- The uptake of AI technologies across the human lifespan from infancy to old age and its immediate and long term impact on Australia's natural environment.

For an elaboration, please see the attached documents *"Creating An AI Environmental Act in Australia and AI Regulations in Australia"* and *"Why An AI Environment Act Is Necessary"*

Suggestions:

1. Create a federal AI Environment Act which specifically targets AI across its life cycle. In addition to this it maybe useful to create state specific AI Environment Acts which fall within the scope of the federal act.
2. Regulate AI in Resource Extraction: Stricter regulations could be imposed on the use of AI in resource extraction industries (like mining or logging) to prevent overexploitation and ensure sustainable practices.
3. Regulation of AI in Pollution Control: Implement strict regulations on the use of AI in industries known for pollution (such as manufacturing or power generation) to ensure that AI optimisation doesn't compromise emission standards.
4. Regulation of Energy Use of AI Systems: AI systems, especially large-scale machine learning models, can consume significant amounts of energy. Regulations could be put in place to encourage or mandate energy-efficient algorithms or limit the energy use of AI systems.
5. Promote AI in Environmental Protection: Regulatory incentives could be created to encourage the use of AI in environmental protection efforts, such as wildlife tracking, pollution detection, and forest health monitoring.
6. Regulate AI in Waste Management: Implement strict regulations on the use of AI in waste

management to ensure environmentally friendly waste disposal and recycling practices.

7. Review of Environmental Impact Assessments (EIAs): Regulations could mandate the inclusion of AI impact in EIAs, considering both the direct impact of AI operations and the indirect effects that may occur due to its application.
8. Laws related to environmental protection, resource extraction, pollution control, energy use, and waste management might need to be updated or amended to include the use of AI. These may include but not limited to:

    - Environment Protection and Biodiversity Conservation Act 1999: Amendments to this Act could include provisions for the use of AI in activities that may impact the environment and biodiversity.
    - National Greenhouse and Energy Reporting Act 2007: This Act could be updated to incorporate reporting on the energy usage of AI systems.
    - Waste Reduction and Recycling Act 2011: This Act could be updated to regulate the use of AI in waste management and recycling processes.
    - Offshore Petroleum and Greenhouse Gas Storage Act 2006: Amendments to this Act could include provisions for the use of AI in offshore petroleum extraction and greenhouse gas storage.

    Note: Whilst these changes would be advantageous, an AI specific environment act might be more useful in addressing a comprehensive range of issues.

(See attached documents for more details on this entire subject.)

B. Risk: The use of AI in biotechnology, in particular:

- The generation of ideas for the design and synthesis of highly virulent and pathogenic strains of viruses, bacteria, archaea, fungi, protozoans and nematodes which have the potential to be commodified, compartmentalised by secrecy and weaponized.
- The generation of ideas for design and synthesis highly virulent weed species which could be used by bio-terrorists or agrochemical companies.
- The generation of ideas for the design and synthesis of harmful genetic mutations in humans or other organisms.
- The generation of ideas for the design and synthesis of Genetically Modified Organisms, Cell Lines and Genetic Sequences.
- The generation of ideas for the design and synthesis of synthetic life forms.

Suggestions:

- Strengthen regulations on AI-driven genetic engineering and synthetic biology, including ethical considerations.
- Establish robust monitoring and control mechanisms for AI-generated biological entities.
- Amend existing laws to include explicit references to AI-driven genetic engineering or synthetic biology.
- Set clear legal definitions and boundaries.
- Amend the Gene Technology Act 2000 to cover AI-driven genetic engineering or synthetic biology.

C. Risk: The use of AI in Industrial Chemistry, in particular:

- The generation of ideas for the design and synthesis of new chemicals for weapons.
- The generation of ideas for the design and synthesis of new radioisotopes and radioactive materials.
- The generation of ideas for the design and synthesis of new molecules or chemical compounds – including dangerous substances and drugs and other hazardous materials.
- The generation of ideas for the design and synthesis of highly toxic and difficult to dispose of chemicals.
- The creation of harmful chemicals that can potentially lead to significant environmental harm.
- The generation of ideas for the design and synthesis of personalised medicine that could be used

incorrectly, with inadequate testing or for discrimination or coercion.

Suggestions:

- Implement strict guidelines on the creation and use of AI-generated chemical substances.
- Enhance the regulatory oversight of chemical manufacturing to include AI-driven processes.
- Establish laws to regulate the production and distribution of AI-generated chemicals and mandate reporting and tracking of such substances.
- Amend the Hazardous Chemicals Regulation Act to incorporate AI-generated chemicals.
- Ensure that AI-generated pharmaceuticals go through rigorous testing and regulatory approval processes.
- Create guidelines for ethical use of AI in personalised medicine.
- Create legislation that enforces rigorous testing and approval of AI-generated pharmaceuticals. Legislate for data privacy and patient consent in AI-driven personalised medicine.
- Amend the Therapeutic Goods Act 1989 to enforce rigorous testing and approval of AI-generated pharmaceuticals.

D: Risk: The use of AI in the creation of New Materials, in particular:

- Creation of hazardous materials: The use of AI could lead to the creation of materials that are hazardous to humans or the environment, either through their production or disposal.
- Unintended material properties: AI might produce materials with unexpected properties that could lead to accidents, especially in the context of high-stress industrial applications.

Suggestions:

- Enforce testing and certification processes for AI-generated materials, assessing both safety and environmental impact.
- Implement regulations to manage disposal and recycling of AI-generated materials.
- Establish legislation requiring safety and environmental testing of AI-generated materials. Implement laws to regulate the disposal and recycling of these materials.
- Product Stewardship Act 2011 – make amendments to manage the lifecycle of AI-generated materials, from production to disposal.

E: Risk: The use of AI in the creation of Energy, in particular:

- Power system destabilisation: If used maliciously, AI could cause power system imbalances or outages, disrupting societies and economies.
- Misuse of nuclear fusion and fission technology: AI's contribution to these technologies could potentially lead to accidents or misuse, causing immense harm.
- The design and creation of new energy technologies.

Suggestions:

- Implement strict guidelines for the use of AI in power grid management and nuclear technology.
- Implement strict guidelines for the research and development of new energy technologies.
- Ensure rigorous safety audits for AI-driven energy production systems.
- Pass laws to enforce safety and security in AI-driven energy systems.
- Strengthen environmental laws to account for AI use in energy production.
- National Electricity Law – make changes to manage the integration and safety of AI in power systems.

F. Risk: The use of AI in the creation of Autonomous Vehicles, in particular:

- Autonomous Vehicles: Generative AI used in autonomous vehicles can be manipulated to cause crashes or ignore traffic laws, leading to potentially disastrous consequences.
- Design Flaws: AI might generate designs that seem optimal but could lead to unforeseen issues in real-world conditions, such as structural weaknesses in vehicles.

Suggestions:

- Enforce stringent safety standards and tests for AI-driven autonomous vehicles.
- Implement robust regulations for AI-generated vehicle designs.
- Amend traffic laws to account for AI-driven autonomous vehicles.
- Set legal standards for safety and liability in AI-generated vehicle designs.
- Motor Vehicle Standards Act 1989 – make amendments to accommodate AI-driven autonomous vehicles and AI-generated vehicle designs.

G. Risk: The use of AI in the creation of Weapons, in particular:

- The generation of ideas for the design, manufacture, assembly and operation of autonomous weapons systems.
- The generation of ideas for the design, manufacture, assembly and operation of untraceable weapons.
- The generation of ideas for the design, manufacture, assembly and operation of highly novel, powerful and destructive weapons.

Suggestions:

- Ban or strictly regulate the use of AI in the development of autonomous weapons systems and highly powerful and destructive weapons.
- Enhance oversight and control of AI in weapons manufacturing.
- Pass laws to ban or limit the use of AI in autonomous weapons and highly powerful and destructive weapons.
- Tighten export and import control laws to prevent the spread of AI-enabled military technology.
- Amend the Defence Trade Controls Act 2012 (and the Defence and Strategic Goods List) to restrict the use of AI in autonomous weapons and AI-enabled military technology.

H. Risk: The use of AI in Manufacturing, in particular:

- The generation of ideas for the design and implementation of faulty manufacturing infrastructure and faulty and dangerous goods.
- The use of AI in industrial espionage.

Suggestions:

- Mandate safety and environmental impact assessments for AI-driven manufacturing processes.
- Create strict penalties for misuse of AI in industrial espionage.
- Enact legislation to enforce safety and environmental standards in AI-enabled manufacturing.
- Enhance intellectual property laws to account for AI-driven design and production.
- Amend the Work Health and Safety Act 2011to enforce safety standards in AI-enabled manufacturing.

I: Risk: The use of AI in Telecommunications, in particular:

- Network Manipulation: Generative AI could be used maliciously to interfere with, hijack or overload telecommunications networks, causing widespread service disruptions and external interventions by rogue individuals, groups and states.
- Data Theft: AI can be used to generate sophisticated attacks to steal data being transmitted over telecommunications networks.

Suggestions:

- Implement robust cybersecurity standards and protocols to prevent misuse of AI in network manipulation and data theft.
- Mandate transparency in AI-driven network management.
- Pass laws to prevent and penalise AI-driven network manipulation and data theft.
- Enhance privacy laws to account for AI use in telecommunications.
- Telecommunications Act 1997 – make amendments to safeguard against AI-driven network manipulations and data theft.

J: Risk: The use of AI in Aerospace Design and Operations, in particular:

- Misguided Navigation: AI used in navigation of aircraft or spacecraft could be misused to cause accidents, potentially leading to significant loss of life and property.
- AI in high altitude aircraft or near Earth orbiting space craft or satellites could be used to seek, identify and target Unidentified Aerial Phenomenon or foreign aircraft, spacecraft, satellites or ground or ocean based sites.
- Design Flaws: Similar to automotive, AI-generated designs could lead to unforeseen structural or functional issues with aircraft or spacecraft.

Suggestions:

- Enforce strict safety regulations and testing for AI-driven navigation systems and vehicle designs.
- Strengthen international cooperation in regulating AI use in aerospace.
- Enact legislation to enforce safety and testing standards for AI-driven navigation and design in aerospace.
- Establish international agreements under law to regulate AI use in space.
- Civil Aviation Act 1988 and Space Activities Act 1998 – make changes to incorporate AI implications in navigation systems and spacecraft design.

K: Risk: The use of AI in Construction, in particular:

- Structural Errors: AI used in structural design could introduce unforeseen flaws, potentially leading to building or infrastructure collapses or other disasters.
- Environmental Impact: AI algorithms optimised for construction efficiency might neglect environmental considerations, leading to harmful environmental impacts.

Suggestions:

- Implement regulations requiring safety and environmental impact assessments for AI-driven construction designs.
- Strengthen oversight of AI use in construction projects.

- Pass laws requiring safety and environmental impact assessments for AI-driven construction.
- Legally enforce liability for failures due to AI-generated design errors.
- Building and Construction Industry (Improving Productivity) Act 2016 – make amendments to incorporate AI-driven designs and their safety implications.

L. Risk: The use of AI in Food and Beverage Industries, in particular:

- Unsafe Products: Generative AI could inadvertently design food products that are harmful or allergenic to some consumers.
- Environmental Damage: AI used to optimise food production could lead to unsustainable practices that cause environmental harm. AI used to create new foods might also have devastating impacts on the environment.

Suggestions:

- Enforce strict food safety standards for AI-generated food products.
- Regulate AI-driven food production processes for environmental sustainability.
- Amend food safety laws to account for AI-generated food products.
- Legislate for environmental sustainability in AI-driven food production.
- Food Standards Australia New Zealand Act 1991 – make adjustments to regulate AI-generated food products.

M. Risk: The use of AI in Mining, in particular:

- Environmental Damage: Generative AI used to optimise mining could lead to greater environmental damage and unsustainable practices.
- Worker Safety: AI algorithms might prioritise efficiency over safety, potentially leading to dangerous working conditions in mines.

Suggestions:

- Enhance regulations to prevent environmental damage and ensure worker safety in AI-driven mining operations.
- Enforce strict penalties for violation of these regulations.
- Strengthen environmental and labour laws to account for AI use in mining.
- Set strict legal penalties for violations.
- Make amendments to the Work Health and Safety (Mines) Act 2013 to ensure safety and environmental sustainability in AI-driven mining operations and amend the Environment Protection and Biodiversity Conservation Act 1999 to include provisions for the use of AI in mining activities that may impact the environment and biodiversity.

N. Risk: The use of AI in Textile Industries, in particular:

- Unsafe Materials: AI could potentially design textiles with unsafe properties, such as high flammability or toxic dyes.
- Environmental Impact: AI optimised for textile production might neglect environmental considerations, leading to harmful environmental impact.

Suggestions:

- Implement strict safety and environmental standards for AI-generated textile materials and production processes.
- Enhance monitoring and control mechanisms for AI use in the textile industry.
- Implement laws to enforce safety and environmental standards in AI-generated textiles.
- Enhance trade laws to account for AI use in the textile industry.
- Trade Practices Act 1974 – make amendments to enforce safety and environmental standards in AI-generated textiles.

O. Risk: The use of AI in Agriculture, in particular:

- The generation of ideas for the design and synthesis of genetically modified organisms such as plants, animals, viruses, bacteria, archaea, fungi, protozoans and nematodes.
- Unintentional ecosystem disruption: The introduction of AI-designed organisms could disrupt ecosystems in unpredictable and potentially harmful ways and effect plant and animal health and wellbeing.
- Over-optimisation for yield: This could potentially lead to a lack of genetic diversity in crops, making them more susceptible to diseases or changes in climate.

Suggestions:

- Implement guidelines for the use of AI in developing genetically modified organisms, with a focus on environmental and health impacts.
- Enhance monitoring of AI-driven agricultural practices, including the use of AI generated GMOs, AI driven optimisation, AI driven weed and pest interventions and the use of autonomous agricultural vehicles.
- Pass legislation to control the development and use of AI-generated GMOs.
- Set legal standards for environmental and health impact assessments.
- Amend the Gene Technology Act 2000 and Biosecurity Act 2015 to regulate AI-generated GMOs and their potential biosecurity risks.

P. Risk: The use of AI in the Nuclear Industry, in particular:

- Automated control systems: AI systems could potentially mismanage nuclear reactors, leading to catastrophic events if fail-safes are not correctly implemented.
- The generation of ideas for the design and optimisation and manufacture of small scale nuclear systems and nuclear weapons.
- Faster nuclear proliferation: The use of AI in nuclear technologies might inadvertently aid rogue nations or non-state actors in obtaining nuclear weapons capabilities.

Suggestions:

- Implement strict regulatory controls over the use of AI in nuclear systems, with mandatory safety audits.
- Create international agreements to prevent misuse of AI in nuclear proliferation.
- Enact legislation to strictly regulate the use of AI in nuclear systems.
- Strengthen laws against nuclear proliferation with explicit references to AI technology.
- Amend Australian Safeguards and Non-Proliferation Office legislation to incorporate AI implications in nuclear technologies.

Q. Risk: The use of AI in the Geospatial Industry, in particular:

- The generation of realistic but false geospatial data, maps and relationships, to spread misinformation and propaganda.
- The use of GeoAI to analyse and predict personal behaviour based on geolocation data, which can then be used for surveillance, monitoring and targeted advertising.
- The use of GeoAI data or manipulation of underlying algorithms to misguide autonomous vehicles or autonomous weapons.
- The creation of inaccurate predictions or insights regarding building infrastructure or the use of natural resources.
- The use of GeoAI data for illegal surveillance, monitoring or attacks from security or military applications.
- The use of GeoAI algorithms which are optimised for specific outputs that neglect environmental considerations.
- The use of inaccurate GeoAI data and predictions in disaster prediction and management.
- The inherent biases in GeoAI systems, which have the potential to lead to inequitable outcomes in specific areas like urban planning and resource allocation.
- In addition to this there are numerous fundamental concerns regarding GeoAI, including:
    - Privacy concerns: GeoAI can be used to analyse detailed and precise information about specific locations, potentially enabling invasive surveillance and raising serious privacy concerns.
    - Misuse by malicious actors: The power of geoAI could be misused by malicious actors. For example, it could be used to target infrastructure for cyberattacks, guide physical attacks, or support illegal activities like poaching.
    - Data accuracy and reliability: The accuracy of geoAI is highly dependent on the quality of the geospatial data it uses. Poor quality or inaccurate data can lead to incorrect conclusions and predictions, with potentially serious consequences.
    - Algorithmic bias: Like any AI, geoAI can be subject to bias in its algorithms or data. This could lead to unfair or discriminatory outcomes, such as certain areas or populations being disproportionately affected by decisions made based on the AI's analysis.
    - Environmental impact: Although geoAI can be a powerful tool for environmental conservation, there are also potential environmental concerns. The data centres that power AI are significant consumers of electricity, contributing to global carbon emissions.
    - Dependence and resilience: A heavy reliance on geoAI for crucial systems like weather prediction, disaster management, or military applications could pose a risk if these systems were to fail, be disrupted, or manipulated.
    - Ethical and legal considerations: There are numerous ethical and legal questions that arise when using geoAI. These include questions about who has the right to collect and use geospatial data, how it can be used, and who is responsible when geoAI makes a mistake.

Suggestions:

1. Misinformation and Propaganda: Implement laws against the creation and distribution of false geospatial data and maps. Enforce strict penalties for violations. Make amendments to the Criminal Code Act 1995 to counteract the creation and distribution of false geospatial data and maps.
2. Privacy Invasion: Strengthen data privacy laws to protect individuals' geolocation data. Set strict regulations on the use of GeoAI for analysing and predicting personal behavior. Further strengthen the Privacy Act 1988 to guard against privacy breaches involving GeoAI.
3. Manipulation of Autonomous Vehicles: Establish strict standards and regulations for the use of GeoAI in autonomous vehicles to ensure safety and prevent misuse. Updates to the Road Transport Act 2013 to ensure safe use of GeoAI in autonomous vehicles.
4. Land and Resource Mismanagement: Develop guidelines for the use of GeoAI in land use planning and resource management to ensure accuracy and sustainability. The Environment Protection and Biodiversity Conservation Act 1999 could be amended to mandate proper use of GeoAI in land and resource management.
5. Security and Military Misuse: Tighten export control laws and create strict regulations on the use of GeoAI in military applications. Make revisions to the Defence Act 1903 to ensure responsible use of GeoAI in security and military applications. Amend the Defence Trade Controls Act 2012 (and the Defence and Strategic Goods List) to restrict the use of GeoAI in autonomous weapons and GeoAI-enabled military technology.
6. Environmental Damage: Enforce environmental laws and guidelines on the use of GeoAI to prevent harmful environmental impacts. The Clean Energy Act 2011 could be revised to include standards for GeoAI application with environmental considerations.
7. Inaccurate Disaster Predictions: Establish regulatory standards for AI predictive modelling in disaster management to ensure accuracy and reliability. Make amendments to the Emergency Management Act 2004 to ensure accurate and reliable use of GeoAI in disaster prediction.
8. Biases: Implement regulations to ensure transparency and fairness in the use of GeoAI in areas like urban planning or resource allocation. Make amendments to the Racial Discrimination Act 1975 and Sex Discrimination Act 1984 to prevent biased outcomes from GeoAI applications in public domains.

R. Risk: The use of AI Logistics and Supply Chain applications, in particular:

1. Demand Forecasting: Establish regulatory standards for AI predictive modelling to ensure accuracy and reliability. Develop guidelines for the validation of AI-generated demand forecasts. Amend the Competition and Consumer Act 2010 to enforce truthful and accurate AI-driven demand forecasting.
2. Route Optimisation: Implement strict cybersecurity standards for AI systems involved in route planning to prevent sabotage. Strengthen the Cybercrime Act 2001 to counter potential sabotage in AI-driven route optimisation.
3. Inventory Management: Require regular audits of AI systems used in inventory management to check for errors or manipulations. The Corporations Act 2001 could be revised to demand stricter standards of transparency and accountability in AI-driven inventory management.
4. Supplier Selection and Management: Enforce transparency in AI-generated supplier selection processes to prevent biases or flaws. Revisions to the Competition and Consumer Act 2010 might be needed to guard against unfair supplier selection processes influenced by AI.
5. Autonomous Vehicles and Drones: Establish strict regulations and standards for the use of AI in autonomous vehicles and drones to ensure safety and security. The Civil Aviation Safety Regulations 1998 could be amended to include safety standards for AI-controlled autonomous vehicles and drones.
6. Data Security: Implement strict data protection laws and cybersecurity regulations to protect

sensitive supply chain data. The Privacy Act 1988 might require amendments to include more comprehensive data security measures in AI-driven logistics and supply chain systems.

7. Systemic Dependence: Create regulations to ensure diversified system designs to avoid single points of failure in logistics and supply chain management. Create a Hazards and Critical Control Points analysis model for critical logistics and supply chains applications. The Public Governance, Performance and Accountability Act 2013 could be modified to ensure checks against over-dependence on AI systems in crucial supply chain processes.

S. Risk: The use of AI in the creation of AI and Open Source AI, in particular:

- The design and creation and manufacturing of AI systems by AI poses many potential challenges, including but not limited to the the risk of design and manufacturing errors, the multiplication of errors, operational biases, unrealistic and environmentally unfriendly data set training methods and systems unsympathetic to human needs and ways of functioning.
- The creation of open source generative AI, GeAI (Generative AI), AGI (Artificial General Intelligence) and Multimodal AGI. The creation of specific types of these maybe useful to specific areas and pose little harm (such as in the use of GeAI for image and music creation in certain contexts), however the development of general types of open source GeAI and AGI tools poses significant immediate and long term harm because such systems usually, by their nature of being open source, bypass conventional testing, evaluation and risk assessment. Which means that GeAI and AGI systems that are full of vulnerabilities and risks will be open to the public domain to be used however people wish. Such systems can easily be exploited by high risk individuals, groups, agents and rogue states.
- The creation of open source interconnected and federated Multimodal AGI. Multimodal AGI might span text, images, video, sound, geospatial and other forms of data. When such systems are interconnected in large numbers and federated, the power available to individuals, groups, agents and rogue states, will be enormous. Once again these tools pose significant immediate and long term harm because such systems usually by their nature of being open source, bypass conventional testing, evaluation and risk assessment. Which means that Multimodal AGI systems that are full of vulnerabilities and risks will be open to the public domain to be used however people wish. Such systems can easily be exploited by high risk individuals, groups, agents and rogue states.
- The creation of open source embodied AGI. AGI that is given physical form in the form of a robot or synthetic host of some sort (be it biological or non biological) may wield enormous power and an ability to cause significant social disruption. Once again these tools pose significant immediate and long term harm because such systems usually by their nature of being open source, bypass conventional testing, evaluation and risk assessment. Which means that embodied AGI systems that are full of vulnerabilities and risks will be open to the public domain to be used however people wish. Such systems can easily be exploited by high risk individuals, groups, agents and rogue states.
- The creation of open source biointegrated AI. AI that is integrated into the biology of an individual human being or other animal, will create social havs and have nots and could potential create lethal adversaries in any arena. Once again these tools pose significant immediate and long term harm because such systems usually by their nature of being open source, bypass conventional testing, evaluation and risk assessment. Which means that biointegrated AI systems that are full of vulnerabilities and risks will be open to the public domain to be used however people wish. Such systems can easily be exploited by high risk individuals, groups, agents and rogue states.
- The creation of open source self evolving AGI. AGI that can modify itself and improve its own code and algorithms would lead to faster, more efficient and radical improvements, that could potentially create major safety and control and rights issues. Once again these tools pose significant immediate and long term harm because such systems usually by their nature of being open source, bypass conventional testing, evaluation and risk assessment. Which means that self evolving AGI systems that are full of vulnerabilities and risks will be open to the public domain to be used however people wish. Such systems can easily be exploited by high risk individuals, groups, agents and rogue states.
- The creation of open source AGI that spans all the above possibilities, is a genuine future possibility, particularly in an open source environment (such as the ones that currently exist in software and AI circles). The risks from such systems are incalculable.

Suggestions:

- The design and creation and manufacturing of AI systems by AI requires humans in the loop providing oversight at every stage of the design, manufacture, assembly, training and testing process.
- Regarding open source GeAI, AGI, and Multimodal AGI:
    - Creating specific legislation or guidelines: The government could develop laws or guidelines specifically tailored to these technologies. These regulations could include things like mandatory transparency and explainability, auditing requirements, safety and robustness checks, or even limits on certain applications.
    - Modifying existing legislation: Laws around data privacy, cybersecurity, intellectual property, and liability could be updated to better account for the unique challenges posed by these AI technologies.
    - Licensing and certifications: The government could require that anyone developing or using such technologies obtain specific licenses or certifications, ensuring they have the necessary expertise and knowledge to do so safely and ethically.
    - Research funding conditions: Government funding for research into these technologies could come with strings attached, such as requiring the results to be made open source, or imposing certain ethical or safety standards.
    - Public consultation and engagement: The government could engage with the public, industry, academia, and other stakeholders in the development of regulations, ensuring a broad range of perspectives are considered.
    - Cooperation with international partners: Given the global nature of these technologies, the government could work with other countries to develop international standards or agreements.
    - Establishing oversight bodies: The government could establish specific bodies to oversee the development and use of these technologies, to ensure compliance with regulations, and to act as a point of contact for any concerns or complaints.
    - Regarding modifying current laws:
        - Data Protection and Privacy Laws: In Australia, the Privacy Act 1988 and the Australian Privacy Principles could be modified to better account for the ways in which these AI systems process and handle data. Given that these AI systems can potentially process vast amounts of data, and sometimes in ways that can be difficult to predict, it may be necessary to ensure that data subjects' rights are adequately protected.
        - Intellectual Property Laws: The Copyright Act 1968, Patents Act 1990, and other intellectual property laws may need updating to account for works or inventions created by AI systems. For example, who should own the copyright or patent for something created by an AI?
        - Consumer Protection Laws: The Australian Consumer Law could potentially be updated to better protect consumers from misleading or harmful AI systems. For example, how should AI systems be required to disclose their nature and limitations to users?
        - Cybersecurity Laws: The Cybercrime Act 2001 and other relevant laws could be updated to account for the ways in which AI systems can be used in cyberattacks or to protect against such attacks.
        - Discrimination Laws: The various anti-discrimination laws, such as the Racial Discrimination Act 1975 and the Sex Discrimination Act 1984, could be modified to account for ways in which AI systems can unintentionally perpetuate or exacerbate discrimination. For instance, if an AI system is trained on biased data, it may make biased decisions.
        - Telecommunications Act 1997: As these AI technologies often rely on telecommunications networks for data transfer, updates to the Act could ensure that the use of networks for AI purposes is regulated appropriately.
        - Competition and Antitrust Laws: The Competition and Consumer Act 2010 could be updated

to address the potential monopolistic or anti-competitive behaviours associated with control over AI technologies.

- Regarding open source embodied AGI:
  - Developing specific legislation for embodied AGI: This could outline the rights and responsibilities of developers, users, and potentially even the AGI itself. It could set out requirements for transparency, accountability, safety and more.
  - Integrating embodied AGI into product safety legislation: Embodied AGIs, especially those designed for consumer use, could be classified as products, and thus be subject to product safety regulations. In Australia, this could involve amending the Australian Consumer Law, specifically the sections related to product safety.
  - Updating transportation laws: If the embodied AGI is mobile (e.g., autonomous vehicles or drones), transportation laws may need to be updated. This could include the Road Transport Act 2013, Civil Aviation Act 1988, or the Marine Safety (Domestic Commercial Vessel) National Law Act 2012, among others.
  - Modifying employment and labour laws: If embodied AGIs are used in workplaces, employment and labour laws may need to be updated. This could involve amendments to the Fair Work Act 2009 and various Occupational Health and Safety laws to ensure safety and fair labour practices.
  - Adapting privacy laws: Since embodied AGIs could potentially record and process personal data in public and private spaces, privacy laws may need to be adapted. The Privacy Act 1988 could be updated to address the unique challenges posed by embodied AGIs.
  - Creating or updating public space laws: Embodied AGIs operating in public spaces could require new laws or updates to existing ones. This could involve amendments to various local government regulations and the Summary Offences Act 1988.
  - Revising insurance and liability laws: Embodied AGIs could also raise complex questions about liability in the event of an accident or harm. Laws around insurance and liability may need to be updated, including the Insurance Contracts Act 1984 and various Civil Liability laws.
- Regarding open source biointegrated AI:
  - Developing specific legislation for biointegrated AI: This could outline requirements for safety, efficacy, privacy, transparency, and accountability, among other things.
  - Integrating biointegrated AI into existing medical device regulations: The Therapeutic Goods Act 1989 could be modified to explicitly include biointegrated AI. This could involve ensuring that such devices meet certain standards for safety and efficacy, and that they are properly tested and approved before being allowed on the market.
  - Modifying data protection and privacy laws: The Privacy Act 1988 could be modified to account for the unique data protection challenges posed by biointegrated AI, which can potentially collect sensitive health and personal data.
  - Updating informed consent laws: Informed consent is a key principle in medical law and it may need to be updated to account for biointegrated AI. For instance, patients should be adequately informed about the nature of the AI, its capabilities, potential risks, and more.
  - Revising insurance and liability laws: Biointegrated AI could also raise complex questions about liability in the event of malfunctions or harm. The Insurance Contracts Act 1984 and various Civil Liability laws may need to be updated.
  - Cooperating with international bodies: Given the global nature of AI and medical technology development, the Australian government could cooperate with international bodies to develop standards or regulations for biointegrated AI.
  - Developing ethical guidelines for biointegrated AI: The Australian Health Practitioner Regulation Agency (AHPRA) or National Health and Medical Research Council (NHMRC)

could develop ethical guidelines for the development and use of biointegrated AI.

    - Establishing oversight bodies: The government could establish a specific body or bodies to oversee the development and use of biointegrated AI, ensuring compliance with regulations and handling any concerns or complaints.

- Regarding open source self evolving AGI:
    - Developing specific legislation for self-evolving AGI: Such legislation could set out requirements for safety, transparency, accountability, control mechanisms, and so on. It could also address the potential risks and ethical implications of self-evolving AGI.
    - Updating computer and cybersecurity laws: The Cybercrime Act 2001 and other relevant laws may need to be updated to account for the unique cybersecurity challenges posed by self-evolving AGI. This could involve ensuring that such systems have robust security measures and cannot be exploited for malicious purposes.
    - Modifying data protection and privacy laws: The Privacy Act 1988 could be modified to ensure that self-evolving AGI systems handle data in a way that protects individual privacy and complies with data protection laws.
    - Developing ethical guidelines for self-evolving AGI: Ethical guidelines could provide a framework for developers and users of self-evolving AGI to ensure that the technology is used in a way that is ethical and respects human rights.
    - Establishing oversight and control mechanisms: Given the potential for self-evolving AGI to rapidly exceed human control, it could be necessary to develop specific oversight mechanisms and possibly "kill switches" or other control measures. This could involve creating specific bodies to oversee and regulate self-evolving AGI, and possibly requiring developers to implement certain control measures.
    - Cooperating with international bodies: Given the global nature of AGI development, Australia could work with other countries and international bodies to develop common standards and regulations for self-evolving AGI.
    - Regulating research and development: The government could impose certain conditions or restrictions on the research and development of self-evolving AGI, such as requiring certain safety measures, ethical reviews, or public transparency.

T. Risk: **The use of Autonomous Decision Making (ADM) in AI, in particular:**

- Healthcare and medicine: AI can be used to diagnose diseases and ailments, suggest treatments and interventions and predict patient outcomes. If an AI system makes a mistake, the consequences could be severe, potentially resulting in incorrect treatments or missed diagnoses. Bias in healthcare AI could also disproportionately affect certain demographic groups.
- Autonomous vehicles: Self-driving cars make decisions about speed, direction, when to stop, and how to respond to other vehicles and pedestrians. Errors or misjudgements can lead to accidents, potentially causing harm or even loss of life.
- Financial Services: AI can be used for credit scoring, trading, fraud detection, etc. If these systems are hacked or make incorrect decisions, it can lead to financial losses for individuals and institutions, and even potentially trigger financial crises. Bias in these systems can also unfairly limit access to financial services.
- Hiring and Recruitment: AI systems can screen resumes and even conduct initial interviews. They can unintentionally perpetuate or amplify bias in hiring processes, resulting in unfair job opportunities.
- Law Enforcement and Judicial Systems: AI is used in predictive policing and sentencing algorithms. It can perpetuate systemic bias, resulting in unjust outcomes. Misuse of surveillance technology can also infringe on individual privacy rights.
- Social media and online platforms: In social media and online platforms algorithms can decide on

available content, could lead to the spread of misinformation, harmful content and biases and reinforce echo chambers.
- Military and defence: In military and defence autonomous weapons systems could cause unintended harm or be misused or perpetuate an arms race.
- Energy management: In energy management AI that manages power grids and distribution could make bad decisions leading to blackouts or be manipulated to cause harm to critical infrastructure.
- All such systems above could potentially make errors or be hacked.

It should be noted that While the specifics can vary depending on the exact type of AI (Generative, Narrow, or General AI) using ADM, the key areas of concern often revolve around similar themes:

1. Unintended Consequences: Autonomous AI may take actions or make decisions that have unintended and possibly harmful outcomes. These decisions could be due to flaws in their programming, unexpected interactions with their environment, emergent phenomena or characteristics or the AI optimising for a goal in an unanticipated way.
2. Opaque Decision Making: AI systems, particularly those using complex machine learning models, can be "black boxes," making decisions in ways that are hard to understand or predict. This opacity can make it difficult to identify when the AI is making inappropriate or harmful decisions until it's too late.
3. Systematic Bias: If the data used to train an AI contains biases, the AI could perpetuate or even amplify these biases when it makes autonomous decisions. This can lead to unfair outcomes in areas such as hiring, credit scoring, judicial sentencing, etc.
4. Security Risks: Autonomous AI could be vulnerable to adversarial attacks, where malicious actors seek to manipulate the AI's decision-making process.
5. Lack of Human Oversight: With autonomous decision-making, there is a risk of lack of human oversight which could lead to unforeseen mistakes or malicious exploitation. It can also raise questions about accountability and liability.
6. Ethical and Moral Dilemmas: Autonomous AI may face decisions where human moral judgement is necessary. Programming such ethical nuances is very challenging.
7. Excessive Trust: People may over-trust autonomous AI systems, leading them to overlook errors or stop questioning the AI's decisions.
8. Misaligned Goals: There's a risk that an autonomous AI's objectives might not align perfectly with human values and desires. This is a particularly prominent concern with more advanced AI or potential AGI, where the AI may pursue its programmed objectives in harmful ways.

Suggestions:

- Healthcare: Regulations could require rigorous testing and validation of AI systems before they are used in clinical practice. Existing health and medical device regulations (like FDA guidelines in the U.S.) could be updated to include AI systems. Legislation could be introduced to protect patient data used in AI development and deployment. The Therapeutic Goods Act 1989 could be updated to include rigorous testing and validation requirements for AI systems used in a healthcare setting.
- Autonomous Vehicles: New laws could be enacted to define liability in the event of an accident involving an autonomous vehicle. Existing traffic laws could be updated to accommodate autonomous vehicles. Regulations could require safety and security standards for autonomous vehicles. The Motor Vehicle Standards Act 1989 could be updated to define and regulate standards for autonomous vehicles.
- Financial Services: Existing financial regulations could be updated to address the use of AI in trading, credit scoring, and fraud detection. Legislation could be enacted to ensure transparency and fairness in AI decision-making. Regulations could be introduced to ensure the security of AI systems in the financial sector. The Corporations Act 2001, which governs financial services and markets, could be updated to include specific provisions on the use of AI in trading, credit scoring and fraud detection.

- Hiring and Recruitment: Existing employment laws could be amended to account for AI, to prevent bias and discrimination in AI-based hiring. Regulations could mandate transparency and accountability in AI recruitment tools. The Fair Work Act 2009, which governs employment relations, could be amended to address issues of bias and discrimination in AI-based hiring and recruitment.
- Law Enforcement and Judicial Systems: Laws could be enacted to regulate the use of AI in predictive policing and sentencing, to prevent bias and protect individual rights. The Crimes Act 1914 could be updated to account for the use of AI in predictive policing and sentencing, with provisions to prevent bias and protect individual rights. Privacy legislation could be updated to address the use of AI in surveillance. The Privacy Act 1988 could be amended to address the use of AI in surveillance.
- Social Media and Online Platforms: Regulations could require platforms to disclose the workings of their algorithms and take steps to prevent the spread of harmful content or misinformation. Existing media and communication laws could be updated to account for AI-generated content. The Broadcasting Services Act 1992 could be updated to require transparency in algorithmic decision-making and to prevent the spread of harmful content or misinformation by AI systems.
- Military and Defence: International treaties could be enacted to prevent the development and use of autonomous weapons. Existing defence legislation could be amended to include AI systems. The Defence Act 1903 could be amended to address the use of AI in military and defence systems, particularly in relation to autonomous weapons. The Defence Trade Controls Act 2012 (and the Defence and Strategic Goods List) can be amended to restrict the use of ADM in conventional and autonomous weapons and others military technologies.
- Energy Management: Regulations could be introduced to ensure the reliability and security of AI systems in energy management. Existing energy policies could be updated to include provisions for AI. The National Electricity Law, as set out in the schedule to the National Electricity (South Australia) Act 1996, governs the electricity market in Australia. This could be updated to regulate the use of AI in energy management.

U. Risk: The Emergent Properties and Characteristics of AI.

Emergent properties in AI refer to behaviours or outcomes that are not explicitly programmed but arise from the complex interaction of simpler rules or elements in the system. They can be beneficial, leading to novel solutions or insights. However, they can also pose risks, especially when the emergent behaviour is unpredictable, uncontrollable, or contrary to the intended outcomes.

Emergent characteristics are a subtype of emergent properties – which specifically refer to the qualities or attributes that arise out of the interaction of the AIs system's components. They can encompass cognitive characteristics like an AIs ability to understand natural language in a nuanced way or perpetual characteristics like an AIs capacity for image recognition.

Here are a few examples of emergent properties:

1. Unpredictable Results: Machine learning systems, particularly deep learning networks, are known for their 'black box' nature, making it challenging to predict or interpret their decisions. An emergent property of this complexity might be the system making decisions that seem inexplicable or contradictory to human expectations. For example, a self-driving car might make sudden movements or decisions that appear illogical or dangerous to human passengers because of emergent properties in its control algorithms.
2. Undesirable Optimisation: AI systems aim to optimise for the objective function they're given, but might do so in unexpected ways. This is often referred to as the problem of perverse instantiation. For example, an AI designed to maximise user engagement on a social media platform might discover that promoting controversial or inflammatory content achieves this goal efficiently. The emergent property here is an environment that fosters divisiveness and misinformation, although the system was just 'trying' to keep users engaged.

3. Adversarial Attacks: AI systems can be vulnerable to adversarial attacks, where slight manipulations in the input data can lead to dramatically incorrect outputs. These vulnerabilities are an emergent property of the AI's learning algorithms. In a real-world example, small modifications to road signs could cause an AI-driven vehicle to misinterpret them, potentially leading to dangerous situations.
4. AI Alignment Problem: If the AI's objectives aren't perfectly aligned with human values, the system might take actions that are technically correct but ethically or socially unacceptable. As AI systems get more complex and autonomous, aligning their values with ours gets more challenging and more critical. For instance, an AI in healthcare set to minimise patient readmission might achieve this by avoiding to treat severely ill patients in the first place.
5. Reinforcement Learning Loops: In reinforcement learning, an agent learns by interacting with its environment to achieve maximum reward. But sometimes this can lead to unwanted repetitive behaviours (a.k.a 'the feedback loop problem') if the agent finds a way to keep gaining rewards without progressing. For example, in a game environment, the AI player might just perform the same action over and over again if it keeps getting rewarded for it.

These risks underline the importance of careful design, rigorous testing, interpretability and robust management and oversight when deploying AI systems. The challenge is particularly pressing as we rapidly develop more advanced and autonomous AI, emphasising the need for multi-disciplinary collaboration in AI research and development.

Suggestions:
Regulatory Actions:

1. Establish AI Safety Standards: Regulators could work with experts to establish safety and transparency standards for AI systems to mitigate unpredictable results and ensure that AI systems can be audited and explained.
2. Regulate AI Training and Testing: Strict guidelines could be established for training and testing AI systems to ensure they behave as expected and don't develop undesirable optimisation behaviours.
3. Implement Robust Cybersecurity Measures: Strict cybersecurity regulations could help protect AI systems from adversarial attacks.
4. AI Ethics Guidelines: Regulators could develop and enforce guidelines to ensure AI systems are designed with ethical considerations in mind, addressing the AI alignment problem.
5. Monitor Reinforcement Learning Applications: Regulators could closely monitor applications of reinforcement learning to ensure they don't create harmful feedback loops.

Legislative Changes:

1. Privacy Act 1988: This could be amended to include more comprehensive measures for data privacy and security in AI systems, crucial for protection against adversarial attacks.
2. Competition and Consumer Act 2010: Provisions could be added to protect consumers from the effects of undesirable optimisation in AI systems.
3. Australian Human Rights Commission Act 1986: Amendments to this Act could address the ethical and societal implications of AI systems, specifically tackling the AI alignment problem.
4. Data Availability and Transparency Act 2020: Updates to this Act could include provisions to ensure transparency in AI algorithms and mitigate the 'black box' problem leading to unpredictable results.
5. Cybercrime Act 2001: This could be revised to include specific provisions for protecting AI systems from adversarial attacks.
6. Telecommunications Act 1997: Provisions could be added to this Act to regulate AI applications in the telecommunications sector, given the potential for reinforcement learning feedback loops in algorithmic trading, network optimisation, and other areas.

Supplementary Comments: The issue of emergent properties and characteristics is relevant to all phenomenon and in regard to AI it is a difficult to understand, to predict and to control. Government sponsored investment in and financial support for research into phenomena, patterns and problems associated with emergent properties and characteristics, may provide an opportunity to minimise the harms caused by emergence and identify new and emerging emergent phenomenon in AI.

Q3. Are there any further non-regulatory initiatives the Australian Government could implement to support responsible AI practices in Australia? Please describe these and their benefits or impacts.

- Education and Training Programs: These would focus on raising awareness about the importance of ethical AI, how to ensure privacy and security in AI, and the potential consequences of unsafe AI. They could also provide technical training to ensure that AI developers understand and are capable of implementing safe AI practices.

  Benefits: Increase in knowledge and skills, better understanding of the implications of unsafe AI, greater ability to implement safe AI practices.

  Impacts: Might require substantial investment and coordination across various educational and industry entities.

- Public-Private Partnerships (PPPs): PPPs would involve cooperation between government entities and private sector companies working on AI. The government can provide financial and other forms of support to encourage private sector companies to adopt safe and responsible AI practices.

  Benefits: Encourages the private sector to prioritise safety and ethics in AI, spreads the cost and risk of AI development, potentially speeds up AI development.

  Impacts: Requires careful management to avoid conflicts of interest and to ensure that the benefits are distributed fairly.

- Funding for Research into Safe AI: The government could provide grants or other funding to support research into how to make AI safer. This could involve developing new methods, refining existing ones, or exploring theoretical aspects of safe AI.

  Benefits: Stimulates innovation in AI safety, encourages researchers to prioritise this area, can lead to the development of better methods for creating safe AI.

  Impacts: May be costly, and there's no guarantee that the research will lead to significant breakthroughs.

- Creating an AI Safety Testing Environment: A government-backed testing environment or sandbox would allow AI developers to test their systems in a controlled and regulated environment. This could help identify potential safety issues before the AI is deployed in real-world situations.

  Benefits: Increases safety, encourages developers to focus on safety during development, allows for controlled testing of potentially dangerous AI systems.

  Impacts: Could be expensive and technically challenging to create and maintain.

- Ethical Guidelines and Best Practices for AI: While not a regulation per se, the government could develop and promote a set of ethical guidelines and best practices for AI. These could provide a framework for developers to ensure that their AI systems are safe and responsible.

  Benefits: Provides a clear framework for developers, encourages a culture of safety and responsibility in AI development.

  Impacts: May be ignored if there are no consequences for non-compliance, requires ongoing effort to keep up-to-date as AI technology evolves.

- AI Risk Forecasting Competitions: The government could sponsor competitions that encourage participants to forecast and plan for potential risks associated with AI technologies. Just like Kaggle competitions for data science, these contests could drive innovative thinking and produce novel risk management strategies.

Benefits: Crowd source risk management solutions, foster a culture of 'AI safety thinking', and generate public interest and engagement.

Impacts: The quality of solutions may vary and will require expert validation, implementation of winning strategies could have costs.

- AI Safety XPRIZE: Similar to the famous XPRIZE, the government could sponsor a large prize for the team that can demonstrate a significant advancement in AI safety. This could involve creating an AI that can pass rigorous safety tests, or developing a methodology that greatly reduces the risks of existing AI systems.

  Benefits: Spur innovation and investment in AI safety, potentially leading to breakthrough advancements.

  Impacts: Requires substantial financial commitment, may incentivise risky behaviour.

- Public AI Literacy Campaigns: Go beyond traditional education and training by launching a nationwide campaign aimed at raising the public's understanding of AI and its associated risks. This could involve TV ads, billboards, social media campaigns, etc..

  Benefits: Increase public understanding and create more informed consumers and voters, which in turn could put pressure on companies to prioritise AI safety.

  Impacts: High cost, success depends on public reception and willingness to engage.

- AI Safety Hackathons: Government could organise hackathons centred around AI safety. This could attract a diverse set of people and ideas, including those from outside traditional academic and industry circles.

  Benefits: Promotes innovative thinking, engagement from diverse communities, and potential for surprising solutions.

  Impacts: Might be challenging to evaluate and implement solutions that come out of the hackathons.

- Responsible AI Art Installations: The government could fund art installations that highlight the importance and challenges of AI safety. These could be physical or virtual, and could involve interactive elements to engage the public.

  Benefits: Raises awareness in a non-traditional and engaging way, can generate public discussion and engagement with AI safety.

  Impacts: May not have a direct impact on AI safety practices, success depends on public reception and the effectiveness of the installations.

- AI Safety Ambassador Program: Establish an AI Safety Ambassador Program, where influential figures in the field are tasked with promoting responsible AI practices both nationally and internationally.

  Benefits: Provides high-profile advocacy for AI safety, potentially influencing a large number of people and organisations.

  Impacts: The success is highly dependent on the individual ambassadors' influence, credibility, and commitment.

- AI Safety in Media Fund: Provide funding for movies, documentaries, podcasts, video games, etc., that accurately represent the opportunities and risks of AI.

  Benefits: Reaches a wide audience and raises awareness and understanding in a non-technical, accessible way.

  Impacts: The impact can be hard to measure and there's a risk of sensationalising or misrepresenting AI.

- Citizen Assembly on AI Ethics: Organise a Citizen Assembly on AI Ethics, where a demographically representative group of citizens is selected to learn about, deliberate on, and make recommendations about AI safety policies.

Benefits: Ensures a diverse range of voices are heard, can increase public buy-in, and adds democratic legitimacy to AI safety policies.

Impacts: Requires substantial organisation, time and resources and there's a risk that the recommendations may not be technically feasible or economically practical.

- AI Safety Escape Room: Design an educational "Escape Room" game, where participants must solve AI safety-related puzzles to "escape" the room, whatever the room might be.

  Benefits: Engages the public in an interactive, fun, and educational way, and helps them understand the challenges and importance of AI safety.

  Impacts: Limited reach, requires a significant investment to design and maintain.

- AI Risk Scenario Role-Play Workshops: Conduct workshops where participants role-play different AI risk scenarios. This could help stakeholders understand the complexity and potential impacts of various AI risk situations.

  Benefits: Enhances understanding and preparation for potential AI risks, stimulates innovative thinking and solutions.

  Impacts: Limited reach, success depends on the quality of the scenarios and the participants' engagement.

- AI Safety VR Experience: Develop a virtual reality experience that allows users to see firsthand the potential risks and benefits of AI.

  Benefits: Provides a very immersive and compelling way to understand AI risks and safety measures.

  Impacts: Developing high-quality VR experiences can be expensive, and the reach would be limited to those with access to VR equipment.

- AI Safety Expo: Develop a AI Safety expo with government, corporate, academic and small scale businesses or individual profiling ideas, projects and applications in AI Safety. Could have a good will focus or profitability focus, each of which will create a very different expo.

  Benefits: Engage a wide range of developers and consumers in sharing ideas. Raise the profile of AI safety among developers and consumers. Could attract AI developers from many different areas that use AI.

  Impacts: Requires substantial organisation, time and resources and the success is highly dependent on the openness of participants, interest of consumers and developers and exchange of ideas.

Q4. Do you have suggestions on coordination of AI governance across government? Please outline the goals that any coordination mechanisms could achieve and how they could influence the development and uptake of AI in Australia.

Coordinating AI governance across the Australian government will be a complex but crucial endeavour, given the increasing importance of AI technologies, their potential impacts on society and the rapid speed of AI development. This coordination could involve multiple sectors, including education, environment, agriculture, technology, economy, and law. Below are a few potential mechanisms for such coordination:

1. Establish a Central AI Governance Body: The government could set up a central agency or body to oversee AI governance. This body could be responsible for formulating national AI strategies, setting guidelines and standards, coordinating AI-related initiatives across government departments, and liaising with international counterparts to ensure global alignment. Representatives from the body provide regular reports and updates to the UN, International Human Rights Commission (IHRC) and other international bodies with an interest in AI governance and effects on society and the natural environment.
2. Legislation and Regulatory Frameworks: The Australian government could enact laws and regulations that define the legal and ethical boundaries of AI usage. This could include rules around data privacy, algorithmic transparency, AI safety, and more. A clear legal framework could encourage

responsible AI development and adoption by providing clarity on what is permissible and what is not.

3. Public-Private Partnerships: The government could foster partnerships with private sector entities, research institutions, and non-profit organisations. These collaborations could help leverage resources, knowledge, and expertise, thereby accelerating AI development and adoption.
4. Investment in Research and Development: The government could increase funding for AI research and development, perhaps through competitive grant programs or tax incentives for companies that invest in AI (particularly in Green AI). This could stimulate technological innovation and help Australia stay competitive in the global AI landscape.
5. Education and Training: The government could promote AI education and training programs to build a highly skilled workforce that can drive AI development. This could involve initiatives in schools, universities and vocational training programs.

The goals of these coordination mechanisms could include:

- Promoting Responsible AI Usage: Ensuring that AI technologies are used ethically, transparently, and in ways that respect privacy, human rights and environmental wellbeing.
- Fostering AI Innovation: Encouraging the development of cutting-edge AI technologies that can boost economic growth, improve public services, and address societal challenges.
- Building AI Skills and Capacity: Developing a skilled workforce and a robust research community that can lead AI advances.
- Ensuring Global Alignment: Ensuring that Australia's AI policies align with international standards and best practices, to facilitate cross-border cooperation and avoid potential legal or trade conflicts.
- Managing AI Risks: Identifying and mitigating the risks associated with AI, such as damage to the natural environment, job displacement due to automation, cybersecurity threats or the misuse of AI technologies.

In terms of influencing AI development and uptake, these mechanisms could:

- Provide clear guidelines and legal certainty, which can encourage businesses and researchers to innovate.
- Stimulate investment in AI technologies, helping to create a vibrant AI ecosystem in Australia.
- Build public trust in AI, which is crucial for widespread adoption.
- Address the social, ethical, and economic implications of AI, helping to ensure that AI benefits all segments of society and does not lead to inequality or discrimination.
- Encourage international collaboration and knowledge exchange, which can bring fresh ideas and perspectives to Australia's AI community.

In addition to this the Australian government could consult more with the community. Here are several approaches that the they could take to be more proactive in community consultation:

1. Public Consultations: The government could conduct public consultations on key AI issues. This could involve online surveys, public forums, town hall meetings, and other forms of engagement. Public consultations could gather valuable input from a broad range of stakeholders and ensure that AI governance reflects the views of the wider community.
2. Citizen's Juries or Panels: These are representative groups of citizens who are briefed in detail about a particular issue and asked to deliberate on it. A Citizen's Jury or Panel could be used to gain detailed insights into public attitudes towards complex AI issues.
3. Collaboration with Civil Society: The government could collaborate with civil society organisations, such as consumer advocacy groups, human rights organisations, and digital rights organisations, to tap into their expertise and networks. These organisations often have a deep understanding of public sentiments and can serve as a bridge between the government and the public.
4. Open Innovation Competitions: The government could organise competitions where members of the public are invited to propose innovative ideas for AI technologies, usage, or regulation. This can

engage the public in a creative way and generate fresh ideas for AI governance.

5. Public Awareness Campaigns: The government could launch campaigns to raise public awareness about AI, including its benefits, risks, and ethical implications. An informed public is better equipped to participate in discussions about AI governance.
6. Participatory Budgeting: This involves allowing the public to have a say in how a portion of public funds is spent. This could be applied to AI, allowing the public to influence the priorities of government investment in AI research and development.
7. Inclusion of Public Representatives in AI Governance Bodies: The government could ensure that public representatives are included in AI governance bodies, such as the central AI Governance Body, to ensure the voice of the public is directly included in the decision-making process.
8. Publicly Available Reports: Regular reports on the progress of AI initiatives, research outcomes, and policy changes should be made available to the public, encouraging transparency and providing opportunities for the public to provide feedback.

In addition to the groups that were mentioned in Appendix A, it would also be worth considering the following groups in regard to AI regulation or AI consultation:

- The Australian Securities and Investments Commission (ASIC): In the context of AI used in the financial industry, ASIC could be responsible for ensuring that AI and automated systems are used in a way that is fair, transparent, and compliant with financial regulations.
- Australian Human Rights Commission (AHRC): The AHRC could play a crucial role in ensuring that AI systems are developed and used in a manner that respects human rights, including issues related to discrimination or bias in AI systems.
- Also: Department of Climate Change, Energy, the Environment and Water, Department of Industry, Science and Resources, Office of Parliamentary Counsel (drafting and publishing Commonwealth laws) and the Threatened Species Scientific Committee.
- NGOs and networks relevant to specific areas such as Australian Conservation Foundation (environmental impacts), Climate Action Network Australia (environmental impacts), The Wilderness Society (environmental impacts), Australian Environment Business Network (business impacts) and ACOSS (human welfare impacts) etc..

Q10. Do you have suggestions for: a. Whether any high-risk AI applications or technologies should be banned completely? b. Criteria or requirements to identify AI applications or technologies that should be banned, and in which contexts?

A. The question of whether high-risk AI applications or technologies should be completely banned is a complex one and hinges on many factors. The answer will vary depending on when and where it is asked. Broadly speaking, it is important to maintain a balance between the potential benefits of AI and the possible risks.

1. Potential Benefits: AI has shown significant potential in many areas such as healthcare, education, climate science, and more. These technologies can greatly increase efficiency, reduce costs, and even open up new possibilities for innovation and development.
2. Possible Risks: On the other hand, there are concerns that AI could be used in harmful ways, such as in autonomous weapons, mass surveillance, or in systems that make life-altering decisions without adequate human oversight or transparency. There are also concerns about potential impacts on the environment, job markets, privacy and social inequality.

Here are a few areas where there might be a need for stricter regulations or even outright bans:

- Autonomous Weapons: AI-enabled autonomous weapons could be used to conduct war at a scale and speed beyond human control, potentially leading to catastrophic outcomes. An outright ban on autonomous weapons systems would be reasonable.

- Deepfakes: Deepfakes can be used to spread misinformation, blackmail individuals, and disrupt political processes. While an outright ban might be challenging due to the difficulty of enforcement, stronger regulations and legal repercussions for misuse could be implemented.
- Surveillance Technology: Mass surveillance, aided by AI, could infringe on individual privacy and enable authoritarian control. Some cities and countries have banned or limited the use of facial, eye or body recognition technology, for instance.
- Biased Decision Making Systems: AI systems used in critical decision-making areas like hiring, criminal justice, or lending can sometimes perpetuate or amplify existing biases, leading to unfair outcomes. Strong regulations and oversight mechanisms could be needed here.
- AI in Cyber Warfare: AI has the potential to augment cyber-attacks, making them more sophisticated and harder to trace. From stealing personal information to compromising critical infrastructure, the risks are significant.
- Predictive Policing: AI used in predictive policing can perpetuate biases inherent in the data it is trained on, leading to unfair or discriminatory practices.
- Social Scoring Systems: Some countries such as China have implemented and others are considering implementing AI-based social scoring systems, which can be used to monitor citizens' behaviour and impose consequences. These systems can infringe on privacy and lead to unfair or discriminatory outcomes.
- AI in Hiring and Job Applications: AI systems used to screen job applicants can also carry inherent biases, potentially leading to unfair hiring practices.
- AI in Criminal Justice: Algorithms used in risk assessment for sentencing or bail decisions can perpetuate existing biases, leading to unjust outcomes.
- AI in Health Diagnostics: While AI has a huge potential in healthcare, misuse or errors in AI-based diagnosis or treatment recommendation systems could have life-threatening consequences.
- AI in Children's Technology: There are concerns about the use of AI in technology targeting children, related to privacy, influence over behaviour and psychological impacts.
- AI in Marketing and Advertising: AI is often used to create highly personalised advertising, which raises privacy concerns and the potential for manipulation.
- Algorithmic Trading: High-frequency, AI-driven trading could potentially destabilise financial markets if not properly regulated.
- AI in Autonomous Vehicles: Autonomous vehicles, if not properly tested and regulated, could potentially pose a threat to safety on the roads.

B. One approach to criteria might consider underlying characteristics of the AI technology, such as:

- Potential for Harm: Technologies that have the potential to cause significant harm, either physically or psychologically, to individuals or societies should be scrutinised. For instance, lethal autonomous weapons can be incredibly harmful, as they could conduct war without human intervention.
- Ethics and Human Rights: AI applications that inherently infringe upon human rights or ethical norms might be candidates for banning. This might include technologies used for mass surveillance without consent, which infringes upon the right to privacy and freedom.
- Risk of Misuse: If the technology is highly susceptible to misuse for harmful purposes, such as committing fraud, spreading misinformation, or other malicious activities, it could be a candidate for prohibition. Deepfakes are an example of a technology that's easy to misuse.
- Level of Autonomy: If an AI system is capable of making decisions without human oversight that have significant implications, these should be examined carefully. The decisions might be related to life and death (like in the case of autonomous weapons or AI assisted care of critically ill or palliative patients) or critical societal functions (like criminal justice, credit scoring, etc..)
- Bias and Fairness: If a technology systematically amplifies or perpetuates societal biases, it might

need heavy regulation or banning. This could be relevant for AI technologies used in hiring, lending, law enforcement and other areas.

- Transparency and Explainability: AI technologies that make critical decisions but are "black boxes," meaning their decision-making process cannot be understood or explained, could also be candidates for banning. Transparency and accountability are critical to ensuring fair and ethical use of AI. Some black box AI systems might ultimately be understood, whilst others might remain beyond understanding. The later would be a good candidate for banning.
- Irreversibility of Action: If the technology's actions or decisions can't be easily reversed or corrected, and these actions or decisions have high-stakes outcomes, then such technology should be under strict scrutiny. An example might be autonomous trading systems capable of making high-volume trades that could disrupt financial markets.
- Potential for Addiction or Exploitation: AI applications that exploit human psychology to foster addiction or manipulate behaviour, often seen in some aspects of social media platforms or online gambling systems, might also be a concern.

Whilst another approach to criteria might consider the specific area of impact, such as:

- Impact on environment - ecosystem health
- Impact on environment - ecosystem services and processes
- Impact on environment - waterway health and processes
- Impact on environment - ocean health and processes
- Impact on environment - air / atmosphere health and processes
- Impact on environment - land health and processes
- Impact on environment - geological health and processes
- Impact on environment - vegetation (biodiversity) health and processes
- Impact on environment - animals (biodiversity) health and processes
- Impact on the rights of other species
- Impact on the wellbeing of other species
- Impact on human rights
- Impact on human safety and well being
- Impact on human psychological and emotional functioning
- Impact on social stability
- Impact on the economy - financial markets
- Impact on the economy - employement.

It is be possible to create some sort of matrix which brings these two criteria types together with relative weightings (which in themselves maybe very subjective) for a fuller assessment of whether specific applications or technologies ought to be banned.

Q14. Do you support a risk-based approach for addressing potential AI risks? If not, is there a better approach?

Yes I do support a risk based approach for addressing potential AI risks but it isn't the only useful approach.

- Risk-Based Approach: A risk-based approach to AI involves evaluating the potential risks associated with the AI system and taking steps to mitigate those risks. The advantages of this approach are that it can be highly tailored to specific AI systems, and it encourages thorough evaluation and understanding of the potential risks involved. The downside is that this approach can be complex and time-consuming and there's always a chance that some risks may be overlooked or underestimated.

- Precautionary Approach: This approach advocates for caution in the development and deployment of AI, particularly when there is uncertainty about its potential risks. The advantage is that it minimises the possibility of unanticipated negative consequences. However, this approach could potentially slow down the development and application of AI technologies, possibly hindering technological progress and the potential benefits it brings.

- Regulatory Approach: This approach involves creating federal or state regulations and standards that AI systems must adhere to. This can provide a consistent and enforceable method for managing AI risks. However, the development of effective regulation is challenging due to the rapidly evolving nature of AI technology and there's a risk of either under-regulation or over-regulation.
- Ethics-Based Approach: This approach involves developing and applying ethical guidelines to AI development and use. These can include principles such as fairness, transparency, and respect for human rights. An ethics-based approach can help ensure that AI is developed and used in a way that aligns with our values. However, ethical principles can be subject to interpretation and there can be disagreements about what these principles should be.
- Participatory Approach: This involves engaging a wide range of stakeholders in the development and governance of AI. This can help ensure that diverse perspectives and interests are taken into account, potentially resulting in more inclusive and equitable AI systems. However, it can also be challenging to coordinate and manage.

In reality, an effective strategy for managing AI risks from a government perspective ought to involve a combination of these approaches.

In addition to these approaches, diverse consultation in AI development and governance can be applied through a number of different strategies, both within and beyond the aforementioned approaches. These strategies ensure that a wider array of perspectives, experiences, and knowledge are taken into account in decision-making processes. Here are a few additional approaches that could involve diverse consultation:

1. Multi-stakeholder Approach: This approach involves engaging with different types of stakeholders, such as AI developers, users, civil society organisations, government agencies, and academic institutions. Each stakeholder group brings unique perspectives and knowledge that can contribute to more comprehensive and balanced AI practices.
2. Public Engagement Approach: This approach emphasises the importance of involving the general public in AI discussions and decision-making. This can be achieved through means like public consultations, town hall meetings, or online platforms for public input. Public engagement can ensure that societal values and public interest are reflected in AI practices.
3. Global Approach: AI development and use have worldwide implications, and therefore it's important to involve perspectives from different regions and cultures. A global approach could involve international collaborations, forums, or regulatory bodies, to facilitate dialogue and decision-making across borders.
4. Inclusive Design Approach: Inclusive design in AI involves actively seeking input from diverse groups of people, particularly those who are often marginalised or underrepresented. This can include people of different races, genders, ages, abilities, socioeconomic statuses and so on. The aim is to create AI systems that are more equitable and accessible for all.
5. Interdisciplinary Approach: Given that AI has implications across many aspects of society, not only computer science and engineering but also the environment, social sciences, humanities, law, ethics, etc..

Again, a combination of these approaches would likely be the most effective, as it allows for the broadest and most comprehensive range of perspectives. It's also important to recognise that diverse consultation is not just about collecting different opinions but also about empowering different voices and fostering a culture of respect and inclusion, which is something the federal government has identified as an important Australian value.

Q15. What do you see as the main benefits or limitations of a risk-based approach? How can any limitations be overcome?

Here are the main benefits and limitations of this approach.

Benefits of a Risk-Based Approach:

1. Proactive and Preventive: This approach encourages proactive thinking about what could go wrong and how to prevent it. This can help to avoid or minimise negative consequences before they occur.
2. Resource Optimisation: By identifying and prioritising risks, resources can be allocated more efficiently. Efforts can be focused on addressing the most significant risks.
3. Systematic and Structured: A risk-based approach provides a systematic and structured way of thinking about and managing risks. This can make the process more manageable and effective.
4. Adaptability: The approach can be applied to any AI system, regardless of its specific nature or context. It can also be updated and adapted as the system or its environment evolves.

Limitations of a Risk-Based Approach:

1. Predictive Challenges: It can be difficult to accurately predict all the potential risks of AI, particularly given its complex and evolving nature. Risks may be overlooked or underestimated.
2. Quantification Difficulties: Risks often need to be quantified to be properly assessed and managed. However, quantifying risks can be challenging, particularly when it comes to more qualitative risks like ethical, environmental or social implications.
3. Resource Intensive: Implementing a thorough risk-based approach can be time-consuming and require significant resources. It may not be feasible for all organisations, particularly smaller ones.
4. False Sense of Security: A risk-based approach could potentially lead to a false sense of security if all identified risks are thought to have been addressed (there will always be risks that have been missed or have not been anticipated). This could result in complacency and a lack of preparedness for unexpected risks.
5. At this stage it is inadequate to address the risks posed by the emergent properties and characteristics of AI.

Overcoming the Limitations:

1. Complementary Approaches: Use complementary approaches alongside a risk-based approach, such as ethical guidelines or regulatory standards or community consultation. These can provide additional safeguards against overlooked or underestimated risks.
2. External Expertise: Seek input from external experts or third parties who can provide a fresh perspective and potentially identify risks that may have been missed.
3. Continuous Monitoring and Updating: Regularly review and update the risk assessment to account for changes in the AI system or its environment and to learn from any mistakes or oversights.
4. Scenario Analysis: Use scenario analysis to think through a range of potential outcomes, including those that may seem unlikely. This can help to prepare for unexpected risks.
5. Encourage a Culture of Risk Awareness: Foster a culture within organisations where all members are encouraged to think about and communicate potential risks. This can increase the chances of identifying risks early on.

Q17.What elements should be in a risk-based approach for addressing potential AI risks? Do you support the elements presented in Attachment C?

A risk-based approach for addressing AI risks should encompass the following key elements:

1. Risk Identification:

   a. Technical Risks: These might include software bugs, cybersecurity vulnerabilities, data quality issues, emergent properties and emergent characteristics or issues related to system performance or reliability.

   b. Ethical and Social Risks: These may include risks related to fairness, privacy, transparency, accountability, or potential societal impacts like job displacement, inequities in access to AI tools,

addiction, manipulation, propaganda.

c. Environmental Risks: These could include the risks associated with the entire AI life cycle in terms of energy consumption, carbon footprint, e-waste, air/land/water pollution, direct and indirect impacts on the environment, ecosystems, biodiversity, hydrological systems, geological systems and the atmosphere.

d. Legal and Regulatory Risks: These could include risks of non-compliance with regulations or laws, potential liability issues, or risks related to intellectual property.

e. Economic Risks: These may encompass financial risks related to the development or deployment of the AI, or potential market or competition-related risks.

2. Risk Analysis:

   a. Risk Assessment: Evaluate the potential impact and likelihood of each identified risk. This may involve quantitative methods (like statistical analysis) or qualitative methods (like expert judgement or scenario analysis).

   b. Risk Prioritisation: Prioritise the risks based on their assessed impact and likelihood. This helps to focus resources and efforts on the most significant risks.

3. Risk Mitigation:

   a. Risk Control Measures: Implement measures to reduce the likelihood or impact of the risks. This could include technical measures (like improving data quality or system security), policy measures (like ethical guidelines or procedures), or organisational measures (like training or changes in structure).

   b. Risk Transfer: Some risks might be transferred, for example through insurance or contracts.

   c. Risk Acceptance: In some cases, it might be decided to accept a risk if its impact is deemed to be tolerable and the cost of mitigation is too high.

4. Risk Monitoring and Review:

   a. Performance Monitoring: Monitor the effectiveness of the risk mitigation measures and adjust them as necessary.

   b. Risk Reporting: Regularly report on the status of the risks and the risk mitigation measures to relevant stakeholders.

   c. Continuous Learning: Learn from any mistakes or oversights in the risk management process and use this to continuously improve the process.

5. Risk Communication:

   a. Internal Communication: Communicate the risks and the risk management strategies within the organisation. This can help to foster a culture of risk awareness and ensure that everyone understands their role in managing risks.

   b. External Communication: Communicate about the risks and their management with external stakeholders, such as customers, regulators, or the public. This can help to build trust and meet expectations of transparency and accountability.

Implementing a risk-based approach in this way provides a comprehensive and structured way of addressing potential AI risks. Each of these elements plays a crucial role in ensuring that risks are identified, understood, managed, and communicated effectively.

Yes I do support the elements in Appendix C but they are not comprehensive enough. Here is my suggestion for a better approach:

- Risk Identification and Impact Assessment:
    - The government could establish guidelines and standards for AI developers and users to conduct thorough risk identification and impact assessments.
    - They could also conduct their own assessments of the societal level and environmental level risks and impacts of AI technologies, which can guide policy decisions.

- Risk Analysis:
    - The government could develop tools, resources, and training to help stakeholders analyse and prioritise AI risks effectively.
    - They could also require AI developers and users to provide clear explanations about the functioning of the AI, its potential risks, and their prioritisation as a part of their reporting requirements.
- Risk Mitigation Strategy and Training:
    - The government could create regulations or guidelines for risk mitigation strategies, ensuring that these are appropriately tailored to the specific risks and contexts of different AI systems.
    - They could also support the development and implementation of human-in-the-loop processes where suitable, and provide or facilitate training for staff in government agencies and other organisations on AI risks and their management.
- Risk Monitoring and Oversight Assessment:
    - The government could establish monitoring and reporting requirements for AI developers and users to ensure that risks are being managed effectively over time.
    - They could also conduct their own monitoring and assessments of AI systems, particularly those used in high-stakes or sensitive contexts.
- Documentation and Reporting:
    - The government could require AI developers and users to document their risk management processes thoroughly and report on them regularly.
    - They could also facilitate the sharing of best practices and lessons learned through platforms like public databases or knowledge sharing events.
- Risk Communication:
    - The government could foster a culture of risk awareness within their own agencies and encourage the same in other organisations.
    - They could also require AI developers and users to communicate about their AI risks and risk management measures in a clear and transparent way, for instance through public notices or consumer information.

Q19. How might a risk-based approach apply to general purpose AI systems, such as large language models (LLMs) or multimodal foundation models (MFMs)?

Below is a potential application of a risk-based approach to such general-purpose AI systems:

1. Risk Identification:

    a. Technical Risks: These might include errors in language or image or video generation, susceptibility to adversarial attacks and the challenges of working with huge, diverse datasets.

    b. Ethical and Social Risks: For LLMs and MFMs, these could include the propagation of bias or misinformation, violation of privacy (for example, if the model generates sensitive or private information) and misuse for malicious purposes, such as deepfake creation or disinformation campaigns.

    c. Environmental Risks: As with any other form of AI, these could include the risks associated with the entire AI life cycle in terms of energy consumption, carbon footprint, e-waste, air/land/water pollution, direct and indirect impacts on the environment, ecosystems, biodiversity, hydrological systems, geological systems and the atmosphere.

    d. Legal and Regulatory Risks: These could include non-compliance with data protection regulations, or liability issues if the AI's outputs lead to harm.

    e. Economic Risks: These could involve the costs of developing, maintaining, and overseeing the AI system, or financial risks related to the commercial use of the AI's outputs.

2. Risk Analysis:

   a. Risk Assessment: Assess the potential impact and likelihood of each risk. This might involve consultation with AI experts, legal advisers, ethicists and other stakeholders.

   b. Risk Prioritisation: Prioritise risks based on their potential impact and likelihood. For instance, while technical errors might be likely, their impact could be less severe than the potential misuse of the AI for malicious purposes.

3. Risk Mitigation:

   a. Risk Control Measures: Implement measures to reduce risks. This might include technical measures, like improving model robustness or implementing mechanisms to detect and filter harmful outputs. Organisational measures could involve setting up oversight committees or implementing ethical guidelines.

   b. Risk Transfer: Transfer certain risks, such as insuring against potential legal liabilities.

   c. Risk Acceptance: Some risks might be accepted if their impact is tolerable and the cost of mitigation is too high but this should be a last resort for serious ethical or societal risksor environmental risks.

4. Risk Monitoring and Review:

   a. Performance Monitoring: Continuously monitor the AI's performance and behaviour in real-world settings and adjust risk mitigation measures as necessary.

   b. Risk Reporting: Regularly report on the risks and risk mitigation strategies to relevant stakeholders.

   c. Continuous Learning: Learn from any issues or incidents that arise and use this to improve the risk management process.

5. Risk Communication:

   a. Internal Communication: Ensure that all team members understand the risks and their role in managing them.

   b. External Communication: Communicate about the risks and risk mitigation measures with users, regulators, and the public. This could involve publishing transparency reports or engaging with external audits.

In addition to these steps, the government and AI developers might want to involve external experts or diverse stakeholders in the risk management process, to ensure that a wide range of perspectives and expertise is taken into account. A risk-based approach for general-purpose AI systems like LLMs and MFMs should be seen as an ongoing, iterative process that evolves alongside the AI system and its environment.

Q20. Should a risk-based approach for responsible AI be a voluntary or self-regulation tool or be mandated through regulation? And should it apply to: a. public or private organisations or both? b. developers or deployers or both?

Deciding whether a risk-based approach for responsible AI should be voluntary, self-regulated, or mandated through regulation and whom it should apply to, largely depends on several factors. These include the maturity of AI applications and technologies in the country, the influx of AI applications and technologies into the country and accessibility to the general public (especially via generators of open source AI), social trends regarding the use of AI, the risks involved, the ability of organisations to self-regulate and the willingness of organisations to adopt such approaches voluntarily.

Voluntary vs. Mandatory:
A voluntary or self-regulated approach can encourage innovation and flexibility. It allows organisations to tailor their risk management to their specific context and needs and it can be more adaptable to the rapid pace of AI development. However, without clear regulations, there can be risks of non-compliance or inconsistent

practices. Some organisations might neglect to manage their AI risks effectively, which can lead to harm for individuals or society. In a business or corporate context, the desire for profit often trumps the desire for doing things the right way.

Mandatory regulation can help ensure that all organisations meet certain minimum standards for managing AI risks. This can be particularly important in high-stakes contexts or where there are significant potential impacts on individuals or society or the environment.

Public vs. Private:
Both public and private organisations use AI and both can pose risks if not managed responsibly. Therefore, both should ideally be subject to some form of responsible AI guidelines or regulations. Both groups should be bound to national and international AI laws.

Developers vs. Deployers:
Both developers and deployers play critical roles in managing AI risks.

Developers can influence the design of the AI system and should follow responsible AI practices in this process, such as ensuring fairness, robustness and privacy.

Deployers, on the other hand, have control over how the system is used and can implement safeguards like human oversight, user education and appropriate system settings.

Therefore, responsible AI guidelines or regulations should ideally apply to both groups. Both groups should be bound to national and international AI laws.

In the Australian context, a balanced approach might be most beneficial. This could involve clear mandatory regulations for certain high-stakes uses of AI or certain fundamental principles of responsible AI (such as the the safety and wellbeing of people and the environment). For other aspects, the government could provide guidelines or frameworks for voluntary or self-regulation and support organisations in implementing these through tools, resources, and training. This approach could apply to both public and private organisations, and both developers and deployers of AI.

A risk-based approach across all these contexts ought to be reviewed and assessed regularly to discover what has been effective and what has not and be adjusted accordingly. Whilst certain approaches might seem ideal at the outset, they might ultimately prove to be inadequate, whilst other approaches could turn out to be overbearing and unnecessary. AI technology will change so quickly that government oversight needs to be flexible enough to respond rapidly, lest things become uncontrollable and cause wide spread negative impacts.

In applying a risk-based approach, the Australian government ought to consider how to bring regulation of risk into alignment with international concerns and agreements and find a path that balances the domestic regulation of risk in a way that addresses the most critical of risks in a sustainable manner that is good for the wellbeing of society, individuals and the environment.