Safe and Responsible AI in Australia Discussion Paper: Consultation Questions

**Source:** [**Safe and responsible AI in Australia: discussion paper (apo.org.au)**](https://apo.org.au/node/322938)

# 

# Definitions

## 1. Do you agree with the definitions in this discussion paper? If not, what definitions do you prefer and why?

We would like to propose the following definitions for consideration:

* **Artificial Intelligence:** “Simulation of aspects (or whole) of human intelligence in machines, particularly computer systems.”. The definition of AI given in the paper is too narrow and is more appropriate for the definition of Machine Learning, which is a subset of AI. ML is by far the most popular sub-branch of AI today, but not all AI systems take ML approach.
* **Machine Learning:** as suggested above, the currently used definition of AI should replace the definition of ML. The definition provided captures some aspects of machine learning, but it is not entirely comprehensive and might be a bit confusing. Specifically, it focuses too much on the output of machine learning (i.e., patterns) rather than explaining the process and purpose of machine learning as a field of study. Here's an alternative definition: "Machine learning is a subset of artificial intelligence that refers to a collection of techniques (algorithms) that allow computers to learn from and make decisions or predictions based on data, without explicit programming."
* **Multimodal Foundation Model (MfM):** this is a minor point, but the fact that MFM can “learn from and understand” multiple forms of inputs is more important to describe than the fact that it can “generate” multiple forms of outputs. This is important because understanding of different forms of data is what gives rise to more powerful cognitive traits that resemble general intelligence.

# Potential gaps in approaches

## 2. 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?

Australia's existing regulatory landscape largely addresses AI-related concerns from a data privacy, cybersecurity, and ethical standpoint. These frameworks, while crucial, may not fully account for the escalating risks that could emerge as AI technology approaches the level of artificial general intelligence (AGI) – a point where AI systems can perform any intellectual task that a human being can.

One hypothetical but highly illustrative concern is the 'Paperclip Maximizer’ thought experiment, coined by philosopher Nick Bostrom. Imagine a super intelligent AI whose single objective is to make paperclips. Given its immense cognitive capacity and lack of human-like values, it could potentially convert the entire planet into paperclips, without any consideration for human life. This simple yet striking example exposes how AGI, without appropriate control mechanisms, could generate catastrophic outcomes. While this is an exaggerated example, it underscores the importance of having safeguards in place. Without adequate regulatory measures, AGI-level AI could, for instance, misinterpret or carry out commands in harmful ways, much like the paperclip machine, but potentially with more devastating consequences.

Considering these potential risks, it may be prudent to consider a regulatory framework where only licensed professionals are permitted to develop and operate high-powered AI systems. Like how professions like medicine, engineering, or law require licenses due to the potential harm arising from malpractice, AI could follow the same path. By ensuring those working with AGI have the necessary training, ethical understanding, and oversight, we can reduce the risk of misuse or unintended consequences.

Though current AI systems are not yet sophisticated enough to warrant such concern, the prospect of a 'fast take-off' scenario justifies early preparation. In this scenario, the progress of AI to AGI could be abrupt and exponential, potentially leaving us unprepared and incapable of implementing effective safety measures post hoc.

While our technological landscape is rapidly evolving, it is essential to adapt our regulations at the same pace. By integrating licensing requirements for advanced AI operation into our regulatory framework now, we can help ensure a future where AGI is not just technologically advanced but is also aligned with our shared societal values and safety.

## 3. 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.

* **Consideration of Workforce Transition & Displacement:** Consider how to handle significant job displacement that will likely happen as result of AI based automation, especially as AI technologies become more powerful
* **Harness Organizational Insights:** Seek to learn from organizations with hands on experience, and disseminate the insights
* **Promote Open Dialogue and Stakeholder Engagement:** The government could establish platforms for dialogue and engagement between various AI stakeholders, including government agencies, AI developers, businesses, academic researchers, civil society organizations, and the public. This could help ensure that a broad range of perspectives are considered in decisions about AI use and governance.
* **Investment in Research and Development:** The government could invest in academic and industrial research focused on ethical AI development and deployment, including AI alignment research. This would accelerate progress in creating AI that respects human values and rights.
* **Public Education and Awareness:** The government could launch public awareness campaigns to educate the public about AI, its uses, and implications. This will promote a better understanding of AI, enabling people to make informed decisions about its use.
* **Skills and Training:** Providing resources and opportunities for AI education and training would help build a skilled workforce capable of developing and overseeing responsible AI. This could include partnerships with educational institutions, vocational training programs, or online courses.
* **Promote Development of Industry Standards:** While not directly imposing regulations, the government can encourage and support the development of industry standards for responsible AI. This can include best practices for AI design, testing, use, and monitoring.
* **Support for AI Startups and Small Businesses:** The government could provide support, such as funding or tax incentives, to startups and small businesses working on responsible AI technologies. This would foster innovation and ensure a diverse AI ecosystem.

## 4. 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.

<No additional feedback>

# Responses suitable for Australia

## 5. Are there any governance measures being taken or considered by other countries (including any not discussed in this paper) that are relevant, adaptable and desirable for Australia?

<No additional feedback>

# Target areas

## 6. Should different approaches apply to public and private sector use of AI technologies? If so, how should the approaches differ?

AI technologies, both in the public and private sectors, should share common principles such as transparency, fairness, bias, privacy, and security. In both sectors, clear regulations and ethical guidelines should be in place to ensure the responsible use of AI.

## 7. How can the Australian Government further support responsible AI practices in its own agencies?

* **Establish Clear Guidelines:** Develop and implement a comprehensive and clear AI ethics framework that guides AI deployment in public services. This should cover principles like fairness, inclusivity, transparency, privacy, and accountability.
* **Regulation and Oversight**: Regular audits should be conducted to ensure compliance with these guidelines. An independent oversight body might be beneficial for conducting these audits and handling grievances related to AI use.
* **Education and Training:** Regularly train government employees on responsible AI use, its potential biases, its impacts, and how to interpret its results. This helps build a culture of responsibility and accountability around AI use.
* **Transparency:** Be transparent about AI use, clearly communicate to the public how data is used, how decisions are made, and provide avenues for redressal if AI decisions negatively impact individuals.
* **Public Engagement**: Involve the public in discussions around AI use in public services. This could be through public consultations or citizen juries. Public engagement can increase trust and acceptance of AI systems.
* **Collaboration:** Collaborate with academia, industry, and international partners to understand best practices, stay updated on latest technologies and to develop robust AI systems.
* **Investment:** Invest in AI research and development, and in technologies that ensure data privacy and security.

It's important to continuously revisit and update these measures as AI technology and its societal impacts evolve.

## 8. In what circumstances are generic solutions to the risks of AI most valuable? And in what circumstances are technology-specific solutions better? Please provide some examples

Our view on this matter is in alignment with the idea of sector-specific regulation. Generic solutions to the risks of AI are most valuable when the challenges are broad-based, consistent across applications, and pertain to basic ethical principles. For instance: Data Privacy, Transparency, Bias Mitigation, etc. These solutions will provide baseline guidance for organizations in absence of industry or domain specific regulations / guidance.

On the other hand, technology-specific solutions are better when the risks are directly tied to the particular characteristics or application of the AI. For example:

1. Autonomous Vehicles: AI systems controlling autonomous vehicles have specific safety and decision-making considerations (like what action to take in case of an unavoidable accident) that are quite distinct from other AI applications. In such a case, technology-specific solutions are needed.
2. AI in Healthcare: AI solutions used for diagnosing diseases or predicting patient outcomes should adhere to specific standards of accuracy, reliability, and interpretability that may be higher than for other types of AI applications. They must also comply with healthcare-specific regulations such as The Privacy Act 1988 (or HIPAA in the US).

In essence, certain technologies and sectors carry inherently greater risks due to their direct impact on human life, safety, and well-being. Autonomous vehicles, for instance, need to be capable of making split-second decisions that could potentially result in life-or-death outcomes. Similarly, AI systems in healthcare may influence critical medical decisions, with any inaccuracies leading to grave consequences. In these and similar scenarios, the potential risks involved are significantly higher and therefore warrant more stringent regulation.

While generic solutions provide a base level of risk mitigation across different types of AI, technology-specific solutions are crucial to address the unique risks posed by AI applications in specific domains where the nature of additional risks on top of common risk are significant.

## 9. Given the importance of transparency across the AI lifecycle, please share your thoughts on:

1. Where and when transparency will be most critical and valuable to mitigate potential AI risks and to improve public trust and confidence in AI?
2. **Training Data**: The dataset used for training an AI model shapes the system's decision-making process and performance. Users have a right to understand what type of information the AI was trained on, as it can influence the outcomes they experience. For instance, an AI trained on biased data could make biased decisions affecting the user.
3. **Data Collection:** Users should be aware of how their data is being collected and used. This can impact their privacy rights and their trust in the system. If users understand the methods and purpose of data collection, they can make informed decisions about their interactions with the AI.
4. **Deployment & Disclaimer:** Clear disclaimers at the point of interaction are essential for managing user expectations and ensuring users understand they are interacting with an AI. This includes being aware of the AI's purpose, its limitations, and the extent of human involvement in the AI’s decisions.
5. **Post-deployment Monitoring & Issue Resolution:** Transparency in how AI systems are monitored post-deployment directly influences users' trust in the system. It matters to users how organizations respond to errors, data breaches, or any other adverse effects that may arise after the AI system has been deployed. Clear, transparent policies on these matters allow users to understand the measures that are in place to protect their interests.

In short, these four areas have a direct impact on users' experiences, their rights, and their trust in AI systems, making them particularly relevant and crucial in promoting transparency in AI.

1. Mandating transparency requirements across the private and public sectors, including how these requirements could be implemented.

Some suggested potential options (not mutually exclusive) include:

* **Legislation and Regulations:** Governments can pass laws requiring certain levels of transparency. These can include obligations to disclose data use, the presence of AI systems, their purpose, and the logic behind their recommendations.
* **Standardized Reporting:** A standardized framework for reporting on AI use can help ensure consistent, comparable information is available across different organizations and sectors.
* **Certifications:** Certifications can be given to organizations that meet high transparency standards, making it easier for the public to identify trustworthy AI-powered organizations, and providing an incentive for businesses to prioritize transparency.
* **Public Engagement:** Encourage public input on transparency standards. This can help ensure the standards are understandable and meaningful to end users.

Mandating transparency can be a powerful way to build public trust in AI, but it needs to be done carefully. Too much technical information can overwhelm non-experts, and too much mandate could stifle innovation, while oversimplified information may not fully represent the AI system's risks.

## 10. Do you have suggestions for:

1. Whether any high-risk AI applications or technologies should be banned completely?

While it's crucial to ensure the safe deployment of AI technologies, we must also consider the balance between risk management and fostering innovation. A blanket ban on certain types of AI applications could inadvertently hamper the advancement of technology and potentially discourage creativity and progress in the field. Therefore, an outright ban on any AI technology should not be the first resort.

However, we should indeed identify certain categories of AI use cases that present critical levels of risk. For these "critically high-risk" scenarios, it may be reasonable to apply stricter regulations to ensure safety and ethical considerations are thoroughly addressed.

For example, we could establish a regulatory process requiring comprehensive risk assessments and a specific approval process before these high-risk AI systems are deployed. The approval could be granted by a dedicated government body or a regulatory authority, following stringent review processes. This will ensure any AI technology posing a high risk is evaluated extensively for its safety and ethical implications before it's allowed to interact with the public.

In addition, it may be beneficial to have ongoing monitoring and regular audits of these high-risk AI applications once they are deployed. This will ensure they continue to operate as intended and any emerging risks can be promptly identified and mitigated.

By implementing such precautions, we can address the unique challenges posed by high-risk AI applications without stifling technological innovation and progress. This balanced approach aims to safeguard the well-being of the Australian public while also encouraging the growth and evolution of AI technologies.

1. Criteria or requirements to identify AI applications or technologies that should be banned, and in which contexts?

As noted before, a stringent regulation might be a better alternative than a general ban for “critically high-risk” applications. “Critically high-risk” applications could include use cases that could pose risk to life (e.g. medical use case, or autonomous weapons) or pose a threat to basic human rights (e.g. public surveillance system with facial recognition, which could have privacy implications). Identifying AI applications or technologies that should be potentially restricted requires a comprehensive assessment of their risks and impacts. Here are some suggestions for criteria that could be considered:

**1. Degree of Autonomy:** The higher the level of autonomy, the higher the risk. AI systems that operate with little or no human supervision in critical areas like healthcare, transportation, or defense could pose significant risks and might be candidates for stricter regulation or potential bans.

**2. Potential for Harm:** AI applications that could cause substantial physical, psychological, financial, or societal harm should be scrutinized closely. For instance, AI technology used to create deepfakes has potential for misuse that could lead to societal unrest or manipulation of public opinion.

**3. Invasiveness:** AI technologies that intrude on personal privacy, such as surveillance technologies or AI that collects sensitive personal information without consent, should be examined critically.

**4. Ethical Concerns:** AI applications that could potentially be used in unethical ways, or in contexts that challenge human rights and freedoms, may be candidates for restrictions or bans. An example might be AI used for social scoring systems, which can lead to discrimination or social inequality.

**5. Reliability and Transparency:** If an AI system lacks reliability, explainability, or transparency, particularly in high-stakes decisions, it could be a candidate for stricter regulation. For example, an AI system used for criminal sentencing that cannot explain its decision-making process could be considered too opaque and unreliable.

**6. Controllability:** AI technologies that can't be easily controlled or shut down, especially those using advanced techniques like reinforcement learning, might pose an unacceptable level of risk.

**7. Legal and Regulatory Compliance:** Any AI applications that can't comply with existing laws and regulations, especially those around data privacy and security, should face restrictions.

The specific context in which an AI technology is used should also be considered. For example, facial recognition technology might be acceptable in certain contexts like unlocking a smartphone, but problematic in others, like mass surveillance without consent. It's crucial to examine not only the technology itself, but also how, where, and by whom it's used. A balanced, nuanced approach to regulation can help ensure that AI serves as a tool for societal good without compromising safety, privacy, and ethical standards.

## 11. What initiatives or government action can increase public trust in AI deployment to encourage more people to use AI?

* **Clear Regulatory Framework:** Governments should establish clear and comprehensive regulations that address potential risks associated with AI. This provides assurance that AI systems are developed and used responsibly.
* **Mandate Transparency**: Governments should mandate transparency in the use of AI. This includes disclosing when and how AI is being used, explaining the decision-making processes of AI systems in understandable terms, and being open about any limitations or potential biases.
* **Education and Awareness:** Governments can sponsor educational campaigns to increase public understanding of AI. This includes explaining how AI works, how it's being used, and the benefits and risks associated with it. An informed public is more likely to trust and use AI.
* **Accountability Measures:** Governments should ensure there are robust mechanisms in place for accountability. This includes clear procedures for dealing with misuse of AI or harm caused by AI systems, as well as avenues for redress for individuals who feel they have been negatively impacted by an AI decision.

# Implications and infrastructure

## 12. How would banning high-risk activities (like social scoring or facial recognition technology in certain circumstances) impact Australia’s tech sector and our trade and exports with other countries?

It’s very difficult to make any reliable prediction about the economic outcome of a certain decision, but as long as Australia doesn’t generally ban application type / use-case type (as opposed to specific application like social scoring), the impact on tech sector and trade may be immaterial.

AI technologies are increasingly becoming generalized, meaning many different applications share common underlying technologies. Therefore, banning specific applications should not stifle development of underlying technologies.

Therefore, banning (or disallowing) specific high-risk activities via regulation (and not categorical ban) in the interest of public well-being is recommended.

## 13. What changes (if any) to Australian conformity infrastructure might be required to support assurance processes to mitigate against potential AI risks?

**Establishment of Dedicated Regulatory Bodies:** To oversee AI development, deployment, and monitoring, specialized regulatory bodies might be necessary. These bodies can set standards, assess compliance, and handle enforcement, ensuring a uniform approach to AI risk management.

The rapid and ever-accelerating pace of AI advancement underscores the need for dedicated regulatory bodies to oversee the various stages of AI development, deployment, and monitoring.

AI is a highly specialized field with a unique set of challenges and risks that may not be fully understood or effectively addressed by existing regulatory bodies, whose mandates are typically broader and cover a range of technologies. The complexities of AI, along with its high rate of progress, necessitate a focus and expertise that a dedicated regulatory body can provide.

Establishing dedicated regulatory bodies for AI would ensure that there are experts on hand who not only understand the technical aspects of AI, but also its ethical, social, and legal implications. These bodies could work in real-time with the AI landscape, reacting promptly to the fast-paced developments and emerging issues that are inherent in the field. They can set AI-specific standards, assess compliance in a nuanced and informed way, and handle enforcement in a manner that respects both the potential and the challenges of AI technology.

In a landscape where advancements can quickly outpace regulations, having dedicated AI regulatory bodies can create a more agile, responsive regulatory framework. This approach will ensure a uniform approach to AI risk management, enabling us to reap the benefits of AI innovations while protecting society from potential risks. This is particularly crucial given that AI technology is evolving faster than most areas of technology, underscoring the need for regulatory bodies that can keep pace with these developments.

# 

# Risk-based approaches

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

Yes. The risk-based approach prioritizes resources and efforts towards areas where the potential impact and likelihood of harm are the highest. It acknowledges that not all AI applications carry the same degree of risk. For instance, an AI chatbot recommending a coffee shop may carry much less risk than an AI system making decisions in healthcare or autonomous vehicles.

This approach is dynamic, allowing for adjustments as the AI system evolves, its uses change, or when the surrounding context shifts. It enables a more nuanced way of addressing AI risks compared to a one-size-fits-all approach.

However, while a risk-based approach has its merits, it's essential to remember:

* It needs to be accompanied by robust risk-assessment methods to accurately evaluate the potential risks associated with different AI applications.
* It should not lead to complacency about lower-risk AI applications, which, at scale or in certain contexts, could still result in substantial impacts.

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

State-of-the-art AI is an ever-evolving field, advancing at a remarkable pace that many leading AI experts believe to be exponential. A risk-based approach might be suitable today, but the duration of its relevance is uncertain given the rapid progress of AI capabilities. Recent trends in AI research consistently reveal the inherent limitations in our understanding of complex AI systems. The more advanced an AI foundation model becomes, the more challenging it becomes to predict and control its behaviors.

This observation implies that the risk levels associated with AI-powered systems are ever-changing, dictated by the rapid advancements in foundation models. Systems that seem innocuous and low risk today, such as a conversational chatbot for a retail company, could potentially become potent vectors for misuse or unintentional harm when their underlying models are upgraded to be more powerful or are deployed at a larger scale.

Given these dynamics, risk-based approaches may eventually become less suitable as AI models approach and exceed the capabilities of Artificial General Intelligence (AGI), transitioning towards Artificial Super Intelligence. Over time, we might need to explore and introduce alternative regulatory approaches, potentially based on the strength of AI models, to maintain effective AI governance.

The future of AI technologies and their societal impact is increasingly complex to predict. To navigate this uncertainty, it's crucial to maintain a tight feedback loop and an adaptable stance. The ability to reassess the landscape rapidly and pivot strategies as necessary may prove to be the most effective approach to tackle these challenges.

## 16. Is a risk-based approach better suited to some sectors, AI applications or organisations than others based on organisation size, AI maturity and resources?

A risk-based approach to AI regulation can be more appropriate for certain sectors, AI applications, or organisations based on factors like organisation size, AI maturity, and resources.

1. **Sectors:** Certain sectors have inherently higher stakes for AI deployment, such as healthcare, autonomous vehicles, and finance. These sectors, where the potential impact of AI failure could be life-threatening or economically disastrous, are particularly well-suited to a risk-based approach that prioritizes stringent regulations and safety measures.
2. **AI Applications:** High-stakes AI applications, regardless of the sector, should also adopt a risk-based approach. For instance, an AI system that makes recruitment decisions requires a different level of scrutiny compared to an AI system that recommends movies to viewers.
3. **Organisation Size:** Large organisations with considerable resources can often afford to adopt more extensive risk management measures. They can invest in comprehensive risk assessment, mitigation, and regular audits of AI systems. Smaller organisations, on the other hand, might struggle with the resource-intensive nature of robust risk-based approaches. This doesn't mean that smaller organisations should overlook AI risks, but rather they might need different, possibly sector-supported, mechanisms to manage those risks.

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

Yes, conditional to frequent re-assessment of currency & suitability.

## 18. How can an AI risk-based approach be incorporated into existing assessment frameworks (like privacy) or risk management processes to streamline and reduce potential duplication?

Incorporating an AI risk-based approach into existing assessment frameworks or risk management processes involves intertwining considerations of AI risks with current procedures to streamline the approach and reduce duplication.

A radical, yet potentially effective, method would be to treat sufficiently advanced AI entities as pseudo-legal entities or "person-like proxies" for regulatory purposes. This does not mean AI would have rights or be considered humans. Instead, it implies that we could use established frameworks of liability and responsibility to regulate AI actions. This approach might be controversial given its novel application to non-human entities, but it can provide a starting point for how we think about the accountability of AI systems.

For instance, if an AI system makes a decision leading to a breach of privacy, it would be held accountable in a manner similar to a human employee who committed the same breach. In this case, the responsibility would typically fall onto the organization that created, owns, and operates the AI system.

This approach can streamline the integration of AI risk management into existing frameworks. It leverages the already-established accountability structures and procedures to ensure the effective governance of AI systems.

However, it's crucial to note that this method should be implemented with careful thought, balancing the need for accountability with the unique characteristics and challenges that AI systems present. Furthermore, such an approach should be complemented by additional safeguards that specifically address AI-specific risks, such as biases in training data or the unpredictability of certain AI behaviors. This is especially important because, unlike humans, AI systems can't understand or adhere to ethical norms or social contracts without explicit programming, and even then, their compliance is limited to their training and design.

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

General purpose AI systems such as Large Language Models (LLMs) or Multimodal Foundation Models (MFMs) have the ability to perform a wide range of tasks across different domains. Given this versatility, the application of a risk-based approach to these systems presents unique challenges and considerations.

In the current state of AI development, the risk-based approach for these systems might not need to significantly differ from those for specific-purpose AI systems. The focus should primarily be on ensuring that these AI systems are used within their design and capability parameters, and that robust safeguards are put in place to mitigate the risks associated with misuse or errors.

However, as we advance towards more potent AI, namely Artificial General Intelligence (AGI) and eventually Artificial Super Intelligence (ASI), a simple risk-based approach may not suffice due to the sheer power and complexity of these systems. At this stage, we might need to adapt or extend the risk-based approach to also consider the underlying capabilities of the AI systems themselves, and not just their specific applications.

One such adaptation could be introducing 'capability-based' regulation, similar to the way we regulate electrical devices. In this scenario, most users would be allowed to operate low-capacity AI systems (analogous to low-voltage appliances), but the operation of more potent and complex AI systems (akin to high-voltage devices) would be tightly regulated and only allowed by licensed professionals. This would add a layer of security and control over the use of powerful AI systems, ensuring they are handled by individuals with the necessary expertise and understanding of the potential risks.

This doesn't mean that we completely abandon risk-based approach, but rather, we complement it with additional measures that recognize the unique risks posed by AGI and ASI-level AI systems. In such a complex, fast-evolving field, a multifaceted approach to risk management will be crucial.

In addition to implementing a more nuanced risk management approach, there's also the consideration of deliberately constraining the capabilities of general purpose AI systems. This concept, often referred to as 'bounded AI', involves restricting the AI's functions to align solely with the intended task or purpose. By limiting the operational scope of these AI systems, we can more predictably manage and mitigate the risks associated with their use.

For instance, a language model deployed for customer service might be bounded to respond to only a specific range of queries related to the service or product, preventing it from offering responses outside its domain of knowledge or purpose. By doing so, the potential for misuse or unexpected behaviour can be significantly reduced.

This is a pragmatic strategy, particularly for near-term AI governance. However, it's important to note that bounding doesn't eliminate all risks and the more we approach AGI-level AI capabilities, the more difficult it will be to enforce effective bounds due to the sheer complexity and adaptability of these systems. Therefore, it's essential to consider bounding as one part of a broader risk management framework, rather than a standalone solution.

## 20. 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:

Managing the potential risks associated with AI responsibly necessitates a combination of both self-regulation and mandated regulatory oversight, with the balance between the two contingent on the level of risk associated with each specific AI application.

For AI applications considered lower-risk, such as recommendation systems or chatbots, self-regulation by the organizations deploying these systems may suffice. Organizations have a vested interest in ensuring their systems work as intended and that they maintain the trust of their users. In this context, self-regulation would involve organizations implementing their own robust internal policies and checks to mitigate any potential risks, guided by industry best practices and ethical standards.

However, when it comes to higher-risk applications, such as autonomous vehicles, AI in healthcare or AI in high-stakes decision making, self-regulation might not be enough. These applications have the potential for significant, even life-altering consequences if things go wrong. Hence, a mandated risk-based approach through regulatory oversight becomes crucial.

To achieve this delicate balance, a close collaboration between AI developers, users, regulators, and other stakeholders is required. This collaborative approach will help create an environment where innovation thrives, but not at the expense of public safety and trust. Ultimately, the goal is to ensure AI serves the best interests of all members of society and their well-being.

1. public or private organisations or both?

Both public and private organisations should adhere to a risk-based approach for responsible AI. The potential risks and impacts of AI applications do not discriminate between public or private sectors. For example, an AI application used in public healthcare has similar risks to one used in private healthcare. Therefore, to maintain public trust and ensure safety, both sectors should adhere to the same standards and regulations.

1. developers or deployers or both?

The responsibilities of mitigating AI risks should be shared by both developers and deployers. Developers, as the creators of AI models, should ensure that these models are robust, fair, and transparent, and that they function as intended within the specified use-case parameters. Deployers, who use these models in real-world applications, should ensure they are used responsibly and within their intended use cases, and that appropriate safeguards are in place to monitor and manage any risks. This dual responsibility creates an end-to-end approach to risk mitigation, which spans the entire lifecycle of AI systems, from development through to deployment and use.