Safe and Responsible AI in Australia:

Questionnaire Responses

Trajecient

Definitions

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

Amended preferred definitions are as follows in italics:

Artificial intelligence (AI) refers to an engineered or bioengineered system that generates predictive outputs such as content, forecasts, recommendations or decisions for a given set of objectives or parameters without explicit programming to a significant level or with equivalent effects. AI systems are designed to operate with varying levels of automation.

The addition of 'bioengineered' is inclusive of emerging bioengineered systems.

The removal of 'human-defined' aims to be more inclusive of potential AI systems generated from other AI systems where objectives or parameters are arguably not human-defined. There is potential for this for neural nets and AI systems itself formed from generative AI, where results may be emergent.

The change from 'without explicit programming' to 'without explicit programming to a significant level or with equivalent effects' is in recognition that the creation of AI systems may involve significant use of explicit programming, both in relation to the algorithms but also operational adjustments, such as pre-defined responses to inappropriate requests or model adjustments to correct perceived bias. To be effective in actual use, it reframes the definition from what counts as being without explicit programming to whether the effect of the AI systems is equivalent to this to account for whenever the relevant risks apply.

Generative AI models generate content such as text, images, audio and code in response to prompts.

The removal of 'novel' is a necessary change as it implies a relationship to originality that is not necessarily the case – especially not as given output may be to produce a plot outline that only uses scène à faire, or elements so universal that they do not rise to the level of originality. Instead of named characters, archetypes may be used.

There is no issue with the definition of machine learning, except the potential issue that 'pattern' may not be as suitable for a plain language definition where either 'outputs' or 'reusable solution' would better convey the meaning.

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?

A great many risks are not covered by existing regulatory approaches, given the current emphasis on a voluntary code. A wider range of approaches are being taken but they are not at this point regulatory approaches, but rather action in relation to consultation and investigation.

The potential risks span the potential for implications for redress and litigation by the potential for a substantial increase in jurisdictional issues; legal gaps in relation to contestability given the scale and nature of AI systems; a lack of transparency and accountability in relation to AI systems; issues related to the protection of wellbeing, safety and intellectual property, including issues related to deliberate misconduct and other crimes; and issues regarding market access and value of outputs.

Trajecient has suggestions for possible regulatory action to properly address these risks and has provided, in a second attachment, a Consultative Submission which outlines recommendations and reasons.

This is done for reasons of length. It is of particular use due to the current process of consultation favouring stakeholders from established bodies and consequently, reduced representation of individuals and other groups who represent those who have yet to enter markets impacted by AI systems. All future creators will be in such a transitional stage at one point and there are certain risks to which future rather than current creators are particularly exposed.

Additionally, the *Discussion Paper on Safe and Responsible Use of AI* in Australia as well as the *Rapid Response Information Report on Generative AI* both did not explore such certain areas which have potential risks crucial to both discuss and consider as part of potential regulatory oversight.

Certain answers here are excerpts or summaries of the Consultative Submission.

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.

As detailed more fully in the Consultative Submission, a guide summarising the relevant laws and other contextual information to assist in the localisation process would help. It would especially help for localisations of AI systems developed by individuals or small groups as to emerge from the free and open source sector where such software already exists. It would also be helpful for other contexts where businesses or other organisations are looking to expand localisation of AI services into Australia but are unaware of what is involved.

The benefits of this is to improve accessibility and fairness is relation to Australians as well as compliance with laws and practices that may otherwise be overlooked, particularly in the case of AI systems which may be developed by hobbyists or internationally (but for use to better serve Australians).

Examples of what such a guide may cover are cultural sensitivities regarding Aboriginal art (which can improve generative AI), the existence of alternate Aboriginal place names for locations (which can improve a range of AI-based recommendation services or search engines which may come to exist) and the Interactive Gambling Act.

In the case of the latter, the distinction between online real money gambling by Australians not being an offence but offering online real money gambling services or advertising such services being an offence is potentially something that can be overlooked by potential hobbyist or smaller developers of AI systems for web design or ad services or recommendation services who may not be as diligent with research. They may check that online real money gambling by Australians is not an offence and not be aware the situation is more nuanced.

Again, there are already open source tools to create AI systems and consideration of regulation needs to bear this in mind.

Such a guide would also assist in voluntary compliance for AI systems which do not serve Australians and so not be covered by any Australian regulatory frameworks which come into place, but wish for such localisation to help improve intercultural respect and understanding.

Furthermore, such voluntary compliance where compliance cannot be compelled even in law is the best means to test broader applicability of voluntary standards.

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.

Fit for purpose coordination would involve the creation of an independent body to ensure active participation of stakeholders in an ongoing manner and to ensure coordination across all levels of government.

A duty of this body would be help and oversight that Australia's AI Ethics Principles is being implemented without compromise across all areas of government.

Robodebt is one example to illustrate that principles do not automatically follow from the operations of government agencies and/or associated partners.

Compromise can come from: political priorities in relation to funding; corruption; ideologically motivated hiring practices; lobbying; ideological motivations on the uptake of AI and/or societal values, or the extent to which human rights can be ignored with sufficient levels of financial compensation and/or strategic benefits.

A duty of such a body would be to enhance both review and development of governance laws, such as in reviewing the risk profile of applications and technologies in various contexts. It can make representations to government about areas of concern and oversee test programs. It has the authority to reverse or at least suspend the approval of grants should they be at risk of being in conflict with Australia's AI Ethics Framework.

Such a body can also coordinate compliance audits and other measures and make investigations to improve contestability, particularly to assist when power dynamics prevent individuals from taking greater action themselves.

Such a body should not be involved in the rollout of AI but be focused more on an oversight role. Existing bodies, such as the National AI Centre can be more focused on the rollout and implementation of AI. This avoids a conflict of interest where a body is asked to scrutinise itself.

This body can be an important means for industry groups and other sectors to play a role in co-development of policy. This can also involve consideration of when it may be fair and reasonable for certain compliance mechanisms to be relaxed, such as may be the case for trusted partners, perhaps due to other licensing requirements in place or due to voluntary codes being shown to be effective.

Such a body may be the existing Australian Competition and Consumer Commission but this would require a significant increase in funding and restructuring.

This body should be financially independent and members cannot be appointed by politicians but rather, membership should include representatives of various stakeholders as either voted by members of relevant peak association bodies or potentially by a more open public vote.

It would be recommended that potential representatives be drawn, or be able to be effectively represent, both larger organisations and businesses and smaller ones.

In light of Robodebt, it would be recommended that such a body include a representative from either the Australian Council of Social Service or in that area.

Representatives should involve representation of the Australian Voice or other levels of a Voice mechanism or equivalent thereof in its absence to represent the interests of Australian Aboriginal people. This is vital given potential for algorithmic bias and other bias and choice of approaches to impact them in particular ways.

Other recommendations for particular representatives would be to be able to represent the interests of Australia Police; of the legal profession and in particular in regards to international litigation; the film and television industry in Australia; separately, the writing industry in Australia (with a need to represent perspectives of both major publishers and independent publishers as well as freelancers); influencers (content creators on social media); business groups; unions; education (with a need to be able to represent the interests of the humanities and not just STEM areas); social scientists; Australians for AI Safetey and medical practitioners (with a need to also represent the views of those working with mental health).

The above representatives could sit on an Oversight Board or some other such mechanism of governance. Day-to-day operations would be run by someone appointed by this board. Representatives should be roughly split in half between those representing those primarily involved in developing or using AI systems and those at greatest risk, from usage, of facing significant negative harms. Any non-even split should favout those at greatest risk of facing signficant negative harms.

The Oversight Board would provide oversight and input into processes but would not need to be involved in day-to-day processes. In relation to reports and other submissions they would be able to make them either unanimously, individually, or in groups such as in the form of a Majority Opinion and a Supplementary Opinion so as to avoid some groups or interests being amplified over others.

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?

Germany's Artificial Intelligence strategy holds that AI applications must augment and supplement human performance. The German AI Inquiry Committee has moreover identified the importance to "ensure that, as social beings, humans have the opportunity to interact socially with other humans at their place of work, receive human feedback and see themselves as part of a workforce". This is thus a recognition that AI deployment could be done in an unsustainable way that causes high levels of backlash to overautomation that ignores psychosocial needs.

Italy has required ChatGPT to have an opt-out for user conversations to be used as training data.

Spain has a Royal Decree, the Rider Law, requiring key elements of algorithms to be disclosed to employee representations, involving the parameters of such algorithms and the motivating logic of such use.

In Portugal, clear limits on employee monitoring through use of AI systems has been put into effect.

At a state level, California is intending to allow opt-out rights for ADM.

Indonesia has required that all private digital services and platforms be registered to avoid being blocked by Internet Service Providers. This may be adapted so that registration at least in certain contexts be needed to assist with accountability and oversight but a lack of registration may not be in itself grounds for being blocked.

Canada has introduced legislation to establish a specific tribunal specific for privacy and data protection. It has proposed that on request, an explanation of how a decision by an AI system was reached be provided. It has also along with the European Union proposed banning certain uses of AI.

The European Union has the right not to be subject to certain forms of ADM but to be desirable a further extension of this right should be considered. There is a requirement to enable individuals to demand and access human intervention.

The AI Liability Directive is particularly laudable, adapting civil liability rules to ensure equal protection for victims of AI systems in relation to liability claims. More broadly, the European Union has established itself as the world leader when it cames to AI on the strength of the General Data Protection Regulation (GDPR) which intersects and interacts with forthcoming AI-specific regulations.

Although imperfect, the GDPR is currently the gold standard when it comes to policies in relation to privacy and data retention. It is also currently the world leader when it comes to enforcement of policies, with actions taken at local and state levels instead of the policy on paper not filtering through to policy of practice.

Whatever one thinks of the principles of the GDPR, there is no doubt that it has had a sweeping impact in setting the standard to which companies and individuals alike have responded. This has reached the extent that a section of the general public prefers to interact with European companies and platforms due to a greater confidence in the security, privacy and ethics of the approach.

Nations including Spain have also sought to establish 'sandboxes' so that these is a means for testing and development of AI systems in a secure, supervised way. This is an approach recommended by the forthcoming EU AI Act.

Target Areas

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

Different approaches should not apply. If something is an issue, it does not matter whether it is taking place in the context of the public or private sector. It is true that if there is a difference in how AI is used between these sectors, then there will br in essence different practices, but that is due to differences in use, not the user.

It is not the case that the public sector is without commercial pressures, such as may need to be done to attract or maintain funding. This is as true of charities and other NGOs as it is with research groups. The recent scandal involving consultancy provided by PwC is evidence public-private interactions are not risk-free.

Additionally, consideration of privacy when it comes to operational details of organisations which may arise as part of compliance and assurance processes applies both in relation to public and private bodies. Whether commercial-in-confidence information for the private sector or to prevent risks of gaming the system or otherwise better learning to engage in fraud, deceptive or misleading conduct in the public sector, there is a need to avoid overdisclosure.

An assumption that the public sector is more trustworthy is problematic. It is also problematic to assume that there is reduced need for transparency and accountability due to a 'public interest' or other 'fair use' defence.

It is through transparency and accountability that it can be determined that what it is being done is implemented to truly be ethical and virtuous.

It has been suggested that there are certain contexts, such as in regards to AI systems used for medical purposes, where consent should not apply to ensure that the systems have sufficiently comprehensive datasets for the greater public good.

This is well-intentioned, but would fail in practice.

Problems with a lack of comprehensive medical datasets has been seen with the case of My Health Record, but there the issue is not of consent, at least not primarily so. System design has been raised as a major issue. Another potential issue is not only the familiarity of pre-existing record systems, but that for private medical practice, separate record systems can be a means of encouraging patient retention, either by raising a passive barrier to going elsewhere to the lack of record sharing or alternatively by having systems be optimised for specific uses and staff.

To the extent that such issues do concern user consent, such as the belief that data could be given without consent to other government or government-linked services, such as the police, it is unwise to respond to such concerns by actually providing information without consent in ways that will be seen as validation.

Doing so only is likely to reinforce resistance to giving consent and thus to also access services, however distant from reality the concerns. Worse, it may be a pathway to radicalisation by adoption of other like-minded fringe beliefs.

This is particularly problematic as it may affect children and other dependents who would be swept up in such impacts.

Such concerns may, as in the case of diaspora groups, be due to experiences and internalised beliefs about the nature or potential nature of police forces shaped by genuinely oppressive and/or discriminatory police. It may be shaped by concerns about data security.

In both cases and other such cases, a better response is to earn trust via outreach, education and at times, an 'arms-length approach' to who stores and provides oversight of data (such as involvement of community groups or designated NGOs). This may of particular import to Aboriginal communities.

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

Although Trajecient recommends the same approaches be used across public and private sectors, there are some means to further support responsible AI practices by government agencies.

The first is to recognise government agencies may fail a 'duress by only resort' test. Hypothetically, let us say that there were no regulations against social scoring and such social scoring was used by Services Australia as a means of setting the level of Mutual Obligations for those on a Jobseeker Payment. Accepting this practice is a requirement for receiving a Jobseeker Payment.

There is no alternative to Services Australia in Australia. There are multiple providers of services to those who are part of the Services Australia system, but that system itself (including the Jobseeker Payment) has no similar alternative. Technically, there are alternatives in the form of getting work or getting direct financial support from family, friends or other individuals, but that is not an equivalently accessible substitute for those in need of such social support.

Therefore, given a 'take it or leave it' choice, people may be forced to 'take it' even if they have grave concerns about algorithmic bias or the potential for AI models to be purposefully, systematically designed in problematic ways. People may be forced to 'take it' even if concerns were actually to be founded in fact. Robodebt is an example of how such concerns can actually be valid.

As a result, the best means for the Australian Government to further support responsive AI practices in its own agencies is to ensure there are higher levels of human oversight and contestability than may otherwise be the case.

This is not calling for a separate approach between the private and public sectors. Rather, the same heightened standards should apply whenever a privacy company also fails a 'duress by only resort' test.

Such heightened oversight should include steps to ensure that oversight doesn't fail due to overreliance on a small number of individuals providing human oversight. Such individuals may be subject to compromise or fail to serve the public interest as they have ambitions and/or beliefs which align with serving power or profit rather than serving the people.

This could be done in part by heightened disclosure requirements when it comes to the receipt of gifts, political affiliations and such aspects. This could be done by the introduction of a Code of Conduct, with appropriate penalties to deter wrongdoing.

There may also be heightened standards in relation to the frequency and intensity of human oversight required and the extent of external human oversight of internal human oversight practices, including access to algorithms.

Although such measures could be desirable in general, it does also pose an added burden on individuals and organisations. The burden is such that it may be disproportionate to the risks when there are other genuine alternatives.

In a marketplace of ideas, there can also be a marketplace of risk. Different people can choose different levels of risk, such as by comparing the alternatives in relation to declared policies and their record.

As the importance of the nature of the operations rises and as the level of risk on a personal level increases, so does the chance of actually comparing alternatives.

When there are alternatives of proper equivalence, scrutiny does not need to be done at as exacting (and burdensome) a standard. Failures, if they occur, or appear more likely to occur, can be better managed as people can seek and find alternatives with less disruption and harm in the process.

Yet, there are times where a market does not properly exist. There may not actually be alternatives to be considered. When this happens, when service provision is so concentrated on one or a few providers, the potential harms of systemic failures are amplified as they can have ripple effects on those reliant on the operations of such an agency or other business or organisation.

The higher damage of failure make justifiable and reasonable a higher standard of accountability and oversight.

Government agencies also can take further steps in relation to how they interact with requests for AI systems or datasets to be used for international purposes, including requests for data to be transferred from one jurisdiction to another as may be the case for emigrants and requests related to cross-border law enforcement.

Responsible and ethical AI practices would be to consider whether there any potential harmful side-effects or consequences to such requests. This is to consider whether the information provided will only be used in a limited way or may be used as part of building capacities of heightened concern.

For example, a request for sharing data on international students from the nation making the request could theoretically be a means to advance the recruitment foreign agents in order to influence outcomes abroad to meet domestic agendas with use of AI models that use data to provide predictive results of who would be at greater likelihood of being successfully recruited.

There should be a mechanism by which intelligence agencies are able to provide information if there is ever such genuine potential for requests to be part of building capacities of heightened concern where cooperation across government agencies and other partners, international and intranational, needs to be of a more nuanced and careful nature so as to consider the level of such risks. A calibrated response would maximise the benefits and minimise potential harms in such areas.

Other potential examples would be requests of data with the side effect of acquiring training data to extend capabilty for known or likely AI-powered, state-sponsored disinformation campaigns, or domestic surveillance capabilities which may also serve a role in empowering forms of governance more likely to act contrary to Australia's interest, such as due to reduced cooperation or active opposition when it comes to foreign policy aims.

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.

Generic solutions to the risks of AI are most valuable when it comes to contexts where the data involved is already of a highly reliable nature. For example, in the use of machine learning which uses information limited to high-quality datasets such as the Australian Census or weather records from the Bureau of Meteorology.

Datasets which are less subject to algorithmic bias are where generic solutions are most applicable. Furthermore, contexts where the results show minimal risk of emergent properties, i.e. In relation to email services. By emergent, this is when the outcomes are more unpredictable due to the nature of the variables involved. Smaller AI systems with more specialised datasets that are more 'locked in' are more likely to be suited to more generic solutions.

Technology-specific solutions are more appropriate if the AI systems involve a continual, active flow of training data, (or just regularly altered databases) or is of a less reliable nature and thus more likely to introduce 'hallucinations' or other algorithmic bias which may require mitigation and correction not simply as it relates to training processes, but 'real-time patches' and other maintenance.

Larger systems with more complex algorithms and metadata arrangements are more likely to be where technology-specific solutions are more likely, as well as those in contexts liable to systemic bias or institutional bias.

Technology-specific solutions are thus more likely to be better in the contexts of recommendation services and all permittable uses of generative AI.

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

a. where and when transparency will be most critical and valuable to mitigate potential AI risks and to improve public trust and confidence in AI?

This question is answered in better detail in the Consultation Submission, but in essence transparency is most critical for generative AI (for permitted uses) due to the ongoing allegations of data scraping and incorporation of material with consent and without compensation, in violation of intellectual property rights, author rights and at times the capacity of individuals to enter into contracts regarding their work.

Furthermore, transparency is also most critical when it relates to disclosure of the existence and type of AI systems, whether dealing in ADM or not, due to the differing levels of risk and harms individuals and groups may face, with particular regard to collection of training data.

. In order for there to be contestability, fairness and accountability, there is a need for people to know when AI systems are in use. There is also a need to be able to scrutinise the type of systems involved, not simply by regulators but also by individuals due to the limitations of regulatory enforcement.

Of equal importance is for individuals to be able to know exactly what data has been collected or may be collected as again, the level of risk of such information can vary depending on a range of individual factors which AI systems and human oversight may not be able to take into account. This can include past cybersecurity breaches and the extent to which a person or group may have a public profile and be at risk of concerted campaigns of abuse and mental health.

Transparency is also needed to some degree when it comes to training processes due to the risks of insufficient training exacebating and/or failing to correct flaws in relation to AI systems, be it algorithmic bias, labelling issues in relation to content or the actual structure of algorithms.

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

This question is answered in the Consultative Submission but in essence there is a need to mandate transparency requirements because thus far there has been a widely recognised absence of transparency for many AI systems and there has already been cases where courts have found deliberate non-compliance.

Furthermore, allegations of data scraping in relation to generative AI in particular has undermined trust and credibility that a voluntary framework would be sufficient for certain companies. This is on top of the Facebook-Cambridge Analytica scandal and other events where information was wilfully collected in deceitful ways.

When it comes to the public sector, there is also a need to mandate requirements as the Sports Rorts, Robodebt and other such occurrences signify that there cannot be automatic trust in the public sector. The electoral importance played the desire of a National Anti-Corruption Commission is evidence of this.

Furthermore, even though there is a mechanism for scrutiny in the National Anti-Corruption Commission and other means for accountability after-the-fact, there is also a need for transparency so there is a means to avert harms, particularly harms of a more lasting and damaging nature. Prevention is better than more costly cures.

Robodebt is also just one example of how accountability mechanisms can take time to run their course. Transparency works from the instance of application and thus strong accountability measures do not replace a need for transparency measures.

Transparency mechanisms can be implemented in the form of disclosure and potentially by either or both data receipts and by conditional access to datasets.

Transparency can also be supported by limiting non-disclosure agreements from applying to training documents in relation to AI, which also has the capacity to provide additional protections should there be international outsourcing when it comes to human oversight of AI systems.

This also avoids banning them or requiring disclosure of training materials of a sensitive nature, especially in the event that governments are hacked (as they have been and in all likelihood shall continue to be in future).

This is also the risk posed by algorithms being shared by governments in any centralised way, as well intentioned as it may be to have a centralised reference point for such algorithms, or key aspects thereof.

10. Do you have suggestions for:

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

These responses may be more specifically in the Consultative Submission, including more detailed reasoning, potential exclusions and enforcement methods.

The more detained reasoning can include external references, which are not used in this document.

- Patents

At present, patent claims must be submitted by a person, but it is unclear to what degree this may be assisted by AI. There are great risks if AI can be permitted to assist writing of patents. This would enable 'patent squatting,' such as by reading research and anticipating future patentable developments without any intent to use the patent directly. Substantial licensing fees would raise barriers to innovation. Patent applications are still based on the premise nobody could beat the inventor to filing an application, but that was before predictive possibilities of AI.

- Generative AI Art as it relates to Aboriginal Art & Dreamtime Narratives

Likely to cause extensive financial harm to vulnerable communities and could contribute to dispossession by the disrespect for the spiritual dimensions of their art, including sacred designs. Technically also religious and racial discrimination.

- Generative AI Art involving Spiritual Expressions Related to Wellbeing

Beyond Australian Aboriginal peoples, there are other cultures around the world, including other vulnerable communities, where their culturally-situated art, including cultural symbols, is used for spiritual expressions which plays a vital role in their well-being. Such practices can be deeply embedded in worldviews.

Such generative AI art is likely to cause financial harm, offence and would also be religious and racial discrimination against their worldview.

- Generative AI to Write Wills

There is no requirement for lawyers to be use to create wills and there are people who already use software to help write wills. Would raise grounds to dispute a person 'really knew and intended what was being generated' especially when as the choice of individual words can be legally important. Even if limited to lawyers there is a risk of overreliance if correct most of the time as seen with self-driving vehicles.

- Social Scoring

Machine learning models are predictive and such prediction involves making certain kinds of assumptions, which may be subject to algorithmic bias or being swayed by propensity evidence which would not be acceptable in other legal contexts as evidence for judgements. There is also ample potential for human intervention and oversight to itself be subject to implicit bias or explicit bias. Because social scoring is in relation to judgements of high impact, the sensitive personal information such systems interact with and being itself inherently about measures of propensity, it is too high risk of use.

- Facial Recognition For Surveillance Systems

Thus far, related AI systems have noted issues with accurate detection of certain individuals but there is also a broader use which cannot be solved even if all algorithmic bias is removed: people can be very similar in appearance. This is obviously the case for identical twins, but also happens without being related. Identifying unknown individuals in large populations, creates heightened risk of systemic bias and overreliance if generally accurate. There is ample potential for human intervention and oversight to itself be subject to implicit bias or explicit bias. More accurate systems can be achieved with more information, but the nature of this information poses great cybersecurity issues disproportionate to benefits.

- Promptless, En Masse Use of Generative AI

AI systems can be created where, instead of using human inputs to provide results, prompts are itself generated by scripts, which may intersect with data scraping, optical character recognition and web crawling processes. Doing so would enable a flooding of material to platforms at a large scale. This can also be done in combination with bots in an attempt to influence recommendation systems, or as part of a kind of DDoS attack on a platform (which may be accompanied by requests of payment to cease and desist).

Use of AI systems in this way would not only be costly to platforms (both in straining technical and human resources and in damaging communities) but poses extensive economic risks not only to human creators, but to prompt engineers. Such use would poison AI systems influenced by recommendation systems, search engine result pages and the like, thus posing broader challenges for fairness, reliability and integrity of a range of systems.

- Generative AI Of Works Via Low Prompt Uses

As with promptless, en mass use, the ability for works, especially longform works to be generated with a minimum of prompts, as opposed to say prompt engineering being needed in relation to each paragraph or equivalent, would have similar effects. It can be used for DDoS attacks and have extensive economic risks by flooding markets particularly when it comes to 'clones' of recent releases or upcoming releases based on excerpts, snippers, trailers or other promotional material.

This in turn would have a stifling effect on innovation, be it of creative works or software, because the costs of development will be undermined by 'clones' provided at low-cost or for free. Furthermore, this would impact the value and thus demand of services related to advertisements, marketing and other promotional activities..

Economic risks of this are even greater for prompt engineers as there is likely to be similarity or overlap in relation to the algorithms and training data used, whether such low prompt usage is done by circumventing limitations on large AI systems or through the creation of other AI systems, such as by adapting open source code.

Such economic disruption to underlying supply and demand is likely to cause significant damage without being connected to growth elsewhere and could see market share being increasingly linked to the funding of individuals and/or groups with connections to the black market and/or organised crime.
The ongoing existence of piracy sites, 'clones' and counterfeit goods shows this is no mere hypothetical. It is related to ongoing damage that could intensify in future.

In the Consultative Submission, this type of generative AI is referred to as being 'piracy by prompt' as it is not the same as direct piracy, but has a similar intent and a similar result.

- AI Systems Which Infer Protected Characteristics

Should AI systems be used to infer protected characteristics, such as through analysis of training data on social media, this serves to increase the risks of enabling further bias (especially if soft law rather than hard law is the case) and heightens the potential severity of cybersecurity breaches. The potential benefits appear to be comparatively minor compared to the risks, particularly given if such details are not directly given with consent, inferring such details indicates the AI system involved is likely to be acting against the wishes of individuals, potentially in direct contradiction of direct refusal to consent to information being collected.

Such AI systems include those seeking to analyse such data for market research, those which may be embedded within algorithms for recommendation systems or in order to better serve advertising or search engine results.• AI Systems Which

Analyse Other AI Systems

It is possible for AI systems to be designed to analyse inputs and outputs to provide predictive information about other AI systems are designed. This becomes more accurate as more information is given and can inform both reverse engineering and to enable jailbreaking of pre-existing AI systems. It is also plausible, for example, to use such a method to test ways to circumvent identification of bots or spam.

There are legitimate research exceptions but in a more general context this is problematic where the potential for harms clearly outweights the potential benefits. It is at least possible, in terms of mitigation, to ban AI systems which directly offer a service of analysing other AI systems to the general public.

- AI Systems Which Analyse Signatures Or Digital Signatures

There is potential for AI systems to be designed to enable digital signature forgery or forgery of physical signatures. There are legitimate cryptographic research purposes to analyse digital signatures and for physical signatures for art historians, but a lack of mitigation in relation to AI systems usable to the general public would be problematic and the foreseeable harms are in clear excess of the potential benefits. Additionally, the ability for AI systems to analyse physical signatures can cause issues regarding forgery, including of art as well as of legal documents.

- Automated Trading

The use of ADM or other automated systems for trading stock or for automated purchase anf selling of goods or other products, or providing recommendation about such decisions (such as how to maximise markups and sales as a reseller) poses extreme risks for distorting markets. This can also relate to issues involving scalping such as that of tickets and limited edition gaming products. This is an issue as it can redirect spending that would otherwise have gone elsewhere, including supporting local businesses and services. On the stock market, it can spark broader issues for the economy by impacting stock markets in disruptive ways.

- Generative AI for Drafting Legislation and Decisions with Case Law Implications

There are limits to the effectiveness of human oversight and human review, particularly when it comes to attention, focus and perception. This is widely known in relation to proofreading, writing and editing. Sometimes you see what you expect to see rather than what is actually there.

Words and phrases are of high importance for interpretation. The systemic impacts of law and legal precedents, together with the potential for things to be overlooked in large documents, Generative AI is too risky in such contexts. At a minimum use of any LLM not designed with specific needs of legal professionals should not be used.

- Post-Launch ADM of Drones & Missiles

In a military context, the use of ADM for drones and missiles for post-launch operations would complicate the capacity to respond to a changing situation. Military situations, even with the best intelligence, can be subject to uncertainty. Potentially, after launch, multiple targets of opportunity moving or responding in different ways can force choices about post-launch direction. There is the potential for doubts to be raised by intelligence as a result of mid-flight footage or where the changing context of collatoral damage may entail reconsideration, such as in the event of human shield/hostage tactics, or positioning of targets in relation to critical infrastructure or sites of great emotional, cultural and other social significance,.

Algorithms are limited in that they can only take account on what is known in the algorithm and so, to quote Donald Rumsfield, they are very good at assigning probabilities to known knowns, but are less capable of assigning probabilities to known unknowns and unknown unknowns. Furthermore, there is the risk of ADM being used to detach decision-making from moral responsibility for outcomes. Considering the irreversible impacts of such decisions, use of ADM for such military applications is at too high a risk of reducing the need for nimble decision-making.

- Generative AI Which Undermines Copyright

Generative AI does not simply have the potential, without safeguards, to result in content being generated which directly violates copyright or trademarks, such as if it were possible to instruct large language models to get an output that appears to be an official letter from an organisation. This could perhaps be done if an AI system can be jailbreaked to circumvent usual restrictions to access training data that involves copyrighted or trademarked material but is not meant to be directly used for outputs.

Generative AI also has the potential to violate the spirit of copyright law and trademark law by the creation, including at large scales, of works that are as similar as possible without actually violating copyright or trademark.

To date, the idea that a style may need protection has been problematic, but the reality of AI systems is that it forces reconsideration. Perhaps the proper balance is to take steps against the intentional (rather than unintentional) recreation of the style of another, at least in certain contexts. This could include if such works are intentionally advertised or promoted as being in the style of another, especially if being offered in direct competition. This could involve considerations of whether such works are being offered at lower prices or for free and whether they are being made available in the same marketplaces, with awareness of the preceding creator.

Should laws not adequately protect against such uses, this will greatly undermine any potential for trust in the use of AI. Even responsible uses will be seen as enabling others to use it irresponsibly through normalisation of usage and by increasing the training data available (as use of such services can itself provide training data). Besides this foundational impact, not protecting against such uses would dilute the very foundations of copyright law and trademark law.

Such law was intended to ensure that innovation was rewarded. AI systems, without due safeguards, will mean only a narrow scope of innovation, related to AI systems, will be rewarded. Other innovative endeavours would not. This would be ultimately more stifling to innovation than due restraints on AI systems, because it relates to the conditions that determine the value of using AI systems themselves. If the outputs of AI systems are subject to being quickly reproduced or otherwise subjected to competition of an overwhelming intensity, the value of the outputs of AI systems, at least in an economic sense, would be minimal regardless of how innovative the construction of the AI systems.

Another reality is that the idea that a style may need protection must be considered to avoid discrimination regarding Australian Aboriginal peoples and other peoples where their cultures and worldviews are ones where collective ownership rather than individual authorship is deeply embedded.

In addition to the above, there is a further need for AI systems, as per existing laws, not to be used in furtherance of illicit acts such as defamation and direct violation of trademark and copyright law.

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

When it comes to criteria or requirements to identify the extent of applications or technologies which should be banned, a number of factors are relevant.

The first is the scale and proportion of harmful and beneficial outcomes. This is not just in relation to the intensity of impact, but also numerosity of impact. Consider a case where the negative impacts are not that severe but such impacts are expected to be predominant. The beneficial impacts are minor and limited in nature.

In such a context there is still a case to be made for banning.

Another factor to be considered are indirect risks. How do these applications or technologies interact with pre-existing behaviour, including behaviour that is already damaging? It is only logical that if new technologies can be used to extend or enact harmful practices which already happen, then those harms will also happen without suitable safeguards.

Such indirect risks include the potential for technologies or applications to be used to spread disinformation, for identity fraud and the risk of cybersecurity breaches. A crucial factor is whether the risks, both direct or indirect, can be controlled by human oversight, data privacy practices, or by limiting access. Where this is possible a general ban is unneeded. Yet there are other circumstances where the risks of significant negative harms from algorithmic bias or intentional use (or misuse) are so pervasive that potential safeguards are too inadequate.

This could be because oversight practices and other mitigating methods are themselves too likely to be compromised by the insidiousness of implicit bias or other kinds of systemic bias. It may be due to human limitations when it comes to perception and attention or due to economic limitations when it comes to applying safeguards in the form of cybersecurity, regulatory oversight or having enough data on smaller groups and intersectional groups to prevent algorithmic bias. It could the existence of other technologies that allow for circumvention of mitigating measures with only a little imagination and minimal expense.

The intended use of such AI applications or technologies may also be simply inappropriate for certain contexts where predictive results from systems which do not actually think are not fit for purpose.

General bans can in the context of general use but with specific exceptions. This is reflected in more general requirements when it comes to databases. There is information of a privileged nature, such as attorney-client information which can only be accessed in the context of that relationship.

Likewise, AI systems may in some cases be limited to researchers, or to law enforcement or the military or to educators or to health professionals, where it relates to information of a more private and sensitive nature, but necessary to a given sector.

Harms must not be narrowly defined, but span all impacts on wellbeing.

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

The biggest barriers to public trust in AI deployment relates to issues regarding a lack of sufficient safeguards, especially in relation to user agency. Studies have noted clear majorities in nations including Australia believe current safeguards are insufficient and that further regulation is needed. In particular, concerns are focused regarding privacy and the collection of use of data.

These views are grounded in concerns about information being without not only without consent, but when consent is effectively forced by services which cannot be unused for professional reasons, including needs in relation to branding, publicity, marketing, networking and public relations.

As seen in relation to responses to recent strike action in America, these views are also grounded in a simple idea: that the claims that increased use of AI systems will create more jobs is not reflected in their experience of economic systems.

It flies in the face of findings that supermarkets, tech and energy companies have increased prices in excess of increased costs in production from world events. It is against the understanding that comes from stagnant wage growth. Whatever claims can be made, it is the hip pocket that does not lie.

Economic experts who are global leaders in their field have pointed the finger at excessive mark-ups. This suggests with sufficient financial incentive, there are companies which will not hesitate to what is unethical but technically legal.

Consequently, strong transparency, accountability and opt-in and opt-out requirements are important. Education on processes and AI ethics is very helpful. Furthermore, sufficient regulation in relation to generative AI is likely to be essential. A failure to deal with concerns related to generative AI to provide sufficient protection to existing or future creators is certain to lead to heightened concerns.

Another strong barrier is the knowledge that politicians are so far not exposed to certain risks the way the general public is. They are not at risk of job displacement because at present, one must be a living human in order to be able to run for office.

Counterintuitively, one of the strongest ways to increase public trust in AI would be to make it possible for AI (or robots) to be eligible to run for office, provided the algorithms are fully public. Being eligible, of course, will not result in any hope of being elected unless they are genuinely perceived to be an improvement on current political representation. But it would address perceptions that politicians, due to the lack of such explosure, can be blinded to an accurate perception of the risks of AI. Although there is not a conflict of interest, it is arguable that there does exist a conflict of non-interest – being influenced by detachment from consequences.

The emotions with which this half-serious suggestion may cause can be instructive.

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?

Trajecient does not have advice to offer in this area.

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

Trajecient does not have advice to offer in this area.

Risk-based Approaches

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

Trajecient recommends a risk-based approach be used to provide an underlying level of mitigation. It should not be used as the only approach. The best approach is to provide agency for individuals and other users in engagement with deployers.

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

The main limitation of a risk-based approach (if rigidly applied) is the lack of flexibility to address more individual cases.

A scenario in the Consultative Submission is a case of domestic abuse, where what would normally be a desirable and safe use of AI for recommendation services is inappropriate in this instance.

In essence, a former partner, sought to control his wife when they were together by directly use her credit card and social media accounts. Consequently, the AI systems made recommendations which not only do not relate to her, but may recommend content could be traumatic by being direct reminders of domestic abuse and reinforce a sense that he is continuing to indirectly control her life. This could without redress plausibly even lead to suicide. If safety and responsibility are aims, this is not an acceptable outcome.

This is but one illustration of how individual factors can alter the calculus of risk. The extent of which an application or technology has for an person can vary according to factors such as the extent of potential impact from algorithmic bias; factors relating to mental health; the extent to which a person may have others actively seeking to harass, control or otherwise cause harm to an individual (or to them as part of a group), the extent of any prior cybersecurity breaches which may compound the potential risks of harm from future breaches; capacity to seek redress in the event of harm; and personal factors that can alter the level of risk such as the heightened severity for reputational damage, financial loss or other harms to which individuals with a public profile may be exposed, for example.

A risk-based approach often relies on normative judgements, based on perceptions and feelings rather than facts.

These may be the judgements made in the course of establishing a hierarchy of risks. These may be the judgements made by developers and/or deployers or potentially by bodies or associations in relation to specific sectors.

Regardless, a risk-based approach of 'top-down' nature mitigates risk in ways which are itself likely to result in unnecessary discrimination towards those who do not fit the normative mould. Besides the potential for systemic discrimination and for algorithmic bias to be reinforced, it is preventable by reasonable accomodations.

Bias includes overlooking diversity to prioritise normativity.

This failing of a risk-based approach is overcome by ensuring agency of customers, clients, consumers and other end users. After all, even when deployers strive to implement policies at a more fine-grained level, such as excluding the use of certain training data or keywords but only in relation to identified risks (such as in relation to known or declared mental health conditions), this has major issues.

The supply of such information of a personal nature would raise the risk profile of cybersecurity breaches to a significant degree – and in the event of deliberate intent to collect information for more criminal purposes. Such information would also often not be provided by meaningful consent. Forcing collection would be likely to have significant negative outcomes.

The use of predictive services or the routine supply of data from other sources to understand such information even with meaningful consent also poses a very high degree of risk and cost for data security.

This being said, a risk-based approach has an important benefit of recognising that different organisations have different limits in terms of a capacity for compliance.

Being able to set priorities in relation to safety measures implemented by developers and deployers is important for enabling them to be able to sustainably maintain a capacity for compliance and safety.

Having a risk-based approach provides regulatory certainty which AI developers and deployers can rely on to know where they stand in relation to requirements. There is lesser need to apply principles or tests particularly with regards to where to draw the line in relation to contestable concepts. There is less need to spend more resources in monitoring oversight decisions leading to changes about implementation.

The limitations of a risk-based approach can be overcome by doing what should be done to respect human-centred values and fairness.

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?

Capacity to follow a risk-based approach, or indeed, any regulation, depends on organisation size, resources, AI maturity and attitudes. That does not make the use of any regulatory approach less suited. Rather, it argues in favour of it.
There is a need to discourage reckless use where significant negative harms results due to ignoring the risks due to cost or burden. A risk-based approach can provide certainty about what these costs and burdens are, allowing for assessment of whether these can be sustainably implemented or if further capacity needs to be built or changes to a business model made to be ready to perform.

The best way to do this is through an approach that is about checks then balances. First determine by simple, streamlined questions what the level of general risk is and then go into a more detailed process if and as necessary.

Requirements under a risk-based approach, or factors determining the level of risk can be conditional according to the scale of deployment, as risk goes up as targets are more attractive targets for hacking, which in turn is dependent on the scale of information involved.

This being said, a risk-based approach can be supplemented by initiatives to help organisations and sectors sustainably adopt a risk-based approach, including in relation to providing help with infrastructure, access to guidance and help with accessing suitably qualified individuals to be involved in training and oversight.

A risk-based approach is better suited to so some applications than others.

This is because some applications involve higher levels of uncertainty. When there is increased uncertainty, it follows that risk assessment will also be less certain, particularly when it regards the impacts of smaller, vulnerable groups and when it concerns indirect impacts.

Consequently, there may be areas of high uncertainty where the extent and severity of harms may be too difficult to predict with sufficient levels of accuracy.
In such areas the choice is either to classify it as an area of higher risk, or to use some other approach. The downside of classifying the area as one of higher risk is that it could be stifling through being overbroad, particularly if there are requirements to check for things that do not apply in a specific context, or when risk assessment doesn’t have sufficient graduations or risk to account, for example, for differences in risk related to size, market share and potential conflicts of interest. The lack of such graduations could result in either or both underregulation and overregulation due too casting too wide a net.

AI systems may not be intentionally designed to be used to generate revenge porn to give one example, but the severity of harm means it is only ethical to consider potential unwanted applications as a possible use and to have sufficient safeguards.

Ignoring how measures can be circumvented or bypassed can mean ignoring the problem itself.

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

Some of the elements presented in Attachment C are inappropriate as even when a risk-based approach is in conjunction with other approaches, some of what is presented here would be contradicted by an standard that is sufficiently safe, unless the intent is that some people are to be considered less human than others.

Due to how the impacts and risks of systems differ depending on individual factors, there is a need for a system explanation to be provided on request, regardless of the risk level, together with a plain language notice. It is reasonable that depending on the level of risk a system explanation may be provided up-front. By provision of an explanation on request, this should be in form of a pop-up or conditional text displayed by user input. It should not require contact via email or via form.

The level of general impact and thus the extent of general processes outside of both opt-in or opt-out measures is a suitable way of organising risk levels. The level of ‘human-in-the-loop’ per risk level is appropriately set.

Indicative use cases is not the best approach to use.

Indicative use cases are most likely to be overrelied upon even when the actual levels of risk differs depending on a specific context. To give a simple example, when it comes to AI-enabled recommendation engines provided recommendations based on preferences, interests and browsing history, the level of risk in relation to the severity of cybersecurity breaches can be shaped depending on what preferences, risks and browsing history is factored in.

The disputes over indicative use cases in relation to their wording and selection is unnecessary as a better solution exists.

Instead of examples or indicative use cases, indicative statements are better. For example: ‘This AI system involves a high-volume, continual stream of data,” would be a statement associated with higher levels of risk, pointing to a greater need for other obligations. On the other hand, ‘AI system does not involve any human data’ would be associated with lower levels of risk.

It is important that such statements be specific enough to be usable without being too open to interpretation.

Such indicative statements have an added advantage in that it helps to capture circumstances where multiple AI systems interact with each other. This could involve sharing data or part of algorithms. It can involve circumstances where outputs of other AI systems, such as but not limited to search engine rankings or results of recommendation services, are used to impact results for other AI systems.

For example, if AI systems involved in hiring or firing processes takes into account search engine results to screen for causes to hire or fire a given person.

Such statements also help to prevent when general purpose AI systems might be segmented in order to fall under lower obligations but are part of the same broader AI ecosystem of system of systems in relation to datasets or data sharing practices.

Certain indicative statements may reappear for multiple categories using the same wording or using adjusted wording.

Monitoring and documentation per Attachment C is generally reasonable, but there should be at all levels a mechanism for external oversight should concerns of a sufficient degree be raised (but not done as a matter of course). Signalling how this process words would add clarity and increase regulatory certainty.

At all levels, there is a need for certain information related to training to be publically disclosable to ensure sufficient levels of accountability and scrutiny especially when it comes to vulnerable populations and in light of limitations of resourcing given to government oversight bodies. Such information related to training process does not need to be public by default, but rather could be done by limiting non-disclosure agreements.

The levels of explanation is reasonable in relation to the default standard of explanation given. Upon request, a specific explanation of an outcome should be available regardless of risk level. For example, in the case of AI used within a computer chess system, it is possible for a system to be biased when it comes to the use of certain strategies or series of moves. There would be a legitimate interest for wanting to check the fairness of the system.

Impact assessment per Attachment C is generally reasonable, but there must be a requirement at all levels to be open to external feedback as it relates to the process of impact assessment. This enables vulnerable populations and other cases where significant negative harms can or will result to make representations when they may otherwise be overlooked, which is more likely to be the case when it comes to groups overlooked by expert advice due to less data or being less represented in the field. This could come with easing of standards for impact assessment done 'in-house'.

The need to be open to feedback should not be limited to any window of time, except if there is any voluntary process of reassessment made on an reasonably regular timeframe to check for the implications of future technological changes, or other events (such as changing cybersecurity risks in relation to conduct of other states). At higher levels, it may be a requirement to not only be open to such feedback, but to actively communicate it. This can be required by public notice on a website or other kinds of public spaces, such as on social media platforms.

Even with opt-in and opt-out mechanisms, impact assessment is still important due to a need to consider the indirect impacts for use, which could have very significant outcomes (such as if the public distribution of an AI system can be reasonably foreseen to result in the creation of malicious AI systems).

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?

In terms of frameworks, this can be done by amendment of relevant legislation which will ripple through to amendment of relevant processes .

In terms of risk management processes, this can be done by incorporating and embedding into existing risk assessment processes checks that the relevant processes in relation to AI systems have been done. In some cases, separate questions will not need to be asked, but rather 'folded in' within answers to existing questions. This is a statement at the most general level and is no substitute to more nuanced understandings of specific industries, fields and circumstances.

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 would in general be higher levels of risk. This is explained in more detail in the Consultative Submission, but essentially the nature of such general purpose systems is that are at much greater risk of being misused. This is from when they are used in conjunction with separate AI systems deliberately designed to circumvent limitations. Such supplementary systems are essentially 'mounted' on top in order to make use of the benefits of general purpose AI systems, such as in relation to generating human-like language, but to override it and turn it to darker uses.

For example, a software program could be created by an individual or a group to automatically substitute words or phrases of praise with hate speech. This is not a large language model and it cannot produce human-like speech.

When fed the results of a large language model prompted to produce a congratulatory letter or email after say, reading a social media post and coming to a realisation about them, this nevertheless would result in human-like results with hate speech rather than praise. Through use of a variety of prompts, it would be possible to produce works of hate speech at larger scales, including works which shall not be readily identifiable as being the work of bots. Instead, the results would be of real harm because they would seem to be from real people, perhaps even specific people known to engage with a person.

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:

a. public or private organisations or both?
b. developers or deployers or both?

A risk-based approach for responsible AI, should one be implemented should apply to both developers and deployers, although what is required in relation to such risks would differ and the extent to which it should apply to developers should be conditional on several factors. These can be reflected in indicative questions.

The reason such an approach if implemented should apply to developers is because of the genuine risks of unregulated AI systems being developed by individuals or other groups who directly adapt or co-opt the work of developers. This may be by means of hacking or reverse engineering, which is more common than is generally known and is most publically evidenced in hacking communities around gaming, including the development of emulation.

As such, if developers create products that are too risky, it isn't enough that developers mitigate the risks if the work of the developers is too easily accessed.

The factors which should impact the extent to which a risk-based approach be applied to developers, to the extent such an approach is used, should first and foremost be how publically accessible the code is. Other potential factors can include whether it relies on bundled, integrated datasets of a more sensitive nature and whether AI systems or analysis thereof could provide an entry point to copy or otherwise gain access to datasets with potentially problematic consequences.

Furthermore, these factors can include if bundled documentation or the AI system itself provides insights into how algorithms work which can lead to reasonably forseeable harms. For example, the example of an AI system intended to automate certain aspects of an email system that describes the limitations of spam filters in such detail that it could inform circumvention of these systems. This could reasonably be foreseen to enable cybercriminals to increase exposure to scams and potentially scams of a convincing nature.

Developers of open source software can reasonably expect that their code will be looked at and even used by other people. As such, AI systems disseminated in such a public way can logically be expected to be the basis for AI systems developed by individuals and other groups, including cybercriminals, who show less or no regard to regulations which impact developers because of the nature of their use, be it criminal in nature or for academic dishonesty or other purposes. In considering regulatory approaches, the existence of criminals should not be disregarded.

If a risk-based approach were to be implemented, it should apply to developers because a risk-based approach means that, depending on the nature of what is involved there may be little to no requirements.

Consequently, it would not be an onerous burden for a risk-based approach to impact developers of AI systems where risk is minimal. It is true that there could be transitional issues of implementing new compliance standards. This only further suggests that the best way to structure such a system is to see if there is a streamlined way of knowing the level of risk which could involve some combination of infographics, a questionnaire, or other self-assessment tool. This could be in conjunction with resources for more precise assessment afterwards.

Should a risk-based approach be adopted, to the extent it is adopted it should apply to both public and private organisations. In practice, however, it would apply differently to public and private organisations, in relation to different practices happening at different frequencies.

For example, the public release of code is more likely to be done by public organisation than private organisations.

It may also be public and private organisations ultimately differ in terms of the type of AI systems they develop. It may be that in practice public organisations end up tending to develop AI systems which pose less risk or minimal risk (such as due to different practices regarding training data sources).

To the extent that a risk-based approach is adopted, whether it is mandated by law depends on the extent to which transparency, accountability and opt-out or opt-in practices are in place and are mandatory.

It is only logical that if there is insufficient transparency, accountability and either or both of opt-in or opt-out practices and there is also no mandatory regulation, it is to be expected that this could result is significant negative harms to an unwelcome degree occurring when they could otherwise be prevented.

Developers and/or deployers could argue after such harms took place that they were only following all relevant laws and in particular that any risks of harm were accepted by consumers, customers, clients and any other parties from the time they entered into a contractual relationship. Such actions may take place because without accountability and transparency mechanisms, conduct such as providing insufficient training or human oversight may not easily come to light.

After all, developers and deployers may rationalise to themselves and argue after such take place, if this was truly an issue there would be regulations against it. Is that not a central role of government?

Even if this a transparent strategy to deflect blame, it is one that is likely to resonate with the public, who are aware of regulations and laws affecting such matters as wearing seatbelts, zoning laws and product labels.

It is true that in the absence of specific regulation, it would still be possible for actions to be taken under existing laws, such as in relation to reckless and intentional invasions of privacy, or if injury is greater than emotional distress. In the absence of specific regulation of AI systems, relevant law would include that of contract law, privacy law and tort law. These all relate to civil law.

Generally, civil law requires victims to initiate proceedings. The difficulty here is that access to such legal redress can face barriers. For a start, the legal costs alone can be a barrier and particularly so for vulnerable groups. Additionally, while lawyers can take cases with different financial arrangements, such as agreeing to only be paid if a claim is successful, such arrangements are far likelier when a higher potential amount of damage is at stake – and higher damages are more likely when a case involves mandatory laws, such as under privacy laws, being breached and not just a voluntary code or potentially even a contract.

In the case of mandatory laws being breached, the standards to be followed are clearer than in disputes, such as those related to contract law, where there may be some aspects disputed which remain ambiguous. For example, the boundaries of what constitutes performance of a contract in relation to terms about or which relate to the performance or structure of AI systems.

There is greater pressure to settle for limited redress or accept no redress instead of proper redress when what is in breach is voluntary regulation or self-regulation rather than mandatory regulation, as legal action may be deemed to be too difficult. Because of the greater clarity mandatory regulation offers when a case involve such regulations not being followed (and as would be expected to apply to a contract even when unstated) the increased confidence influences the decision to pursue remedies.

Furthermore, sometimes victims still need, going forward, to use the services provided by those who did them harm. In such cases, there is a greater risk of justice denied out of fears that taking legal action would lead to adverse actions of one kind or another, particularly when going to court over those actions is too burdensome.

Mandatory regulation would permit greater scope for enforcement actions.

Furthermore, mandatory regulation would make for stronger and clearer communication when it comes for AI systems created not by companies or larger organisations but out of the actions of hobbyists or other individuals who may be less likely to be concerned about voluntary standards. This could be partly out of a belief such standards are intended for large or commercial projects only and not really for what they do. It could be out of a lack of awareness. It could be out of ideological opposition to regulation or other beliefs or attitudes that lead to a belief that such regulations do not matter to them (for example, this can be the case for developers with dark triad traits).

When it comes to areas of low risk or minimal risk there is potential for a hybrid approach with self-regulation or voluntary regulation applied for lower risk tiers, subject to change if there is sufficient evidence that mandatory regulations are needed for compliance or fairness.

The threshold of risk at which mandatory regulation is needed is lowered depending on the extent of transparency, accountability and opt-in and opt-out mechanisms are at a sufficiently high standard. This is because such mechanisms would protect against liability and thus reduce the potential for legal cases. More issues would be prevented in advance. Additionally, many issues would be pre-empted.

Consider a fire in two aircraft. In both cases smoke is visible before the fire breaks out and can be noticed by passengers. In the first plane, the emergency exits are operational, whereas in the second plane they are not. Passengers in the second plane are stuck where they are and can only hope for the best. Passengers in the first plane can supplement hope with added action, as they can evacuate the plane.

In other words, the more prevention the system enables, the less remedies needed for resulting harms.

When it comes to areas of lower risk, being at or towards the lowest end of the scale depending on the number of tiers, mandatory regulation is not needed because there is less risk of issues which require contestation in the legal system.

At this level it is more likely that sufficient resolution shall generally be the result from internal mechanisms on their own such as via a complaints process or opt-in and opt-out mechanisms. This is in part because at areas of lower risk, the disputes involve lesser potential harms (particularly when other areas of the law are not also in dispute).

To the extent there isn't sufficient resolution resulting from internal mechanisms, then at this level there is the option to not use a service or to go elsewhere. This is since the level of risk necessarily rises with the extent to which a company or other organisation is a dominant player, for this increases both the severity of systemic issues and severity of cybersecurity breaches were something to go wrong.

Furthermore, to the extent issues do arise, they are more likely to be less concerned with AI systems as opposed to other aspects of law. There is also reduced risk of issues in these areas stemmed from interaction of these applications with intentional criminality, such as in relation to hacking and ransomware.