Office of the Victorian Information Commissioner

t 1300 00 6842
e enquiries@ovic.vic.gov.au
w ovic.vic.gov.au

PO Box 24274
Melbourne Victoria 3001

31 May 2019

Artificial Intelligence
Strategic Policy Division
Department of Industry, Innovation and Science

By email only: artificial.intelligence@industry.gov.au

Dear Artificial Intelligence Team

Submission in response to the *Artificial Intelligence: Australia's Ethics Framework* Discussion Paper

The Office of the Victorian Information Commissioner (**OVIC**) is pleased to provide a submission in response to the Department of Industry, Innovation and Science's (**DIIS**) Discussion Paper on *Artificial Intelligence: Australia's Ethics Framework* (**the paper**).

OVIC is the primary regulator for information privacy, information security, and freedom of information in Victoria, and administers the *Privacy and Data Protection Act 2014* (**PDP Act**) and the *Freedom of Information Act 1982* (Vic) (**FOI Act**). As the Privacy and Data Protection Deputy Commissioner, I have a strong interest in matters that affect Victorians' information privacy and the security of data.

Artificial intelligence (**AI**) is a particular area of interest for my office, given the challenges and opportunities that AI poses to information privacy and the fundamental principles underpinning many privacy protection regimes around the world, including Victoria. Notably, OVIC published an *Artificial intelligence and privacy issues paper* (**AI paper**) in 2018, which explores some of the issues relevant to AI and information privacy.[1] In March 2019, OVIC made a submission to the Australian Human Rights Commission and World Economic Forum's White paper on *Artificial Intelligence: governance and leadership*.[2] AI continues to be a big focus area in OVIC's current and future work – in August 2019, my office will be publishing a book on AI, exploring the technical, legal and social aspects of AI, with contributions from prominent experts.

In presenting this submission, I would like to stress that OVIC does not view AI in a negative context. There are many benefits flowing from AI, and it is already in widespread use by consumers, businesses and governments. Many of these uses are beneficial and relatively low-risk – for example, AI is used in business to provide energy savings, with positive effects in terms of cost and carbon emissions. However, where AI interacts directly with individuals and their data, we see a need for increased caution and scrutiny.

[1] Published June 2018, available on the OVIC website at https://ovic.vic.gov.au/resource/artificial-intelligence-and-privacy/.
[2] Available on the OVIC website at https://ovic.vic.gov.au/resource/submission-to-the-australian-human-rights-commission-artificial-intelligence-governance-and-leadership-white-paper/.

This submission is organised around some of the key themes identified in the paper, and while OVIC has an interest in some of the broader issues canvassed in the paper, our remit means the submission is primarily focused on a privacy perspective. In this submission, I would like to highlight the importance of incorporating privacy considerations as part of an AI ethical framework. Privacy provides an important framework for making ethical choices about how we develop, use and regulate new technologies such as AI, and addressing privacy challenges will be essential to the long-term success of AI. Ensuring a balance between technological innovation and privacy considerations will help to promote the development of socially responsible AI that can assist the creation of public value.[3]

Promoting the privacy rights of individuals is also consistent with the Victorian public sector's obligation to act in accordance with the *Charter of Human Rights and Responsibilities Act 2006* (**the Charter**) and the requirement to take individuals' human rights into account – including the right to privacy – when making decisions.[4]

Definitions and terminology

The paper uses a number of different terms when referring to information about or relating to individuals – for example, 'personal data', 'private data', and 'personal information'. While the paper notes the definition of 'personal information' as contained in the federal *Privacy Act 1988* (**the Privacy Act**), it does not include a definition of 'private data' or 'personal data'.

While these terms may encompass the same or similar types of information, it is crucial to have consistent and clearly defined terminology in any proposed AI ethical framework to avoid confusion or ambiguity. Further, the concepts of 'personal data' and 'private data' are not found elsewhere in legislation in Australia, and their introduction in this proposed AI ethical framework is unlikely to assist in ensuring the information privacy rights of individuals are adequately protected. Using terms consistent with those in privacy legislation to describe concepts around personal information (noting that the definition of personal information varies slightly across jurisdictions in Australia) will be important for ensuring that the personal information used in AI development and systems is covered by privacy laws.

This issue of terminology also applies to 'sensitive information', which is distinct from personal information in the Privacy Act and the PDP Act, and carries a higher standard of protection in both. Any AI ethical framework must define sensitive information or data, and similarly, these definitions should be consistent with the definitions under privacy legislation. Further, the framework should distinguish between sensitive information as defined in privacy legislation and information that falls outside this definition, but which is considered to be sensitive for other reasons. For example, under the PDP Act sensitive information does not include financial information, however this type of information is often considered to be sensitive or delicate due to its highly personal nature.

Consent

Consent is a key issue raised in the paper, which notes that "protecting the consent process is fundamental to protecting privacy".[5] While consent is indeed a key tenet in many privacy protection regimes, such as in Victoria, and an important means for individuals to protect their privacy by allowing them to exercise control (in certain circumstances) over their personal information, its role should not be overstated. Under the PDP Act (and similarly, the Privacy Act), consent is not the sole legal authority permitting the collection, use and disclosure of personal and sensitive information. Any AI ethical framework must therefore recognise that, in many circumstances, personal information may be lawfully collected, used or disclosed without individuals' consent. This does not imply that organisations should collect data against the will of individuals; however as OVIC acknowledges below, in some cases the sharing of data is beneficial.

---

[3] *Artificial intelligence and privacy issues paper*, 2018, p 7.
[4] Available at www.legislation.vic.gov.au.
[5] On page 28 of the paper.

While we note that it may not be essential, there are other reasons OVIC questions the viability of a consent model in the context of AI. The limitations of consent are intensified by the complexity of the development, deployment and operation of AI. Obtaining meaningful consent can be particularly difficult in this context, as many individuals will lack the necessary knowledge about AI to provide informed consent. If individuals do not fully understand how AI applications or algorithms operate, or how personal information will be used, their ability to exercise choices about their personal information and provide informed consent is diminished.[6] Moreover, the interfaces in many current forms of AI, for example those used in consumer devices, lack a mechanism for individuals to adequately question or make choices regarding the operation of the AI, posing further difficulties for obtaining meaningful consent.

The dynamic nature of AI and its ability to extract meaning and draw inferences from data also means that even the developers and organisations designing and using AI applications cannot always anticipate how personal information will be used in the future. The potential downstream uses of data and decisions created by or as a result of the operation of the AI are often opaque at best.

Moreover, technological advances and the merging of newly created datasets increase the potential for personal information to be used for purposes beyond those for which it was initially collected.[7] Communicating to individuals all the potential uses to which their personal information will be put at the time of collection is therefore a significant challenge, and likely impossible.

The issue of explainability poses another limitation to a consent model in an AI context. While work is being done to build systems that can explain how a particular output was generated, there is currently no effective working product for explainability. Solving this problem is a significant area of current research, however it is unclear whether it can in fact be solved.[8] The ability to explain how a process has led to an outcome or output is integral to transparency and consent, and equally important to maintaining trust in AI systems; again, if organisations cannot adequately communicate AI processes to individuals, obtaining meaningful consent will be difficult. The need to include explainability as a core principle in an ethics framework before an effective working system for AI explainability actually exists may mean that adoption of AI needs to be delayed. This is a concern not only to regulators and civil society, but one that is also shared with many in software development. My office has met with software vendors who have indicated that a lack of effective interpretability poses a barrier to adoption in many markets where executives are called to account for their decisions.

The issue of explainability in turn raises that of intelligibility. The New Zealand Law Foundation's (**NZLF**) report *Government use of artificial intelligence in New Zealand* points out that even if explanations for how an AI reached a particular decision are available, whether individuals can actually comprehend those explanations is another matter altogether — this links back to the limitations of consent described above and the challenges for individuals in providing informed consent for processes that they do not understand.[9]

---

[6] See, for example, the US cases of consumer and health-care worker bafflement described in Lecher, Colin, *What Happens When An Algorithm Cuts Your Health Care*, The Verge, 21 March 2018, https://www.theverge.com/2018/3/21/17144260/healthcare-medicaid-algorithm-arkansas-cerebral-palsy.

[7] For multiple examples, see Eubanks, Virginia, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor,* 2018, St. Martin's Press.

[8] See, for example, the Defense Advanced Research Projects Agency's Explainable Artificial Intelligence (XAI) Project at https://www.darpa.mil/program/explainable-artificial-intelligence; Kim et al, 2018, *Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)*, available at https://arxiv.org/abs/1711.11279.

[9] See page 41 of *Government use of artificial intelligence in New Zealand*, New Zealand Law Foundation, 2019, available at https://www.lawfoundation.org.nz/wp-content/uploads/2019/05/2016_ILP_10_AILNZ-Report-released-27.5.2019.pdf.

Finally, meaningful explanations may be available, but whether they are accessible is again another issue. Certain economic, political or legal restrictions (for example, intellectual property rights or commercial sensitivity) may prevent the disclosure of the AI algorithm or code, or access to the training data. As the NZLF notes, "algorithms that are otherwise technically transparent may therefore be "opaque" for nontechnical reasons".[10] AI developers and operators will face challenges in obtaining informed consent if they cannot communicate such explanations to individuals, whether for legal or other reasons.

In light of the issues associated with the traditional consent model, OVIC is of the view that relying on consent as the primary mechanism to protect individuals' privacy, as suggested by the paper, has substantial limitations and may be inappropriate in the context of AI. However, in order to protect the human rights of individuals interacting with AI, some form of improved legislative protection will be necessary to provide sufficient controls on the risks identified in both the paper and some of the sources cited within.

Open data sources and re-identification

OVIC strongly agrees that AI's ability to detect patterns and infer information from publicly available 'non-sensitive' data (including data that has been de-identified or is considered to be 'non-personal') can potentially reveal individuals' identities and therefore pose a risk to their privacy. It is therefore important to note that even where information has been de-identified, it may still constitute personal information due to the risk of re-identification.

Although the paper acknowledges the risk of re-identification in the context of 'non-sensitive' data that has been publicly sourced, the potential for re-identification where de-identified data is shared privately between parties, or where de-identified datasets are combined with other datasets, should also be considered.

Further, the risk of re-identification is present not just in relation to data that has been obtained for use in AI systems, whether from an open data source or otherwise. Data generated by AI may also carry a risk of re-identification, again even where such data appears to be non-personal or de-identified. Organisations using AI systems may not be aware of the broader context in which AI-generated data may be used beyond their program, since, as identified above, this is often impossible.

Given the risk of re-identification in the context of AI, OVIC suggests that this issue be explored further and consideration be given to potential safeguards in relation to the collection or use of de-identified data in AI, whether that be part of an ethical framework, standards, guidelines or other instrument. The possibility of strictly limiting downstream collection of data from the outputs of an AI is one area that may be considered.

Principle 1: Generates net-benefits

In relation to Principle 1, OVIC strongly recommends that further consideration be given to the notion of 'net-benefits'. Principle 1 states that AI systems "must generate benefits for people that are greater than the costs",[11] but does not expand on whether these benefits can include benefits only for certain groups (for example, government or industry), or whether the benefits must be for the wider community. Any AI ethical framework needs to clearly define what constitutes a net benefit, and for whom the net benefit is intended.

Consideration must also be given to how net benefits can be measured, and who should be responsible for assessing them. This will require the establishment of criteria for assessing net benefits, which should incorporate privacy considerations — whether an AI system generates a net benefit should be balanced against the potential cost to individuals' human rights, including the right to privacy.

---

[10] Ibid, page 41.
[11] On page 6 of the paper.

It should be noted that the assessment of net benefit is a subject that has been debated long before AI existed. Importantly, it is not a discussion that should be undertaken within the AI community alone. As AI has whole-of-society impacts, discussions around 'net benefits' should involve input from a very wide cross section of the community.

Principle 4: Privacy protection

Principle 4 states that systems (including AI systems) "must ensure peoples' private data is protected and kept confidential plus prevent data breaches which could cause reputational, psychological, financial, professional or other types of harm".[12] As noted earlier, the concept of 'private data' needs to be defined and should be consistent with definitions under privacy legislation, to ensure that such information is protected by privacy laws.

Further clarity around who the data should be protected and kept confidential from would also be valuable, noting, however, that privacy legislation is not intended to prevent the sharing of personal information. It is better seen as an enabler for information sharing, with appropriate limitations to uphold individuals' right to privacy. In some cases, it may be appropriate or necessary for AI systems to share data, however this should be done in accordance with relevant laws, including privacy legislation.

As framed in the paper, this principle appears to be primarily focused on ensuring the security of information. While security is certainly an important element, protecting the privacy of personal information needs to go beyond security to look at the broader principles underpinning information privacy, and which form the basis of many privacy protection regimes, including the PDP Act. These principles include collection limitation (limiting the collection of personal information to only what is necessary), purpose specification (informing individuals of the purpose or purposes of collecting the information, at the time of collection), and use limitation (using or disclosing personal information only for the purposes for which it was collected).[13]

Consistent with the collection minimisation principle, OVIC suggests that any AI ethical framework should encourage the use of the least amount of personal information needed, and the use of less sensitive information where possible and appropriate. Further, in each and every case, an argument should be made for why specific data is required to perform the necessary decision making, or why it should be provided as an output of the AI.

Effective privacy protection is fundamentally about allowing individuals to control their own data. To the extent that an AI makes this more difficult, it will have a diminishing effect on human rights.

Principle 5: Fairness

In principle, OVIC welcomes the inclusion of fairness as a core guiding principle for ethical AI. However, OVIC also recognises that implementing this principle in practice comes with significant challenges, and further consideration and public discussion is needed around what constitutes fairness in AI. Many quantitative measures of fairness rest on implicit assumptions about fairness in society, and these assumptions are often mutually incompatible.[14]

A key point to note is that algorithmic fairness is not simply mathematically calculated. As acknowledged in the paper, notions of fairness vary.[15] Fairness can be socially and culturally determined, and moreover, must be considered contextually – what is considered fair in one context or by one particular group may not be in another context, or by another group. The proposed AI ethical framework should detail how fairness will or should be assessed, and in doing so, identify the lens through which fairness is being

---

[12] Ibid.

[13] Conversely, the nature of AI is challenging these very principles. See pages 9 – 11 of OVIC's *Artificial intelligence and privacy issues paper* for more information.

[14] See Friedler, Scheidegger and Venkatasubramanian, 2016, *On the (im)possibility of fairness*, available at https://arxiv.org/pdf/1609.07236.pdf.

[15] On page 41 of the paper.

considered (for example, whether fairness is being assessed through the lens of certain legislation such as the *Racial Discrimination Act 1975* or the *Sex Discrimination Act 1984*).

Importantly, assessments of fairness should not be based on legal considerations alone. Fairness is not just a legal issue; it is also a philosophical and social issue, and these perspectives should be taken into account when developing criteria to assess algorithmic fairness. Further, there are different dimensions to fairness. The European Commission's (**EU**) *Ethics Guidelines for Trustworthy AI*, which also includes fairness as a key principle for ethical and robust AI, proposes that fairness has a substantive and procedural dimension. The former relates to equal and just distribution of benefits and costs, and a commitment to ensuring individuals and groups are free from unfair bias and discrimination, while the procedural dimension involves the ability to contest and seek effective redress against decisions made by AI systems.[16]

Principle 5 also notes that particular attention must be given to ensuring that training data is free from bias or characteristics that may cause an algorithm to behave unfairly. However, bias is not only present or created in the collection and preparation of data used as input into an AI system – bias can also be created in the design of a system's objectives, or can develop as more data is introduced to the system, especially when it is uncontrolled.[17] As such, fairness should be assessed throughout the lifetime of AI systems, especially when new data is introduced, to ensure that they remain fair and free from bias. However, determining whether bias is present is another challenge in itself, as to date, there is no singular baseline or set of standards to evaluate bias. Bias may emerge even where measures are in place to limit the potential for it to occur, and indeed, it remains to be seen whether bias-free AI can ever be developed.

Bias itself is a somewhat loaded term. Data scientists usually prefer to discuss weighting of features in a dataset, as opposed to bias in the dataset. It is, generally speaking, impossible to have a completely evenly weighted set of features for an AI to act upon – weighting for one feature (to overcome perceived bias) may in turn determine increased weighting on other features, and those features may give rise to unintended consequences.

As mentioned briefly earlier, any discussion of fairness must also refer to the cultural factors involved in an assessment. As the research underpinning the MIT Moral Machine experiment cited in the paper indicated, the fairness of an outcome may depend upon the culture it impacts upon.[18] People from differing socioeconomic or ethnic backgrounds may have very different views on fairness than the operators or developers of AI systems, and adhering to a fairness principle may require AI operators to engage anthropological expertise to ensure they understand these factors when designing or operating systems.

Fairness and bias therefore need to be considered not only in relation to input data or the generated outcomes and output, but also in the very design, development and application of an AI system. Further, where a system is found to contain bias or does not meet this fairness principle, AI developers and organisations using AI will need to consider the possibility that the system or investment may need to be discarded.[19] Biased or unfair AI systems are rarely fixed by 'tweaking' them;[20] they must usually be re-trained from scratch with new data, or discarded entirely. In turn, this could present a political or reputational risk, particularly where investment has been significant.

---

[16] *Ethics Guidelines for Trustworthy AI*, High-Level Expert Group on Artificial Intelligence, European Commission, 2019, available at https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[17] See, for example, the Tay case study on page 31 of the paper.

[18] See page 50 of the paper, which references Awad E, Dsouza S, Kim R et al. 2018. The Moral Machine experiment. Nature, 563(7729): 59-64.

[19] This is reflected in the Termination Obligation of the Electronic Privacy Information Center's *Universal Guidelines for Artificial Intelligence* (2018), which states that an institution that has established an AI system 'has an affirmative obligation to terminate the system if human control of the system is no longer possible'. A broader obligation to terminate where bias cannot be removed without other negative effects may need to be considered. See, for example, Amazon.com and its recruiting algorithm, which was shown to have introduced bias against women and subsequently scrapped by the company after the bias in that feature was removed but the company became aware of other discriminatory weightings – https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

[20] See, for example, Alex Hern, 'Google's solution to accidental algorithmic racism: ban gorillas', *The Guardian* (online, 13 Jan 2018), available at https://www.theguardian.com/technology/2018/jan/12/google-racism-ban-gorilla-black-people.

There must also be consideration as to who should be responsible for assessing the fairness of AI. The public needs to have confidence that systems are fair, however this is challenging with AI systems as their datasets often cannot be released publicly, and in any case, many forms of machine learning that underpin AI systems rely on techniques not amenable to interpretation (for example, so-called 'hidden' layers in neural networking, or Generative Adversarial Networks). A possible solution to this would be to establish an independent regulator with sufficient resources, expertise and oversight to properly assess and audit fairness in AI systems; however, this will only be effective as a control where the regulator has the power to compel information on systems and enforce suspension or removal of approval for the operation of the AI.

Principle 7: Contestability

As noted above, the ability to contest the use or output of a particular AI algorithm is an important dimension of fairness. OVIC therefore welcomes the inclusion of contestability as one of the core principles for ethical AI. However, the ability to challenge the use or output of an algorithm should not be limited to the individual directly impacted by the algorithm. Other individuals should similarly be able to challenge algorithms on behalf of affected individuals (for example, carers, lawyers, human rights groups), and this ability to contest should also extend to individuals who may be impacted by the AI algorithm in the future.

Implementing this principle in practice will require government and businesses to provide pathways to contestability, and these pathways need to be clearly communicated to individuals. Importantly, in order for individuals to contest an algorithm, the entity accountable for the decision or output must be identifiable and decision-making processes should be explicable;[21] as identified earlier, however, explainability remains a significant challenge in AI that is currently not able to be easily solved.

AI Toolkit

In relation to the AI toolkit proposed in the paper, OVIC provides the following comments:

- *Risk Assessment Framework*: while OVIC recognises that the frameworks contained on pages 63 and 64 of the paper are only examples, any risk assessment framework for AI systems should not be overly reliant on consent or consider it to be the only privacy-related risk. Further, any example frameworks provided as part of an AI toolkit (which entities developing or using AI may use as a basis for their own frameworks) that includes consent as a potential risk area needs to carefully consider how this risk is assessed and framed.

  For example, the 'Regulatory and legal Compliance/Likelihood' column of the table on page 64 posits that risk is insignificant where consent is gained for the use of data. However, this is not necessarily the case; privacy risks may arise even where consent is obtained (for example, the risk of overcollection of personal information, or the individual's lack of understanding of what they have consented to). Similarly, the lack of consent does not necessarily mean that the regulatory and compliance risk is 'critical' – as noted earlier, consent is not the only legal basis upon which personal information may be collected, used or disclosed.

  Consideration also needs to be given as to whether risk assessments can be adequately conducted for AI systems at all while explainability remains an issue – how can entities conduct risk assessments if they cannot understand the factors involved in their AI systems?

[21] *Ethics Guidelines for Trustworthy AI*, High-Level Expert Group on Artificial Intelligence, European Commission, 2019, page 13.

Much of the paper anticipates risk in the design or development phase of AI, but as the paper briefly notes, AI risks can continue after the AI has been completed and deployed.[22] Risk assessment may be difficult to conduct in situations where the inputs to the AI cannot be anticipated. While many forms of AI are developed from structured, defined datasets, some AI is designed to be trained using live data inputs. There is a special risk associated with these kinds of AI. For example, an automated vehicle may be manipulated into an inappropriate manouvre if an adversarial attack is made using painted lines on a road, but it might also behave inappropriately if a speed limit sign is removed during maintenance or by an adversary.[23] Adversarial attacks can come in many forms, including the example of the Microsoft Tay chatbot cited in the paper.[24] Failure to adequately assess potential pathways to dangerous AI represents one of the greatest risks in AI development and operation.[25]

Finally, a standard for risk assessment should be indicated to assist AI users to apply best practice assessments. The Victorian Protective Data Security Framework, which my office administers, points to ISO 31000:2009 and ISO 31010:2009 as relevant standards. OVIC's experience is that risk assessment capability varies widely between organisations and individuals, and that a program of education supporting maturity in assessments will be an important part of any risk assessment framework proposed as part of an AI ethical framework.

- *Best practice guidelines*: To support Principle 4 ('Privacy protection'), OVIC suggests that best practice guidelines encompass data breach response plans, or guidance on responding to data breaches. This will assist AI users to effectively and appropriately respond to data breaches specifically involving the use of AI systems or data, which may be more complex compared to other types of data breaches. In the absence of clear interpretability, it is possible that AI operators may be confused by breaches involving AI or AI generated data, and in particular by any re-identification issues, which may impact their willingness to acknowledge issues and respond to them appropriately.

  OVIC's experience in providing guidance and templates for privacy impact assessments (**PIAs**) gives us a further element of concern in relation to this AI toolkit – that guidelines and templates may be seen merely as a "check box" or compliance exercise, rather than as a guide to whole-of-life development and operation of a system. We regularly provide advice to organisations on the assessment of privacy impacts, but we stress that PIA guidelines are only part of the journey and that privacy needs to form an integral part of system design and operation. If this AI toolbox, and the guidelines it contains, are used in a way that sees them as a hurdle to be overcome rather than a set of principles to be continuously applied, the probability of ethical AI development will be reduced.

- *Consultation*: OVIC agrees that public consultation is imperative to ensuring the development, use, and regulation of AI aligns with community expectations.[26] Consultation plays an important role in fostering public acceptance and uptake of AI systems, and helps build social licence for developers, operators and users to collect and use individuals' personal information for AI systems. However, consultation needs to be a continuous process of information sharing and feedback from stakeholders, AI developers and users, not a one-off exercise.

---

[22] On page 31 of the paper.
[23] An argument against this is that adversarial attacks on infrastructure (such as vandalism) occur frequently even when AI is not present. However, in current real-world examples, humans are capable of exercising judgement, and depending upon the circumstances this may not be possible for an AI. Further, there will be a tendency by government and business to blame the vandalism for adverse outcomes, when the question that may be better asked is: *should the design of the AI have anticipated the possibility of adversarial action?*
[24] See page 31 of the paper.
[25] See Yampolskiy, Roman V, 2015, Taxonomy of Pathways to Dangerous AI, in proceedings of 2nd International Workshop on AI, Ethics and Society (AIEthicsSociety2016). Pages 143-148. Phoenix, Arizona, USA. February 12-13th, 2016, available at https://arxiv.org/abs/1511.03246.
[26] On page 62 of the paper.

Public consultation should also cater for minority or disadvantaged groups to whom usual avenues for consultation may not be as accessible (for example, young people or Indigenous groups).

Additionally, the paper notes that "consultation with various stakeholders including the general public, academics and industry members is of critical importance when developing AI regulations".[27] While OVIC strongly supports further public discussions on issues such as net benefits and fairness, and agrees on the need to gather different perspectives from a diverse range of stakeholders, government must ensure that the development of regulations (and other governance mechanisms) takes into account the views and potentially competing interests of each stakeholder group, not only those with the greatest capacity to participate in the consultation process. Further, these interests also need to be balanced against the broader public interest and the protection of individuals' human rights.

General comments

Project governance and ethics

The paper's proposed AI ethical framework currently lacks a detailed discussion on the lifecycle of AI and its impacts on governance and ethics. Many – if not most – IT initiatives in government and business are undertaken as projects, rather than being driven as ongoing continuous development initiatives. There are two issues that arise from this practice.

- *Inadequate governance:* once a business case has been made and funding secured for an AI project, it is in the nature of business and government to expect a return on the investment. However, as noted earlier, a biased AI cannot be easily retrained and must often be discarded. Any AI project will therefore likely need, at a minimum, a precursor project to look at the data anticipated to be used in the development of the AI (the 'training set') and determine whether or not the features in that training set are suitable for the intended purpose, whether there might be features in that set that are problematic from a human rights perspective, or whether the AI may be subject to adversarial attacks based on data inputs.

  In the event that the data available to support the AI is unable to be used without unfortunate features developing in the AI, it may be advisable to terminate the project entirely (see the Amazon HR example cited earlier). While this may be typically regarded as a failure, in data science it is a learning experience. In many domains, a new level of sophistication at senior management and board level will be required for the maturity necessary to support responsible AI development and operation.

  The paper also references the concept of Humans in the Loop (HITL) and Humans on the Loop (HOTL). As much of the research in this area indicates, human-AI supervision is only an effective control when the individuals in or on the loop actually understand the limits and capabilities of the systems they administer.

- *Inadequate knowledge retention:* if an AI is intended to operate for a period of time after the dissolution of the project team involved in the development and implementation the system, special care will be needed to not only document the workings and domain limitations of the AI, but to ensure that management understands the learnings from its development.

[27] Ibid.

In government and business there is a tendency to see IT systems as assets, and often a desire to build on and extend their use to maximise asset value. In AI – more than other areas of information technology – this is fraught with risk, as AIs are (until the advent of General Artificial Intelligence, should that ever occur) remarkably domain specific. In the current state, each AI is trained or defined within a set of defined circumstances. AI operators constantly need to be aware of the constraints of those circumstances.

Recent commercial research also indicates that most companies engaged in AI development struggle with data quality and training, indicating a low level of maturity and a consequent desire to outsource to overcome perceived weaknesses.[28] Outsourcing is almost certain to have unfortunate flow-on effects in terms of an understanding of the domain limits of the AI and the relevant features in the dataset used for training, increasing the risk of inappropriate use or imperfect management of the operation of the AI once the contracted vendor completes the work.

Access to information

In considering the concept of transparency around the use of AI (Principle 6), consideration needs to be given to existing access rights to information (including personal information), which in Victoria is provided for primarily under the FOI Act.[29] Principle 6 states that individuals "must be informed when an algorithm is being used that impacts them and they should be provided with information about what information the algorithm uses to make decision".[30] This is reflected in the object of the FOI Act, which is to extend the public's right to access information about rules and practices that affect them in their dealings with agencies.[31] The public access schemes of other jurisdictions, including at the Commonwealth level, should also be taken into account.

Thank you for the opportunity to provide a submission to the paper. Ethics in AI – and AI more broadly – is a pressing issue and considerable investment and further public discussion on these matters is crucial. OVIC therefore welcomes future opportunities to comment on these matters and will follow progress on the Government's approach to AI ethics in Australia with interest.

I have no objection to this submission being published by DIIS without further reference to me. I also propose to publish a copy of this submission on the OVIC website, but would be happy to adjust the timing of this to allow DIIS to collate and publish submissions proactively.

If you have any questions about this submission, please contact Tricia Asibal at tricia.asibal@ovic.vic.gov.au.

Yours sincerely

Rachel Dixon
Privacy and Data Protection Deputy Commissioner

[28] Dimensional Research, Artificial Intelligence and Machine Learning: Projects Are Obstructed by Data Issues Global Survey of Data Scientists, AI Experts and Stakeholders, May 2019, https://content.alegion.com/dimensional-researchs-survey.
[29] Note the FOI Act only applies to Victorian government agencies. There are equivalent acts in other jurisdictions.
[30] On page 6 of the paper.
[31] Section 3 of the FOI Act.