Horizon Scanning Series

The Effective and Ethical Development of Artificial Intelligence: An Opportunity to Improve Our Wellbeing

Data Collection, Consent and Use

This input paper was prepared by Lyria Bennet Moses and Amanda Lo

Suggested Citation

Bennet Moses, L and Lo, A (2018). Data Collection, Consent and Use. Input paper for the Horizon Scanning Project "The Effective and Ethical Development of Artificial Intelligence: An Opportunity to Improve Our Wellbeing" on behalf of the Australian Council of Learned Academies, www.acola.org.

Introduction

This paper responds to questions raised by the Australian Council of Learned Academies (ACOLA) regarding data collection, consent and use in the deployment of artificial intelligence technologies.

The questions received include: How can consent be given for future uses? How to ensure informed consent for data collection, third parties, and use for additional, dual purposes (e.g. amalgamation with other data sets)? Whether specific regulatory frameworks are required? How should data sovereignty be handled by transnational companies? In what ways might governments recognise the value of data and ensure that benefits are passed onto the public while enabling innovation nationally and internationally?

The response is set out under sections 5.1.1 and 5.1.2 based on ACOLA's draft report structure.

5.1 Collection, consent and use

5.1.1 Including aggregation of data and subsequent use for purposes where there is no consent

Privacy laws in Australia

The principal legislation governing privacy and data protection in Australia is the federal *Privacy Act 1988* (Cth), which regulates the handling of personal information of individuals by the private sector and federal government agencies.[1] It contains 13 Australian Privacy Principles (APPs) based on the 1980 OECD Guidelines and the EU Directive.

The APPs collectively govern the collection, use, disclosure, storage, security, as well as access and correction of personal information, which is defined in the *Privacy Act 1988* (Cth) as "information or an opinion about an identified individual or an individual who is reasonably identifiable: (a) whether the information or opinion is true or not; and (b) whether the information or opinion is recorded in a material form or not."[2]

Impact of big data on current model for data protection

What is big data? Definitions focused on the technological aspects highlight the volume, velocity and variety of data.[3] It can be understood as a process whereby information about individuals are collected from different entities and aggregated over time to identify patterns and/or draw predictive inferences. Kitchin recognizes that big data also implies certain beliefs about ways in which inferences are drawn.[4] Boyd and Crawford expand the definition of big data and view it as a "cultural, technological and scholarly phenomenon".[5]

The first comprehensive review on the impact of big data in Australia was led by the Productivity Commission. They concluded the tremendous potential value of big data could be unleashed to benefit individuals, businesses and society through increased access and use of data.[6] The question is how to ensure the benefits of big data are harnessed without undermining the individual's right to privacy. The rest of this section explains how big data exposes weaknesses in the current approach to data protection.

---

[1] Public sectors of various states and territories are governed by separate legislations: Information Privacy Act 2014 (ACT), Privacy and Personal Information Protection Act 1998 (NSW), Information Act (NT), Information Privacy Act 2009 (Qld), Personal Information and Protection Act 2004 (Tas), and Privacy and Data Protection Act 2014 (Vic). South Australia issued administrative rules requiring compliance with a set of Information Privacy Principles, while in Western Australia, some privacy principles are included in the Freedom of Information Act 1992 (WA). See Office of the Australian Information Commissioner, "Other privacy jurisdictions" at https://www.oaic.gov.au/privacy-law/other-privacy-jurisdictions

[2] *Privacy Act 1988* (Cth), s6.

[3] Paul Zikopoulos and Chris Eaton, *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*, McGraw-Hill, 2011.

[4] Rob Kitchin, "Big Data, New Epistemologies and Paradigm Shifts", *Big Data and Society* Vol 1, 1-12.

[5] Danah Boyd and Kate Crawford, "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon", *Information, Communication and Society* Vol 15, 662-79.

[6] Productivity Commission, "Data Availability and Use", March 2017 at 4.

First, the current privacy framework emphasizes consent, or individual control over personal data. Under the system of "notice and consent",[7] the data subject is given notice, often in the form of a privacy policy, of the purpose of the data use at the time of data collection. However, the "notice and consent" model is problematic for a number of reasons.[8] It is well documented that users often do not read privacy policies.[9] The "transparency paradox" suggests users will not read more detailed policies, but oversimplified policies fail to explain privacy choices meaningfully.[10] Even with sufficient information, users do not act rationally, and are likely to trade off long-term privacy for short-term benefits.[11]

Second, the value of personal information is often unknown at the time of collection when consent is usually requested. This may make it hard for the data controller to specify upfront the types of purposes that the data may be used. Additionally, there could be new data controllers who use the data after collection depending on how the data is combined and processed. Future purposes, and use by new data controllers, are often unexpected and would require amended consent, which is likely a costly exercise.

Third, privacy laws only regulate personal data, which is generally defined as information that makes an individual identifiable. However, it is not easy to determine whether certain information is personal data because individuals can be re-identified from de-identified data when it is cross-matched with other data sets.[12] De-identification can be better understood as a risk management process rather than a single outcome. The problem with the *Privacy Act* and other state and territory privacy legislation, is that information is viewed in binary terms, meaning the data must either be personal or non-personal. But the extent to which information can identify an individual will differ for different data sets given different levels of risks (based on the probability and foreseeable impact of re-identification over time). Legislation in Australia differs in how it deals with this challenge – in some cases, context is relevant in classifying information as personal, whereas in others, the wording suggests whether individuals are identifiable from a dataset is an intrinsic property of that dataset.[13] While the contextual definition is more helpful in ensuring appropriate data governance, it does create a challenge in that the same data set may fall into the definition only at particular times or in particular hands.

Fourth, the current system places a heavy burden on individual users to self-manage their privacy in the face of numerous entities collecting their data.[14] Whereas data controllers are in a position to analyse risks, data subjects generally lack the information or expertise. This imbalance could lead data controllers to exploit the privacy risks to their advantage.

Fifth, weighing the costs of privacy protection and the benefits of big data innovation is not straightforward. The benefits and harms of privacy choices are distributive in a society.[15] One privacy choice could benefit some to the detriment of others. If the consent model places the responsibility on individuals, then individuals are likely to make privacy choices in isolation from others. This may inadvertently create the "tyranny of the minority",[16] where a small number of individuals who volunteer information make it possible for knowledge to be inferred about the majority who have withheld consent. An obvious example of this is the sharing of genetic information, but the same point applies to the sharing of information about social networks or community behaviours, for example.

---

[7] Omer Tene and Jules Polonetsky, "Big Data for All: Privacy and User Control in the Age of Analytics" *Northwestern Journal of Technology and Intellectual Property* Vol 11 Issue 5, 2013 at 260.
[8] Daniel J. Solove, "Privacy Self-Management and the Consent Dilemma" *Harvard Law Review* Vol 126, 2013; Solon Barocas and Helen Nissenbaum, "Big Data's End Run around Anonymity and Consent" in Julia Lane et al. (Eds.), *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, New York: Cambridge University Press, 2014.
[9] Fred H. Cate and Viktor Mayer-Schönberger, "Notice and consent in a world of Big Data" *International Data Privacy Law* Vol 3 Issue 2, 2013 at 67; Helen Nissenbaum, *Privacy in Context: technology, policy, and the integrity of social life*, Stanford: Stanford Law Books, 2010 at 105.
[10] Helen Nissenbaum, "A Contextual Approach to Privacy Online" *Daedalus* Vol 140 Issue 4, 2011 at 36.
[11] Alessandro Acquisti and Jens Grossklags, "Privacy and Rationality in Individual Decision Making" *IEEE Security and Privacy Magazine* Vol 3 Issue 1, 2005.
[12] Australian Computer Society, "Data Sharing Frameworks: Technical White Paper", September 2017.
[13] Unlike legislation in the ACT, NT or (after 2012) the Commonwealth, the definition of personal information in Queensland, Victoria, and NSW states a person must be identifiable 'from the information'. It is possible that these words mean information does not become personal information merely because there is potential for linking with other information. When a similar wording used to exist in the Commonwealth Act, former OAIC guidance suggested such strict interpretation was inappropriate. The former guidance is no longer accessible.
[14] Solove, 1888 and Cate, 68.
[15] Lior Jacob Strahilevitz, "Toward a Positive Theory of Privacy Law" *Harvard Law Review* Vol 126, 2013.
[16] Barocas and Nissenbaum, 61.

Ideas on data governance in an age of big data

Data anonymization

Anonymization is the "process by which information in a database is manipulated to make it difficult to identify data subjects".[17] The most common approaches are k-anonymity, l-diversity and differential privacy.[18] This technique is often used by data controllers to anonymize data before they release personal data to protect the privacy of data subjects. While it is widely used, Ohm has criticised privacy legal scholarship for its "faith" in anonymization.[19] This is because anonymized data can be easily re-identified by linking anonymized information to outside information, thereby revealing the identity of data subjects.[20]

Reforming the consent model

In a summary of views from government, academia, advocacy and industry groups, Cate and Mayer-Schönberger highlighted several themes of reforms to the current legislative framework.[21]

First, they point out the burden of privacy management should be shifted away from data subjects to data controllers, and they consider the emphasis should be placed less on individual notice and more on disclosure to a regulator or a central repository. In addition, they suggest data controllers could demonstrate accountability through "responsible data stewardship", which means doing more than meeting the basic privacy standards for compliance.

Second, their article points out the attention should be around the "use" instead of the "collection" of personal data. The Australian Computer Society (ACS) also suggests the focus should not be on the data itself, but the impact from the use of data. The ACS suggests a move away from examining who "owns" the data, but rather the "rights, roles, responsibilities, and limitations for those who access data in the various processes from collection, use, sharing and storage."[22]

Data licences

At the moment, privacy policies often contain highly individualised and specific terms on how data is collected, processed and used. One idea worth exploring is to communicate terms of data use through licences instead of text. In the copyright domain, there are six standardised Creative Commons (CC) licences, which reflect different combinations of lawful uses and conditions.[23] A similar licensing framework could potentially be applied to personal data, in which a limited number of licence types specify the different terms of data usage, for example:

- Use limited to entity to whom data is provided and purposes closely aligned with purpose of collection. Data deleted when no longer required for that purpose.
- Use limited to entity to whom data is provided, but purposes can be related to primary purpose. Data deleted when no longer required for that purpose.
- Use limited to entity to whom data is provided but uses can be unrelated to primary purpose of collection. Data deleted in accordance with ordinary company policy.
- Data can be shared with related entities and use can be unrelated to primary purpose of collection. Data will not necessarily be deleted after any particular fixed period.
- Data can be broadly shared and used provided it is de-identified under the data controller's data risk governance framework (or some general standard).

---

[17] Paul Ohm, "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization" *UCLA Law Review* Vol 57, 2010 at 1701.
[18] Above 12.
[19] Above 17 at 1704.
[20] Above 17 at 1708.
[21] Above 9.
[22] Above 12.
[23] Creative Commons Australia, "About the Licences", https://creativecommons.org.au/learn/licences/

These could each be associated with more detailed "standard conditions" privacy policies. The advantage of standardisation is that computers could be programmed to communicate directly with each other, without human intervention, automatically negotiating terms of data management between a data subject and a data controller based on relatively simple settings.

Privacy certification and rating

Dr. Alan Finkel AO, Chief Scientist of Australia, proposed the creation of a recognised mark for ethical technology vendors, which would be named the "Turing certificate". He envisions this as a voluntary system suitable only for low-risk consumer technology, such as smartphone applications and digital home assistants.[24] Finkel describes a system that would be informed by standards developed by experts in consultation with consumer and industry groups. The products, procedures and processes of the company would be reviewed by an independent auditor. This paper suggests privacy standards could be embedded into part of the ethical certification process that Finkel proposed.

Another method that may help consumers make better purchasing decisions is privacy labelling. Currently, Energy Star ratings assess the energy efficiency of electronic appliances. A similar system could be used to demonstrate whether a technology application is privacy-friendly. It is possible that the visualization of privacy risks could make privacy choices more accessible to the average consumer and potentially increase the transparency and disclosure of privacy risks by data controllers.[25] Labelling may even encourage competition between data controllers to provide more privacy-friendly solutions, including computer-to-computer negotiations over data management terms.

Legislative reform

The Data Sharing and Release Bill (DSR), which is being drafted based on recommendations from the Productivity Commission, aims to create a new data governance framework that enables researchers to harness the value of government data.[26] Below is a summary of recommendations that were included in the Allens Hub submission to the Issues Paper:[27]

- Rationalise current patchwork of laws about how government shares information internally and externally and clarify the Bill's relationship with existing data protection laws
- Broaden the reform to clarify definitions and concepts in data sharing and release
- Acknowledge further diversities related to data, such as quality, context, community perspectives and amenability of data to reuse
- Ensure government decisions are based on principles of fairness and justice given the risks of uneven data availability and misleading policy objectives
- Develop a data ethics framework and accountability mechanisms and increase education and training to promote a culture supportive of responsible data use

[24] Presentation at the Human Rights and Technology Conference organised by the Australian Human Rights Commission on 24 July 2018 in Sydney, Australia. https://tech.humanrights.gov.au/conference
[25] Inspired by food nutrition labels, Cranor has proposed the design of standardised and simplified privacy nutrition labels to replace privacy policies for consumers. See Lorrie Faith Cranor, "Necessary but not Sufficient: Standardized Mechanisms for Privacy Notice and Choice", *Journal on Telecommunications and High Technology Law* Vol 10 Issue 2, 2012, 273-308.
[26] Department of the Prime Minister and Cabinet, "New Australian Government Data Sharing and Release Legislation: Issues paper for consultation", 4 July 2018. https://www.pmc.gov.au/resource-centre/public-data/issues-paper-data-sharing-release-legislation
[27] The Allens Hub for Technology, Law & Innovation, "Response to Issues Paper on Data Sharing and Release" available on request.

5.1.2 Data sovereignty[28]

Overview of data localisation laws

Data localisation legislation requires network providers to store original or copies of collected data about internet users in the country on servers located within the jurisdiction. These laws have been justified to ensure the privacy and security of citizen's data, provide better information security against foreign intelligence agencies, and support domestic law enforcement activities.[29]

Data localisation measures vary in scope.[30] A handful of countries such as China, Russia, and Indonesia have enacted broad data localisation laws. Most countries have narrow data localisation laws, imposing the restriction only on certain types of personal information and specific industry sectors.[31] For example, Australian laws are narrow in scope and require electronic health records to be stored locally.[32] The transfer, processing or handling of such data outside of Australia is permitted only if such records do not include "personal information in relation to a consumer" or 'identifying information of an individual or entity."[33]

Impact on multinational companies

Technical aspects

Data localisation laws are likely to create technical difficulties for multinational companies seeking to generate business insights from data collected from multiple jurisdictions. Many companies store data in the cloud, which has been noted for its lack of transparency, making it difficult for companies to see where the data is stored and processed.[34] However, to comply with the laws, companies need to know precisely what type of data is stored in what location.

It seems likely that these laws could hinder the application of machine learning to collected data. If most companies conduct data analysis from a centralised database, this would require the free flow of data across national borders. In these scenarios, distributed machine learning[35] could be a potential solution. It helps companies apply machine learning to data collected in a "distributed" manner, for example from different geographies, without first communicating it all to a central location. This, however, potentially slower and more difficult.

Legal compliance

Countries with broad data localisation laws in effect create new privacy standards for data collected within their jurisdiction. This means multinational companies could have the additional burden of complying with privacy standards unique to each country on top of international and regional privacy legislative frameworks. For example, China's Cybersecurity Law (CSL) introduces restrictions on cross-border data transfers that differ from international privacy regimes such as the European Union's General Data Protection Regulation (GDPR) and the voluntary Asia-Pacific Economic Cooperation Cross-Border Privacy Rules (CBPR).[36]

---

[28] The term "data sovereignty" is often used in the context of indigenous data sovereignty, which refers to the "right of indigenous peoples and nations to govern the collection, ownership, and application of their own data." See slide 16, "The Governance of Indigenous Data: Generating a Framework and Principles", US Indigenous Data Sovereignty Network. http://www.ncai.org//3._2018_IDGov_Drafting_Princples_and_Framework_-_NCAI_2018_16x9_FINAL.pdf

[29] John Selby, "Data localization laws: trade barriers or legitimate responses to cybersecurity risks, or both?" *International Journal of Law and Information Technology* Vol 25 Issue 3, 2017, 213–232.

[30] Anupam Chander and Uyên P. Lê, "Data Nationalism" *Emory Law Journal* Vol 64 Issue 3, 2015, 677-739.

[31] For example, in Europe, different governments require different types of data to be stored locally. These range from financial records, gambling winnings and user transactions, and government records as discussed by Selby. Other countries impose restrictions to data collected from specific sectors, such as financial, health and medical information, online publishing, and telecommunications data. See Bret Cohen et al., "Data Localization Laws and their Impact on Privacy, Data Security and the Global Economy", *Antitrust* Vol 32 Issue 1, Fall 2017, 107.

[32] *Personally Controlled Electronic Health Records Act 2012* (Cth), s77.

[33] Ibid., s77(2).

[34] Ruslan Synytsky, "GDPR and Data Localization: The Significant (and Often Unforeseen) Impact on the Cloud", *SC Media US*, December 2017.

[35] Travis Dick et al., "Data Driven Resource Allocation for Distributed Learning", Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), USA, 2017.

[36] Samm Sacks, Paul Triolo and Graham Webster, "Beyond the Worst-Case Assumptions on China's Cybersecurity Law", *New America*, October 2017.

Under the CSL, "network operators" and operators of "critical information infrastructures" are required to store "personal information" and other "important data" that is collected and generated in China within the jurisdiction. Such data can be stored or provided overseas for business reasons only if it is "truly necessary" and the operators conduct a self-security assessment or pass an official security assessment when a threshold test is met.[37] The security assessment is based on a two-pronged test.[38] First, whether the transfer is "lawful, legitimate and necessary". Second, the risk of transfer is evaluated by looking at the nature of the data and the likelihood and impact of security breaches involving such data. While the GDPR and CSL appear to have similar cross-border transfer tests, there are material differences.[39] CSL does not provide for derogations that are found in the GDPR. Neither does the CSL contain mechanisms in the GDPR such as Binding Corporate Rules[40] and standard data protection clauses for companies to gain approval.[41]

Lastly, data localisation laws are likely to increase compliance costs since companies engaged in data collection from different countries will have to build local data centres in each jurisdiction.

This is not to say that data localisation laws may not be rational for individual countries seeking to protect citizen data and ensure local access (for example, by intelligence and law enforcement agencies). However, an international framework with consistent data protections and clear rules for transnational access would resolve some of the issues identified above. Whether this is ultimately feasible or desirable is a question beyond the scope of this paper.

Authors: Amanda Lo, Lyria Bennett Moses

[37] Reed Smith, IP, Tech & Data White Paper "China's Cybersecurity Law", 2018.
[38] Ibid.
[39] Xiaoyan Zhang, "Cross-Border Data Transfers: CSL vs. GDPR", *The Recorder*, January 2018.
[40] Binding Corporate Rules allow multinational companies to transfer personal data out of the European Union within the same corporate group to countries that do not have an adequate level of data protection.
[41] Standard contractual clauses are used to transfer data outside the European Union and are deemed to provide sufficient data protection by the European Commission.