Discussion Paper: International Data Ethics Frameworks

Purpose and scope

1. This paper has been prepared on behalf of the Government Chief Data Steward for the Data Ethics Advisory Group (DEAG). It discusses the current landscape of data ethics frameworks, which includes artificial intelligence, and explores common themes which are particularly relevant to the function of the DEAG. The focus is on the benefits and considerations of frameworks. This paper is not official New Zealand government policy.

2. This paper is intended to support the general discussion of the benefits and considerations of data ethics frameworks for the DEAG. The issues canvassed should not be considered reflective of the position of any specific government agency (including Stats NZ).

Executive summary

3. Data ethics frameworks have proliferated from a variety of stakeholders, primarily in Europe and North America. They tend to focus on artificial intelligence and provide high-level principles or 'deontological ethics' which guide professional cultures and narratives in non-binding ways. They are considered more flexible and can be applied more rapidly than laws or professional codes of conduct. Some frameworks include self-assessments or certifications for compliance.
    a. While generally non-binding, some frameworks can be mandated under specific conditions such as during the of procurement for AI systems (e.g. UK Government Draft Guidelines for AI Procurement).
    b. In most cases frameworks are framed for 'ethics-by-design' and they target new projects during the procurement, development or deployment stages. There is limited discussion on a retro-active application of frameworks.

4. Four common themes of 'privacy', 'transparency', 'bias and discrimination' and 'accountability' emerge from analysis of these frameworks, however, divergences tended to arise in relation to the actors which have formulated them, how they are framed and how they are interpreted. The common themes tend to overlap with areas where technical fixes can be or have already been developed.
    a. Recent frameworks[1] expand on the four common themes with aims for social benefit and they tend to incorporate aspects or refences to human rights. Key themes include the balance of beneficence and non-maleficence ('proportionality'), 'autonomy' or 'self-determination' and 'justice'.
    b. It should be noted that this exploratory work did not find prominent indigenous themes per se for the majority of data ethics frameworks, however, the CARE Principles indicate three key indigenous ethical components: 'proportionality', 'justice' and 'future use'.

5. While there is some consensus at a high level, this has not been achieved at a detailed level. In particular, there are unresolved issues around how these principles should be interpreted, why they should be deemed important, what issue, domain or actors they should pertain to, and how they should be implemented[2].

6. There have been calls to consolidate some existing ethical frameworks, particularly where frameworks are frustrating attempts to achieve industry-level compliance and streamlined processes.[3] In this sense, frameworks should be designed to complement professional codes of conduct or standards (e.g. IEEE) which provide practical guidance for data practitioners and laws which provide procedures to manage and remediate non-compliance.

---

[1] E.g.: Cowls, J. and Floridi, L., 2018. An ethical framework for a good AI society, URL.

[2] Jobin, A., Ienca, M. and Vayena, E., 2019. The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), pp.389-399, URL.; Daly, A., Hagendorff, T., Hui, L., Mann, M., Marda, V., Wagner, B., Wang, W., Witteborn, S., 2019, Artificial Intelligence Governance and Ethics: Global Perspectives, URL.

[3] See page 56: Australian Human Rights Commission, 2019. Human Rights and Technology Discussion Paper, URL.; and Mittelstadt, B., 2019. Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, pp.1-7, https://www.nature.com/articles/s42256-019-0114-4.

There has been a proliferation of international principles (>80 AI ethics principles)

7. There is a high representation (~60%) of frameworks, particularly relating to artificial intelligence, developed in Europe and North America, and over 80% were released since 2016. [1] Private and public sectors have contributed equally (~20% each) while academic, intergovernmental and NGO organisations have contributed to a lesser extent (~10% each). [1]

8. These frameworks can take different forms and levels of engagement. For example, all frameworks include key principles to adhere to, however, some focus on the swearing of an oath to these principles (e.g. Danish Data Ethics Oath), some use a self-assessment tool to comply with principles (e.g. Canadian Algorithmic Impact Assessment) and some include a public facing certification if compliance is met (e.g. Danish Data Ethics Seal).

9. Between 2019 and 2020, several groups have made efforts to identify leading principles which have been included in **Table 1**. The four common themes which are highly represented are discussed in the sections below.

10. The highly represented themes mostly overlapped with themes raised during interviews with DEAG members. However, 'data sovereignty' ('ownership') was one theme raised by DEAG members which were not prominent in ethical frameworks.
    a. One report[1] did briefly note the concept of ownership: while it is considered a solution to data governance, ownership was not easily applicable and does not address issues around data use and impact of outcomes from automated decision making.

Common theme: privacy

11. *Privacy*: is not always defined but generally covers data security and data usage with links to human freedoms such as autonomy. Invasions on human freedoms and privacy tend to be related to AI implementations such as profiling, decision making and surveillance technology.
    a. Privacy is protected in most countries, excluding the US and others, by conventional national laws and is a key theme for personal data protection under the General Data Protection Regulation of the European Union (GDPR).

12. Technical solutions such as privacy by design, data minimization and access control are included in framework recommendations to data practitioners and designers.

13. There are calls for data-subject-focused approaches such as strong consent process, control over the use of data, ability to restrict data processing, right to correction and the right to be forgotten, most of which are covered by the provisions of the GDPR.[2]

CASE STUDY: AI Ethics Principles, Department of Industry, Science, Energy and Resources, Australian Government

In April 2019, the Minister for Industry, Science and Technology released a discussion paper to encourage conversations on "…*how we should design, develop, deploy and operate AI in Australia*". They received 130 submissions from government, business, academia, non-government organisations and individuals and used this feedback to revise and publish the AI Ethics Principles.

Principle "privacy protection and security":

"*Throughout their lifecycle, AI systems should respect and uphold privacy rights and data protection, and ensure the security of data."*

The explanation for this principle suggests that authors of AI systems should ensure 'proper data governance', and management, for all data used and generated by the AI system throughout its

[1] Discussed briefly on page 30: Digital Future Society, 2019. Toward better data governance for all: Data ethics and privacy in the digital era, URL.

[2] See page 21: Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., Srikumar, M., 2020. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI." Berkman Klein Center for Internet & Society, URL.

lifecycle. This includes minimising data which could be used to identify individuals (i.e. anonymisation) and ongoing assessments of 'connections' between data and inferences which are produced by the AI systems. This principle also encourages data and AI security measures are taken to address potential security vulnerabilities, and assurance of resilience to adversarial attacks. Unintended use or potential abuse risks should also be identified and accounted for.

The Australian AI Ethics Principles is a good example of the high-level or 'deontological' approach which is generally taken for ethical frameworks. While the principles go some way to identify the prominent issues, there is little to no guidance to how data-practitioners can action these concepts. This is especially apparent in the description of the privacy principles above, and no reference is made to other more-established sources or laws where there is a lack of clarity. The feedback from consultation stated that the privacy principle "needs improved clarity on how it interacts with current statutes and common law principles" which may be addressed in ongoing revisions of this and other principles.

Common theme: transparency

14. *Transparency*: covers explainability, interpretability or other acts of disclosure and relate to how data is used, and how automated decision are made. Transparency is dependent on how effectively processes, risk and limitations can be understood by end-users and those affected (i.e. data literacy).
    a. The need for transparency is highly dependent on the context and the severity of the consequences if that output is erroneous or otherwise inaccurate. For example, the AI Now 2017 Report includes criminal justice, healthcare, welfare and education sectors within the definition of 'high-stake' domains and the report recommends that core public agencies in these areas no longer use "black box" or poorly transparent AI and algorithmic systems. This recommendation is included in the Toronto Declaration.

15. Some frameworks (e.g. UK Government Data Ethics Framework) encourage that AI and algorithms are developed in the simplest way possible with robust documentation. This is because more sophisticated systems make it difficult to explain how decision making was reached in an understandable way – which is referred to as the 'transparency fallacy'.[1]

16. It is noted that transparency about limitations of data and AI are included in prominent frameworks (e.g. ODI Data Ethics Canvas), however this wasn't clearly identified in the high-level principles of the analyses. These types of frameworks encourage discourse about the limitations of data insights and decision-making processes, particularly for end-users and those affected. This discourse is associated to benefits of public trust and is important to establish if data or AI quality/accuracy is 'fit-for-purpose'.

CASE STUDY: Data Ethics Framework, Department for Digital, Culture, Media & Sport, UK Government

Initially released in 2016, updated in 2018 and undergoing third update currently, the Data Ethics Framework is relatively mature compared with other frameworks. It builds on the core values of the UK Civil Service Code of integrity, honesty, objectivity and impartiality. The framework is accompanied by additional guidance for how to enact the principles of the framework and a Workbook to help record the ethical considerations which were made for compliance and refreshing processes. While this framework is high-level there has been integration with other frameworks to encourage an ethical approach in specific areas such as the Draft Guidelines for AI procurement.

Principle 6 of the Data Ethics Framework addresses transparency and accountability in unison:

> "*You should be transparent about the tools, data and algorithms you used to conduct your work, working in the open where possible. This allows other researchers to scrutinise your findings and citizens to understand the new types of work we are doing*."

[1] Discussed briefly on page 28: Digital Future Society, 2019. Toward better data governance for all: Data ethics and privacy in the digital era, URL.

The concept of transparency and openness of data, and data usage, arises to encourage an improved used of data across government, unless there are reasons for secrecy such as fraud or counter-terrorism. Building trust and enabling peer-review processes are key reasons given to encourage transparency. Feedback tells you what people care about and feel is comfortable. The framework also indicates a need to plan for how you will explain your work to others, ensuring your approach can be held to account. It is considered essential that data systems which guide government policy are based on interpretable evidence in order to provide accountability of policy outcomes.

The guidance component gives practical advice for peer-review such as getting feedback from the internal data science functions, or externally through the Government Digital Service Data Science Community or GitHub for automated testing. It is suggested that if data is non-sensitive and non-personal it should be open and accessible with a digital object identifier. When sharing personal data, you must comply with the Information Commissioners Office Data Sharing Code of Practice, which has been open for consultation and will be updated to be compliant with the Data Protection Act 2018. For artificial intelligence systems, it is recommended that the methods used for training are most relevant for open release, and release of the system itself would also be useful for peer-review and monitoring the evolution of the system through time. However, neither of these should be released if there is a risk to endanger the privacy of those whose data was used to train it or integrity of the task being undertaken. In this case it may still be useful to release metadata about the model on a continual basis, like its performance on certain datasets. In cases where the project is very sensitive, you could arrange for selected external bodies (another government department, academia or public body), approved by your organisation, to examine the model itself in a controlled context to provide feedback.

Principle 4 also is relevant to transparency as it addresses understanding and being clear about the limitations of data:

> "*Data used to inform policy and service design in government must be well understood. It is essential to consider the limitations of data when assessing if it is appropriate to use it for a user need*."

This concept asks users to explore the occurrence and impact of errors, including errors in meta-data, and bias (social bias with algorithms also covered in Principle 5 – Use robust practices and work within your skillset). It also covers the provenance of data and understanding the appropriateness both technically (accuracy, reliability and representativeness) and ethically (e.g. is there a match in the use case between the original and new use).

The guidance component suggests the use of the UK Statistics Authority Quality Assurance of Administrative Data framework to help you understand the data that you are using, how it was collected and any likely quality impacts.

Common theme: bias and discrimination

17. *Bias and discrimination*: covers fairness, monitoring and mitigation of unwanted bias and discrimination in data and AI. These are shaped by historic and current social norms and the impacts of bias and discrimination is dependent on the purpose of technology and how suitable the underlying data is to that purpose.

18. Generally, it is encouraged that the benefits, harms and risks of a given technology should be measurable and proportional across affected demographics. This is a prominent aspect of recent frameworks and tends to be described the themes of beneficence and non-maleficence.

19. Most frameworks make technical recommendations such as setting standards, clear documentation, and auditing throughout the lifecycle of technology. However, these tend to not include what should be documented and how it should be monitored.

20. Oversight and diversity of views is encouraged internally (within management and data practitioner teams), and externally through civil discourse and meaningful interaction with other relevant stakeholders.

a. Advisory processes are encouraged however these generally focus on the development or deployment phases of a project rather than in post-development phases. For example, the Ethical Framework for a Good AI Society states: "*… with the development of a mandatory form of "corporate ethical review board" to be adopted by organisations developing or using AI systems, to evaluate initial projects and their deployment with respect to fundamental principles*."

Common theme: accountability

21. *Accountability*: is not clearly defined by frameworks and generally covers acting with integrity and within legal constraints. The actors which are deemed responsible is project-specific and vary between frameworks, if they are covered at all.

22. Traditional models of accountability tend to fail for automated decision-making systems, and there are diverging schools of thought about whether AI should be held accountable in a human-like manner or whether humans should always be the only actors who are ultimately responsible.[1]

23. Some frameworks emphasise the need for safe processes of whistleblowing when unethical behaviours emerge, especially in cases where there is a high risk of harm.[2]

Considerations for data ethics frameworks:

The uptakes and usages of data ethics frameworks are not well understood

24. Considering a study reporting that the reading of the Association for Computing Machinery's Code of Ethics had little effect on data practitioner's intentions for development,[3] evidence of uptake and usage is needed, however, methods to measure these are not well established.
    a. *Uptake:* can be defined as either an individual or groups 'awareness', 'conscious use' or 'adherence to' a given framework and this could be automated through a website
    b. *Usage:* is more difficult to define and relates to how the framework is interpreted and what the impacts on behaviours are, which requires behavioural research
    c. The UK Government Data Ethics Framework staff are undergoing a manual survey of users to explore both issues and early indications suggest that data ethics practitioners found practical guidance of the 2016 iteration particularly useful. They also found that statisticians and other professional practitioners (such as NHS health experts) often defaulted to their specific codes of conduct rather than using the framework.

Similarities have formed between data ethics and medical ethics

25. According to recent analyses,[4] artificial intelligence ethics has converged on principles that closely represent medical ethics. However, there are significant differences between the fields of medical practice and artificial intelligence. For example, artificial intelligence does not have:
    a. Common aims and fiduciary duties
    b. Professional history and norms
    c. Proven methods to translate principles into practice
    d. Robust legal and professional accountability mechanisms.

---

[1] Jobin, A., Ienca, M. and Vayena, E., 2019. The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), pp.389-399, URL.
[2] Hagendorff, T., 2019. The ethics of AI ethics-an evaluation of guidelines. arXiv preprint arXiv:1903.03425, URL.
[3] McNamara, Andrew, Justin Smith, and Emerson Murphy-Hill. 2018. "Does ACM's Code of Ethics Change Ethical Decision Making in Software Development?" In Proceedings of the 2018 26th ACM Joint Meeting ESEC/FSE 2018, 1–7. New York, ACM Press, URL.
[4] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy and M. Srikumar, 2020, "Principled AI: Mapping Consensus in Ethical And Rights-based Approaches to Principles for AI," Berkman Klein Center for Internet & Society, Cambridge, Massachusetts, URL; and Mittelstadt, B., 2019. Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, pp.1-7, URL.

Differences between frameworks arise from the actors involved, framing and interpretations

26. Jobin *et al.*[1] suggests that these differences indicate that there is uncertainty about how the principles of frameworks should be prioritised and how conflicts between ethical principles should be resolved. An example of conflict between principles can be see between rectifying *bias*, i.e. collecting larger and more diverse data sets, and individual *control* and *privacy* of data which may be impinged when covering a wider base of individuals with different opinions.

Data ethics frameworks do not provide a 'catch-all' for ethical practices

27. Frameworks tend to provide high-level principles which guide professional cultures and narratives in non-binding ways. They are considered more flexible and can be applied more rapidly than laws or professional codes of conduct. Because of these factors, they tend to have limited governance over:
    a. Practical guidance for data practitioners
    b. Enforcement of principles and guidance with robust accountabilities.

28. Overlapping ethical frameworks can frustrate attempts to achieve industry-level compliance and streamlined processes regardless of agendas.[2] There have been calls (e.g. from the Australian Human Rights Commission) for consolidating some existing ethical frameworks, perhaps either at an international or industry level.
    a. In the vacuum of established professional codes and laws, the principles and frameworks are the easiest to implement. They set a 'precedent' for the more binding aspects to enforce as they are implemented. It is important that ethical considerations are not used as a substitute for regulation – which is referred to as 'ethics-washing'[3].

Indigenous themes are not prominent in data ethics frameworks

29. None of the prominent literature analyses that are included in this paper contained references to indigenous themes per se, and references of "culture" were generally limited to professional cultures.
    a. This could be explained by the observation that frameworks appear to be driven by consensus (e.g. consensus and improvement in the areas of the common themes), and this may result in underrepresentation for themes that involve diverse, and potentially conflicting views.
    b. This absence of indigenous themes could also be explained by the high-level approach of frameworks which generally do not include details that are specific any one demographic, but rather attempt to provide wide coverage of populations.

30. In support of this finding, the IEEE Ethically Aligned Design for automated and intelligent systems has identified a "Western" monopoly on ethical traditions. This practitioner code describes that there is a need to broaden traditional ethics from a contemporary "Western" ethical foundation to include other traditions of ethics, and suggests the inclusion of concepts inherent to Buddhism, Confucianism, and Ubuntu traditions. The code recommends an acknowledgement of where there may be differences in ethical approaches and efforts to find intercultural commonalities of what makes up responsible innovation practice.

31. While diverse viewpoints may not be reflected in the high-level framework principles, it should be noted that 'diversity' in the data practitioner workforce was a theme present in many frameworks. In addition, some frameworks did include mentions of cultural aspects such as the Montreal Declaration which states: "…*AIS [AI systems] must be compatible with maintaining social and cultural diversity and must not restrict the scope of lifestyle choices and personal experience*".

---

[1] Jobin, A., Ienca, M. and Vayena, E., 2019. The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), pp.389-399, URL.

[2] Page 56: Australian Human Rights Commission, 2019. Human Rights and Technology Discussion Paper, URL.

[3] 'Ethics washing' is defined as when ethics is used as a substitute for regulation, see Wagner, B. (2018). Ethics as an Escape from Regulation: From ethics-washing to ethics-shopping? In M. Hildebrandt (Ed.), Being Profiling. Cogitas ergo sum. Amsterdam University Press, URL.

There are unresolved tensions which arise from framework ambiguity and different values

32. Gardam 2019[1] notes that there are specific areas where different ethical values result in tensions and inconsistencies of how ethical frameworks can be applied. Uncovering and resolving the ambiguity in ethical frameworks is needed before actions and interpretations can be made more consistent. Some tensions include:
    a. *Accuracy vs. fairness*: using algorithms to make decisions and predictions more accurate versus ensuring fair and equal treatment
    b. *Personalisation vs. solidarity*: reaping the benefits of increased personalisation in the digital sphere versus enhancing solidarity and citizenship
    c. *Efficiency vs. privacy*: using data to improve the quality and efficiency of services versus respecting the privacy and informational autonomy of individuals.

There are preconceived narratives which exist in the area of data and AI ethics

33. Greene et al. 2019[2] discusses that there are narratives which occupy the technological landscape which may influence the development of ethical frameworks. These narratives may or may not be beneficial or sustainable long-term and alternative narratives may need to be developed. For example, some narratives include:
    a. *Determinism*: it is assumed that the technologies: a) are coming and b) will replace a broad range of human jobs and decisions. This has limited the ethical debate to 'appropriate' design and implementation, as the current advance of technology is perceived to be "*unstoppable and irresistible*".
    b. *Technology as the focus of ethical scrutiny*: the ethical frameworks target technology (particularly 'high-risk' technology such as facial recognition or profiling), however, the wider ethical issues about commercial control and business ethics can be marginalised.

[1] Gardam, T., 2019. Data science and the case for ethical responsibility. Ada Lovelace Institute, URL.
[2] Greene, D., Hoffmann, A.L. and Stark, L., 2019, January. Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. In Proceedings of the 52nd Hawaii International Conference on System Sciences, URL.

Appendix 1

Table 1 *Simplified principles from literature analysis. The sector involvement has been included with either AI for Artificial Intelligence or DE for Data Ethics foci.*

Simplified Principles from Analyses					
Academic - AI [1] (ranked by importance)	Academic - AI [2] (ranked by inclusion)	Academic - AI [3] (ranked by inclusion)	Multisector – DE [4] (non-ranked)	Gov/business – DE [5] (non-ranked)	Business – DE [6] (non-ranked)
Transparency	Privacy	Privacy	Privacy by design	Privacy	Respect for persons behind the data
Justice and fairness	Accountability	Accountability	Open source by default	Transparency	Downstream uses of data
Non-maleficence	Transparency and Explainability	Fairness, non-discrimination, justice	New data governance models	Bias and discrimination	Data provenance
Responsibility	Fairness and Non-discrimination	Safety, cybersecurity	Accountability for unethical data use	Governance and accountability	Privacy and security safeguards
Privacy	Human control	Common good, sustainability, well-being	Favourable conditions for private sector shift		Follow and exceed legal obligations
Beneficence	Professional responsibility	Human oversight, control, auditing	Data literacy and education		Data minimisation
Freedom and autonomy	Human values	Solidarity, inclusion, social cohesion	Diverse and interdisciplinary AI workforce		Equal benefits and impacts
Trust		Science-policy link			Explicability of methods
Sustainability		Legislative framework, legal status of AI systems			Accurate representation of literacy
Dignity		Responsible/intensified research funding			Design for: • transparency,
Solidarity		Public awareness, education about AI and it's risks			• configurability
		Future of employment			• accountability
		Dual-use problem, military, AI arms race			• audibility
		Field-specific deliberations (health, military, mobility etc.)			Internal and external ethical review
		Human autonomy			Robust practices which are reviewed regularly
		Diversity in the field of AI			
		Certification for AI products			
		Cultural differences in the ethically aligned design of AI systems			
		Protection of whistle-blowers			
		Hidden costs (labelling, clickwork, content moderation, energy, resources)			

Appendix References

[1] A. Jobin, M. Ienca and E. Vayena, “The global landscape of AI ethics guidelines,” *Nature Machine Intelligence,* vol. 1, pp. 389-399, 2019.

[2] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy and M. Srikumar, “Principled AI: Mapping Consensus in Ethical And Rights-based Approaches to Principles for AI,” Berkman Klein Center for Internet & Society, Cambridge, Massachusetts, 2020.

[3] T. Hagendorff, “The Ethics of AI Ethics An Evaluation of Guidelines,” *arXiv preprint arXiv,* p. 1903.03425, 2019.

[4] Data Future Society, “Toward better data governance for all: Data ethics and privacy in the digital era,” 2019.

[5] Deloitte, “Government Trends 2020 – The rise of data and AI ethics,” 2019.

[6] Accenture, “Universal Principles of Data Ethics: 12 guidelines for developing ethics codes,” 2019.