Response to the "Data Strategy Consultation Paper" of the COAG Energy Council's Energy Security Board

Associate Professor Ben Rubinstein[1]

November 27, 2020

The ESB Data Strategy Consultation Paper presents public interest use of energy data across Core Bodies, motivating a review of legislation, policy, and governance for sharing or releasing energy data. This response focuses on privacy implications of sharing or releasing energy data and refers to the Data Strategy Consultation Paper (DS) and the ESB Data Strategy Preliminary Legal Report (LR).

1. The Data Sharing Principles (aka. Five Safes)

The Data Strategy calls for adoption of the Data Sharing Principles[2] (the Office of the National Data Commissioner's naming of the Five Safes framework) as a governance framework for management privacy protection. The Principles suffer from a range of serious flaws, including:[3]

1. No preference given to protections providing verifiable security properties or defence against identified threat models.
2. Encouragement of a risky *perimeter defence* mentality instead of *defence in depth.*
3. Privacy by superficial compliance over careful analysis.

While the framework acknowledges that sources of risk come from people, projects, settings, data and outputs of data sharing or release, nothing in the framework recommends one protection over another. When releasing data, for instance, the Principles do not favour best practice differential privacy - a framework with a verifiable security property protecting against a concrete threat model (see Section 2) over merely stripping identifiers from detailed micro data. The former limits re-identification while the latter is highly susceptible to re-identification.

Five Safes is advertised as providing "Principles [that] are much more flexible in nature and each principle can be applied on a sliding scale based on risk" (p. LR-38). This interpretation of Five Safes is not unique to the Data Strategy. It is inappropriate to configure risk management along the dimensions of Five Safes *independently* as these dimensions interact. Access to trusted data users such as Core Bodies within an

---

[1] BR is with the School of Computing & Information Systems, University of Melbourne. Trained in machine learning and statistics at UC Berkeley, Rubinstein's relevant expertise is in data privacy. He has pioneered research in the theoretical foundations of privacy, e.g. differential privacy, and has contracted in the Australian government (ABS, OVIC, Transport NSW) and private (big four) sectors on the practice of robust privacy enabling careful data release. With Culnane and Teague, he demonstrated reidentification of patients and doctors in the 2016 Medicare/PBS 10% data release and of public transport users in the 2018 Myki card release. His submissions and research have been cited in the Senate Inquiry into the *Reidentification Offence Bill* and the Australian COVIDSafe App's PIA.

[2] Office of the National Data Commissioner, *Data Sharing Principles*, March 2019. See https://www.pmc.gov.au/resource-centre/public-data/data-sharing-principles Accessed: 5 Nov 2020.

[3] Chris Culnane, Benjamin I. P. Rubinstein, and David Watts. *Not fit for Purpose: A critical analysis of the 'Five Safes'*. arXiv:2011.02142 [cs.CR]. Nov. 2020. https://arxiv.org/pdf/2011.02142

on-premises secure research environment may obviate certain protection measures, however this is the exception not the rule, and has been standard data holder policy internationally long before Five Safes.

Planned adoption of Five Safes, in the *DAT Bill* for example, suggests self-assessment of Five Safes or assessment by data curator. Much discussion has been made over accreditation. However, no proposed legislation involving Five Safes instructs *appropriate vs. inappropriate* application of the framework. As long as a data sharing or release has passed through a Five Safes assessment, so the thinking goes, the data holder is free of liability. This is unlikely to be the case when the first privacy breaches through Five Safes occur, particularly regarding institutional reputation. LR-3.3.4 Question 6 “should there be a defence from liability...if it is satisfied with a Five Safes assessment?” should be answered in the negative, as being satisfied with a Five Safes assessment does not mean that suitable protections have been put in place.

The word “safe” is used frequently in the Data Strategy (“safe, de-identified analysis” p. DS-33; “safe protected analysis of meter data” p. DS-33) as is “robust” (p. LR-52) while nothing in the framework makes data sharing or release inherently safer than without it. While the Data Sharing Principles might hold appeal, it is no “privacy-by-design” (p. DS-39); t is not the case that “consensus is forming around the use of the Five Safes Framework” (p. DS-42); nor is it “internationally recognised” (p. LR-75) in any real sense of the phrase. The framework has not undergone any substantive peer review by data privacy experts or legal scholars.

Recommendation 1. The Data Strategy should not rely on the Five Safes for protecting privacy.

2. From “De-identified” to Security Properties and Threat Models

Non-falsifiable or unverifiable definitions of “de-identification” (p. DS-19, 32, 33, LR-17) invite ad-hoc privacy solutions as witnessed by this author in the 2018 Myki[4] and 2016 Medicare/PBS[5] releases, alongside (obsolete but once) peer-reviewed proposals such as *k*-anonymity[6]. Merely making data in an “aggregated form” (p. LR-17) is well known to be inadequate due to differencing[7] and reconstruction[8] attacks which are typically able to recover micro data. Stating that “confidential information is omitted” (p. LR-17) assumes that data related to a person can be classified as not identifying or sensitive - the fallacy of the quasi-identifier - with enough entropy about a person, that person can be identified. Finally, harm does not always require unique identification at all as has been shown in homogeneity attacks.[9]

Common to these approaches is a false intuition that a dataset has been rendered somehow anonymised without any consideration of what *security property* is being asserted, nor consideration of a *threat model* of the kinds of attacks that the “de-identification” process will successfully protect against. Technical

[4] Chris Culnane, Benjamin I. P. Rubinstein, and Vanessa Teague. *Stop the Open Data Bus, We Want to Get Off*. Aug. 2019. arXiv: 1908.05004 [cs.CR]. https://arxiv.org/pdf/1908.05004

[5] Chris Culnane, Benjamin I. P. Rubinstein, and Vanessa Teague. *Health Data in an Open World*. Dec. 2017. arXiv: 1712.05627 [cs.CY]. https://arxiv.org/pdf/1712.05627

[6] Pierangela Samarati, Latanya Sweeney, *"Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression"*, Harvard Data Privacy Lab, 1998.

[7] Chris Culnane, Benjamin I. P. Rubinstein, and Vanessa Teague. *Privacy Assessment of De-identified Opal Data: A report for Transport for NSW*. May. 2017. arXiv: 1704.08547 [cs.CR]. https://arxiv.org/pdf/1704.08547

[8] Irit Dinur and Kobbi Nissim, *Revealing information while preserving privacy*. In Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, 2003, pp. 202–210

[9] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. *"l-diversity: Privacy beyond k-anonymity."* ACM Transactions on Knowledge Discovery from Data (TKDD) 1(1), 2007.

measures that do not provide a security property cannot be argued to provide *any* measurable privacy. Most well documented (technical) privacy failures can be traced back to a failure of threat model thinking.

Recommendation 2. For technical privacy measures in data sharing or release, the Data Strategy should prefer those that provide security properties that assert protection against an identified threat model.

A common but flawed counter argument is that formal approaches to privacy, by contributing stronger protection (or measurable protection at all), must necessarily sacrifice benefits of data sharing or release. Cryptographic protocols are commonplace and protect against computationally bounded adversaries when storing data on untrusted storage devices, transmitted data through untrusted networks, or processing data on untrusted cloud infrastructure. Differential privacy[10] has successfully been employed at scale by the U.S. Census Bureau for the 2020 U.S. Census[11] and Google's COVID-19 Community Mobility Reports[12] among many organisations releasing data or sharing data with untrusted third parties.

The Data Strategy calls for "best practice" (p. DS-7), "appropriate privacy and security safeguards" (p. DS-6). The best practice in privacy protection today involves consent, cryptographic protocols, and differential privacy. This, coupled with secure research environments for higher-risk sharing with trusted recipients ("secure digital labs" p. DS-58) would constitute a starting point for "managing de-identification and the risk of re-identification" (p. DS-43; Best Practice Approach Step 6).

3. Data Minimisation and Transparency

The Data Strategy entertains a culture change to the current "default, to one which authorises protected data sharing where these are safe controls and clear benefits for all Australians" (p. DS-5), that "rejecting a data sharing request under the Five Safes framework should not be a common occurrence" (p. LR-77), while claiming that "consent processes vary widely and cause constraints" (p. DS-17).

Public benefit is a strong motivation to streamline legislation across the Core Bodies. However a key tenet of privacy-by-design is that of "data minimisation": collecting and storing only the data for which consent has been freely given, and which has a clear need and purpose for the customer and their service.

Recommendation 3. Where it does not fulfil a clear service or public interest need, data should not be collected. Where it is collected, customer consent should first be obtained (and freely given) for all intended uses, even when strong privacy protections have been put into place. In rare circumstances, Core Bodies might decide (or be legislated in such a way) that public interest supersedes consent; in such circumstances, customers must be actively informed of how their data is shared or released.

The Data Strategy calls for "approaches to privacy and security [that] are robust and transparent" (p. DS-21). Transparency is a recurring theme to promote market efficiency and improve customer protections. Privacy protections that adopt concrete security properties typically eschew "security through obscurity": their protections do not rely on keeping implementations secret. As such data subjects are typically made

[10] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. *"Calibrating noise to sensitivity in private data analysis."* In Theory of Cryptography Conference, 2006, pp. 265-284.
[11] John M. Abowd, *"The US Census Bureau adopts differential privacy."* Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018
[12] Google LLC, *"Google COVID-19 Community Mobility Reports"*, https://www.google.com/covid19/mobility/ Accessed: 27/11/2020

aware of specific privacy protections (or even able to view published implementations): transparency leading to more robust implementations and engendering trust in data holders and users.

Recommendation 4. Prefer privacy protections and more broadly data release mechanisms whose details and operation can be transparently shared with data subjects. Invite independent auditing.

Technical privacy measures such as cryptography and differential privacy adopt threat models that eschew "security through obscurity" - they do not require opaque, secret, implementations, and provide protection even when code implementations are published. This is popularly known as Kerckoff's Principle. An important benefit, apart from more reliable implementations with fewer bugs, is transparency. Transparency of technical privacy measures, in the case of differential privacy, permits post-release improvement to accuracy of data analysis on shared data.[13] Transparency engenders trust from data subjects and develops the social license to collect and share data.

4. The Uncertainty Barrier

Uncertainty as to whether privacy protections such as ad-hoc de-identification really work is an understandable barrier to data sharing, but one which cannot be fixed by legislating that data should be shared irrespective of adopted protections.

The consultation paper offers an alternate framing of an uncertainty barrier, identifying "complexity leading to uncertainty" (p. DS-5), "a culture...from 'what are the risks of disclosing data?'" (p. LR-9), "tendency to not disclose data where uncertainty exists" (p. LR-25), "lack of guidance or certainty" (p. LR-29).

The recurring emphasis on uncertainty frames not only inconsistencies in legislation but legislated privacy protections and public interest tests as leading to uncertain outcomes for proposals to data sharing and release: "despite having broad statutory rights, it is clear from the RFI process that Core Bodies face difficulties sharing data between themselves" (p. LR-20).

It is understandable that uncertainty in data sharing and release frustrates legitimate needs to share energy data. Vague and ill-defined privacy protections lead to uncertain PIAs, uncertain scope for customer consent, and ultimately unacceptable uncertainty around the risk of releasing or sharing data. If regulation were to recommend privacy protections with clear security properties and concrete threat models, the arguably inherent uncertainty in current sharing and release practices would be reduced if not eliminated.

5. Data Leadership

It is advisable that the Coordination Group (viz. Recommendation 23 and Question 3 "wider representation" and Question 5) includes in its membership expertise in data privacy to complement energy domain experts, economists, data scientists and other stakeholders in "fostering best practice and capacity development" (p. DS-53). A significant refresh to legislation, policy and culture of data sharing and release is an opportunity for the Australian energy sector to join organisations such as Apple and the U.S. Census Bureau as privacy innovators or fast followers; effectively managing data for positive outcomes for the sector and consumers alike.

[13] Raj Chetty and John N. Friedman, *A Practical Method to Reduce Privacy Loss When Disclosing Statistics Based on Small Samples,*in American Economic Review Papers and Proceedings, 109, pp. 414-420, 2019.