CCG WORKING PAPER

(April, 2024)

Use of Anonymised Data in AI system – A comparative perspective

Centre for Communication Governance
National Law University Delhi

Authors: Jhalak M. Kakkar[1], Nidhi Singh,[2] Prateek Sibal[3] and Charitarth Bharti[4]

NOTE:

This is a working paper written for publication in a book titled 'Challenges and Opportunities in AI Regulation and Data Governance: An examination into Singapore, India, and China's approaches'. This is an early draft of the chapter, published to gain feedback on the scholarship.

Please reach out to us at nidhi.singh@nludelhi.ac.in for any comments/feedback on the contents of this chapter.

Acknowledgements: This paper was made possible by the generous support we received from the National Law University Delhi (NLUD). The Centre for Communication Governance (CCG) would therefore like to thank our patrons, the Vice Chancellor Professor (Dr.) G.S. Bajpai and the Registrar Prof. (Dr.) Ruhi Paul of NLUD for their constant guidance. We would also like to thank our Faculty Director Dr. Daniel Mathew for his continuous direction and mentorship.

[1] Executive Director, Centre for Communication Governance at National Law University Delhi.
[2] Project Manager, Centre for Communication Governance at National Law University Delhi
[3] Programme Specialist, Digital Innovation and Transformation at UNESCO
[4] Co-Chair - Toronto IAPP KNet Chapter

ABOUT THE NATIONAL LAW UNIVERSITY, DELHI

The National Law University Delhi is one of the leading law universities in the capital city of India. Established in 2008 by an Act of the Delhi legislature (Act. No. 1 of 2009), the University is ranked second in the National Institutional Ranking Framework for the last five years. Dynamic in vision and robust in commitment, the University has shown terrific promise to become a world-class institution in a very short span of time. It follows a mandate to transform and redefine the process of legal education. The primary mission of the University is to create lawyers who will be professionally competent, technically sound and socially relevant, and will not only enter the Bar and the Bench but also be equipped to address the imperatives of the new millennium and uphold the constitutional values. The University aims to evolve and impart comprehensive and interdisciplinary legal education which will promote legal and ethical values, while fostering the rule of law.

The University offers a five year integrated B.A., LL.B (Hons.), a one-year postgraduate masters in law (LL.M), and a Ph.D. program, along with professional programs, diploma and certificate courses for both lawyers and non-lawyers. The University has made tremendous contributions to public discourse on law through pedagogy and research. Over the last decade, the University has established many specialised research centres and this includes the Centre for Communication Governance (CCG), Centre for Innovation, Intellectual Property and Competition, Centre for Corporate Law and Governance, Centre for Criminology and Victimology, and Project 39A. The University has made submissions, recommendations, and worked in advisory/consultant capacities with government entities, universities in India and abroad, think tanks, private sector organisations, and international organisations. The University works in collaboration with other international universities on various projects and has established MoU's with several other academic institutions.

ABOUT THE CENTRE FOR COMMUNICATION GOVERNANCE

The Centre for Communication Governance at the National Law University Delhi (CCG) was established in 2013 to ensure that Indian legal education establishments engage more meaningfully with information technology law and policy and contribute to improved governance and policy making. CCG is the only academic research centre dedicated to undertaking rigorous academic research in India on information technology law and policy in India and in a short span of time has become a leading institution in Asia. Through its academic and policy research, CCG engages meaningfully with policy making in India by participating in public consultations, contributing to parliamentary committees and other consultation groups, and holding seminars, courses and workshops for capacity building of different stakeholders in the technology law and policy domain. CCG has built an extensive network and works with a range of international academic institutions and policy organisations. These include the United Nations Development Programme, Law Commission of India, NITI Aayog, various Indian government ministries and regulators, International Telecommunications Union, UNGA WSIS, Paris Call, Berkman Klein Center for Internet and Society at Harvard University, the Center for Internet and Society at Stanford University, Columbia University's Global Freedom of Expression and Information Jurisprudence Project, the Hans Bredow Institute at the University of Hamburg, the Programme in Comparative Media Law and Policy at the University of Oxford, the Annenberg School for Communication at the University of Pennsylvania, the Singapore Management University's Centre for AI and Data Governance, and the Tech Policy Design Centre at the Australian National University.

The Centre has had multiple publications over the years including reports on Intermediary Liability in India, a report Mapping the Blockchain Ecosystem in India and Australia, a UNDP Guide on Drafting Data Protection Legislation, a book on Privacy and the Indian Supreme Court, Hate Speech Report, and most recently two essay series, one on Democracy in the Shadow of Big and Emerging Tech, and a second on Emerging Trends in Data Governance. The Centre has launched freely accessible online databases - Privacy Law Library (PLL) and High Court Tracker (HCT) to track privacy jurisprudence across the country and more than sixteen

jurisdictions across the globe in order to help researchers and other interested stakeholders learn more about privacy regulation and case law. CCG also has an online 'Teaching and Learning Resource' database for sharing research oriented reading references on information technology law and policy. In recent times, the Centre has also offered courses on AI Law and Policy, Technology and Policy, and first principles of cybersecurity. These databases and courses are designed to help students, professionals, and academicians build capacity and ensure their nuanced engagement with the dynamic space of existing and emerging technology and cyberspace, their implications for the society, and their regulation. Additionally, CCG organises an annual International Summer School in collaboration with the Hans Bredow Institute and the Faculty of Law at the University of Hamburg in collaboration with the UNESCO Chair on Freedom of Communication at the University of Hamburg, Institute for Technology and Society of Rio de Janeiro (ITS Rio) and the Global Network of Internet and Society Research on contemporary issues of information law and policy.

ABSTRACT

The use of artificial intelligence systems and big data analytics has affected all sectors of the economy. The increased use of data driven AI-models has created new sectors and revolutionised others. The use of data has therefore become an invaluable economic asset for companies/states which are working to extract maximum value from data.

This increased use of data has also led to growing concerns surrounding the privacy and security of data. Data processors have historically used de-identification techniques like pseudonymisation and anonymisation to remove personal identifiers from data so that it can be used, while respecting the privacy of data subjects. Anonymisation has been considered a 'silver bullet' solution to irreversibly de-identify data in a way which can exempt it from the ambit of personal data protection laws. However, new developments have shown that re-identifying anonymised data is possible using auxiliary information.

The failure of anonymisation has led to widespread debates in academia, wherein scholars now argue about whether there should be a distinction between personal and non-personal data. Many argue that personal data can never be irreversibly anonymised and converted into non-personal data and that States must therefore consider risk-based approaches to the protection of personal data being used to train AI systems. States have started recognising that anonymising data cannot be considered to be sufficient protection, and there has been an increase in the implementation of additional administrative and regulatory safeguards to data protection regimes.

This paper looks at the legal and regulatory regimes surrounding anonymisation in developed and emerging economies around the world with a special focus on three Asian countries, namely India, China and Singapore. The paper covers laws from Europe, US, Singapore, China, and India. It compares the approaches of Asian countries with other data protection regimes and tries to bring the Asian perspectives and context into the debate surrounding the use of anonymisation globally.

INTRODUCTION

The use of artificial intelligence systems and big data analytics has affected all sectors of the economy. The use of data and AI systems have created new economic opportunities with an expected contribution of 15.7 trillion USD to the global GDP by 2030.[5] The use of data has therefore become an invaluable economic asset for companies and States which are working to derive maximum value from data.

The large-scale collection and use of data about individuals has led to the exploitation of personal data in a variety of ways, from training machines, to revealing voter preferences and influencing shopping behaviour.[6] On the one hand, practises that rely on data analytics are of great economic value,[7] on the other hand, competitive advantages gained by data-driven models contribute to the hoarding, commodification and monetisation of data that is at times detrimental to the privacy interests of data subjects.[8]

"De-identification" is the general term for the process of removing personal information from a record or data set.[9] In the context of this paper, we use the term de-identification as a general term which covers different types of de-identifications techniques such as anonymisation, pseudonymisation, etc. The process of de-

---

[5] 'Data Is Giving Rise to a New Economy' [2017] *The Economist* <https://www.economist.com/briefing/2017/05/06/data-is-giving-rise-to-a-new-economy> accessed 17 May 2022; Maxwell Wessel, 'How Big Data Is Changing Disruptive Innovation' [2016] *Harvard Business Review* <https://hbr.org/2016/01/how-big-data-is-changing-disruptive-innovation> accessed 17 May 2022; Frank Holmes, 'AI Will Add $15 Trillion To The World Economy By 2030' (*Forbes*) <https://www.forbes.com/sites/greatspeculations/2019/02/25/ai-will-add-15-trillion-to-the-world-economy-by-2030/> accessed 17 May 2022.

[6] Louise Matsakis, 'The WIRED Guide to Your Personal Data (and Who Is Using It)' *Wired* <https://www.wired.com/story/wired-guide-personal-data-collection/> accessed 17 May 2022.

[7] 'Measuring the Economic Value of Data and Cross-Border Data Flows: A Business Perspective' (*OECD*) <https://www.oecd.org/digital/measuring-the-economic-value-of-data-and-cross-border-data-flows-6345995e-en.htm> accessed 17 May 2022.

[8] 'The Rise of Data Capital' (MIT Technology Review 2016) <http://files.technologyreview.com/whitepapers/MIT_Oracle+Report-The_Rise_of_Data_Capital.pdf> accessed 17 May 2022.

[9] 'Deidentification Guidelines for Structured Data' (Information and Privacy Commissioner of Ontario 2016) <https://www.ipc.on.ca/wp-content/uploads/2016/08/Deidentification-Guidelines-for-Structured-Data.pdf> accessed 17 May 2022.

identification is used to protect the privacy of individuals when their personal data is being used for machine learning or AI systems.[10]

Traditionally, the practice of anonymisation, as a technique of de-identification, was used to distinguish between personal and non-personal data, where anonymised data was considered to be non-personal data.[11] However, in the context of AI and Big Data, recent evidence shows that anonymisation is not sufficient to protect privacy as data accretion or the combination of anonymised data can lead to re-identification of data subjects.[12]

The conversation around anonymisation is rapidly evolving within the privacy and data protection landscape. This paper attempts to map key ideas emerging from both academic and regulatory discourse to better inform ongoing regulatory initiatives in this domain.

The paper contributes to the Asian level dialogue on AI and data protection in two ways:

I. It discusses the academic literature that underpins the global discourse on de-identification as a means to balance privacy and the economic benefits of big data by reviewing developments concerning de-identification in computer science and law.

II. It examines the state practice surrounding the use of de-identification globally and maps the risk-based approaches found in several privacy and data protection regulations.

The paper is organised as follows. Section one discusses technical concepts related to de-identification, anonymisation and pseudonymisation. Section two discusses the

---

[10] ibid.

[11] Section one of the paper discusses the definitions in detail.

[12] Luc Rocher, Julien M Hendrickx and Yves-Alexandre de Montjoye, 'Estimating the Success of Re-Identifications in Incomplete Datasets Using Generative Models' (2019) 10 Nature Communications 3069.

academic literature on the efficacy of different de-identification practices in the context of big data and artificial intelligence. It argues that de-identification by itself is not sufficient for personal data protection. Section three builds upon the arguments in section two and discusses risk-based approaches to privacy and data protection proposed in the academic literature as potential solutions. Section four discusses state practices concerning de-identification from different jurisdictions including Europe, USA, Singapore, China, and India. Section five considers the trends from academia and state practices and concludes the paper. It looks at the emerging models based on risk-based approaches which are now emerging in certain States and provides recommendations that the States must consider while formulating their own data protection regimes.

1. TECHNICAL CONCEPTS RELATED TO DE-IDENTIFICATION

This section discusses the technical concepts related to de-identification, including personal and non-personal data, anonymisation and pseudonymisation.

1.1. Personal and Non-Personal Data

Data is defined as "information, especially facts or numbers, collected to be examined and used to help decision-making or information in an electronic form that can be stored and used by a computer".[13] Given the all-encompassing definition of data, data protection laws typically categorise data into personal and non-personal data so as to better define the scope of legal protection.

Personal Data contains identifiers that allow direct or indirect identification of individuals (data subjects).[14] Direct identifiers include information such as name, phone number, email and indirect identifiers include information like pincode, date of birth, gender, etc.[15] Non-personal data is any set of data which does not contain personally identifiable information.[16]

---

[13] 'Data' <https://dictionary.cambridge.org/dictionary/english/data> accessed 17 May 2022.

[14] Boris Lubarsky, 'Re-Identification of "Anonymized" Data' [2017] Georgetown Law Technology Review <https://georgetownlawtechreview.org/re-identification-of-anonymized-data/GLTR-04-2017/>.

[15] 'Deidentification Guidelines for Structured Data' (n 9).

1.2. De-identification and De-identification Techniques

The International Organisation for Standardisation (ISO) defines de-identification as "any process of reducing the association between a set of identifying data and the data subject".[17]

Two widely used techniques of de-identification are anonymisation and pseudonymisation. Anonymisation is defined as the "process by which personal data is irreversibly altered in such a way that a data subject can no longer be identified directly or indirectly, either by the data controller alone or in collaboration with any other party".[18]

On the other hand, pseudonymisation is another type of de-identification that both "removes the association with a data subject and adds an association between a particular set of characteristics relating to the data subject and one or more pseudonyms".[19] [20]

Anonymisation and pseudonymisation are different categories of de-identification.[21] It is important to note the distinction between different de-identification techniques as, in many cases, the type of de-identification technique used determines the restrictions on further processing of the data.[22] For example, anonymisation is often

---

[16] Michèle Finck and Frank Pallas, 'They Who Must Not Be Identified - Distinguishing Personal from Non-Personal Data Under the GDPR' (2020) 10 SSRN Electronic Journal 11; Aashish Aryan, 'Explained: What Is Non-Personal Data?' *The Indian Express* (New Delhi, 27 July 2020) <https://indianexpress.com/article/explained/non-personal-data-explained-6506613/> accessed 1 June 2022.

[17] 'Health Informatics — Pseudonymization' (*International Standards Organization*, 2017) <https://www.iso.org/obp/ui/#iso:std:iso:25237:ed-1:v1:en> accessed 1 June 2022.

[18] ibid.

[19] ibid.

[20] The GDPR defines pseudonymisation with both technical and organisation parameters as "the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person".

[21] 'Opinion 05/2014 on Anonymisation Techniques' (Article 29 Data Protection Working Party 2014) <https://www.pdpjournals.com/docs/88197.pdf>.

considered to make the data subject completely unidentifiable, and therefore such data falls outside the ambit of legislation such as the General Data Protection Regulation ('GDPR'); however, the use of pseudonymisation techniques does not exempt data controllers from the ambit of the GDPR.[23]

Anonymisation removes any direct or indirect identifiers from personal information, such that the data subject can no longer be identified.[24] Anonymised data is considered to be non-personal data. Such data falls outside the scope of privacy regulations, whereas pseudonymised data is still considered to be personal data due to its high risk of re-identification. Pseudonymised data is therefore subject to privacy regulations, with some leeway.[25] [26]

However, studies have shown that anonymisation does not guarantee irreversible de-identification.[27] As a result, concerns have been raised about the exemption of

---

[22] Oleksandr Tomashchuk and others, 'A Data Utility-Driven Benchmark for De-Identification Methods' in Stefanos Gritzalis and others (eds), *Trust, Privacy and Security in Digital Business* (Springer International Publishing 2019) <https://link.springer.com/chapter/10.1007/978-3-030-27813-7_5>.

[23] Recital 26 'General Data Protection Regulation' (*General Data Protection Regulation (GDPR)*) <https://gdpr-info.eu/> accessed 1 June 2022.

[24] 'Guidance on Anonymisation and Pseudonymisation' (Data Protection Commission 2019) <https://www.dataprotection.ie/sites/default/files/uploads/2019-06/190614%20Anonymisation%20and%20Pseudonymisation.pdf> accessed 1 June 2022; 'Introduction to Anonymisation: Draft Anonymisation, Pseudonymisation, And Privacy Enhancing Technologies Guidelines' (Information Commissioner's Office 2021) <https://ico.org.uk/media/about-the-ico/consultations/2619862/anonymisation-intro-and-first-chapter.pdf> accessed 1 June 2022.

[25] Gerald Spindler and Philipp Schmechel, 'Personal Data and Encryption in the European General Data Protection Regulation', (2016) 7 <https://www.jipitec.eu/issues/jipitec-7-2-2016/4440/spindler_schmechel_gdpr_encryption_jipitec_7_2_2016_163.pdf> accessed 1 June 2022.

[26] Recital 26 of the GDPR provides that data is considered to be anonymised if it is 'reasonably likely' that it cannot be traced to an identified or identifiable natural person. Whether it is 'reasonably likely' to trace the data to an identifiable person is to be determined based on objective factors, including the time and cost required for identification and the available technology. Thus, anonymisation under the GDPR is based on a standard of reasonability that considers the practical likelihood of identification, and not an absolute standard based purely on the technical possibility of identification, implying that the test under Recital 26 embraces a risk-based approach. Where there is a reasonable risk of identification, the data is pseudonymised and thus treated as personal data, whereas if the risk is reasonably unlikely/negligent (even if not excluded with absolute certainty), the data is anonymised and the principles of personal data protection do not apply to it.; See also art. 4(5), GDPR, which indicates that pseudonymisation prevents direct identification of a person through attribution with the data, but not through any other means reasonably likely to be used to identify an individual; Michèle Finck and Frank Pallas, 'They Who Must Not Be Identified—Distinguishing Personal From Non-Personal Data Under The GDPR' (2020) 10 International Data Privacy Law.

[27] 'Opinion 05/2014 on Anonymisation Techniques' (n 21).

anonymisation under different data protection legislation.[28]

In this paper, anonymisation and de-identification are not used interchangeably, however, and the term de-identification is used as a broader classification which includes both anonymisation and pseudonymisation.

1.3. Re-identification

Re-identification is the process of associating data in a de-identified data-set to the data principal or subject.[29] A re-identification attack is an operation performed on a dataset with the intention of identifying the de-identified information in a dataset. It is primarily done by linking large publicly available datasets or other auxiliary data[30] to the anonymised dataset.[31]

Regulators recommend implementing a "motivated-intruder test," where organisations assess the possibility of individuals being re-identified from anonymized data by an intruder with reasonable competence, access to resources like the Internet, and the employment of investigative techniques, such as consulting people with additional knowledge of the data subject's identity. This test assumes that the intruder lacks specialised knowledge.[32] It prompts organisations to consider potential intruders (internal or external) and their motivations, such as disgruntled employees, efforts to discredit the research team or funder, or investigative journalists, and evaluate the protective measures against these threats.[33]

---

[28] Shruti Dhapola and Aashish Aryan, '"No Data Is Permanently Anonymised": Experts Warn of Re-Identification Risks' *The Indian Express* (9 March 2021) <https://indianexpress.com/article/technology/tech-news-technology/personal-data-protection-bill-no-data-is-permanently-anonymised-experts-warn-of-re-identification-risks-7219992/> accessed 1 June 2022.

[29] 'Privacy Enhancing Data De-Identification Terminology and Classification of Techniques' (*International Standards Organization*) <https://www.iso.org/obp/ui/#iso:std:iso-iec:20889:ed-1:v1:en> accessed 1 June 2022. Section 3.31.

[30] Auxiliary data is information that helps in re-identification (for example census data, product reviews, or comments posted online).

[31] Elizabeth Brasher, 'Addressing the Failure of Anonymization: Guidance from the European Union's General Data Protection Regulation' (2018) 2018 Columbia Business Law Review 209.

[32] Bridget Treacy, 'Big Data: How Firms Can Exploit It without Breaking Privacy Laws' (*ZDNet*, 29 November 2012) <https://www.zdnet.com/article/big-data-how-firms-can-exploit-it-without-breaking-privacy-laws/> accessed 6 June 2022.

Luc Rocher, Julien M. Hendrickx & Yves-Alexandre de Montjoye have surveyed recent re-identification attacks on supposedly anonymised datasets.[34] For instance, in 2016, a journalist and a data scientist procured a database of 3 billion URLs from 3 million German users from a data broker. They were able to identify some users directly from the information available in the URL, even though the broker had claimed that the data was anonymised. For the other users, they built a probabilistic model that could uniquely identify someone by inferring information about them from only about 10 URLs.[35]

Researchers at the University of Melbourne have re-identified individuals from a de-identified dataset containing 30 years of medical billing data using publicly available information or by combining the data with other commercial datasets.[36]

A study on geospatial data with taxi trips data made available by the New York Taxi and Limousine Commission to inform traffic management and city planning found that poorly applied pseudonymisation techniques exposed unique taxi IDs in the dataset. This taxi ID data when matched with other publicly available datasets could reveal sensitive information about taxi drivers.[37] Another study involving an anonymised dataset containing 3 months of credit card data for 1.1 million people was able to uniquely re-identify 90% of the individuals using just four spatiotemporal data points.[38]

The next section discusses the academic literature on the effectiveness of anonymisation.

---

[33] UCL, 'Anonymisation and Pseudonymisation' (*Data Protection*, 24 April 2019) <https://www.ucl.ac.uk/data-protection/guidance-staff-students-and-researchers/practical-data-protection-guidance-notices/anonymisation-and> accessed 16 April 2024.

[34] Rocher, Hendrickx and de Montjoye (n 12).

[35] Alex Hern, '"Anonymous" Browsing Data Can Be Easily Exposed, Researchers Reveal' *The Guardian* (1 August 2017) <https://www.theguardian.com/technology/2017/aug/01/data-browsing-habits-brokers> accessed 1 June 2022.

[36] Chris Culnane, Benjamin IP Rubinstein and Vanessa Teague, 'Health Data in an Open World' (arXiv 2017) arXiv:1712.05627 <http://arxiv.org/abs/1712.05627> accessed 1 June 2022.

[37] Marie Douriez and others, 'Anonymizing NYC Taxi Data: Does It Matter?', *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (2016).

[38] YA de Montjoye and others, 'Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata' [2015] Science 347 <https://dspace.mit.edu/handle/1721.1/96321> accessed 1 June 2022.

2. ACADEMIC DEBATE ON THE EFFECTIVENESS OF ANONYMISATION

The academic literature on de-identification and re-identification has been growing since the early 2000s with contributions from computer scientists, legal scholars and policy experts. Some scholars argue that due to the failure of anonymisation as a technique to irreversibly protect privacy of individuals, it is futile to draw a distinction between data as personal and non-personal data in law. While recognising the limits of anonymisation, other scholars argue that it is still a useful technique to protect privacy, along with other process based safeguards like limited release of data, and access controls among others, mitigate the risk of re-identification of individuals in anonymised data.

2.1. The Failure of Anonymisation and the Futility of Characterising Data as Personal and Non-Personal

On the one hand are scholars who consider anonymisation to have failed in protecting privacy against data accretion and combination made possible by AI and Big Data analysis based methods. They argue that the ability to re-identify individuals by combining different datasets from de-identified data has rendered the distinction between personal and non-personal data found in data protection legislation irrelevant.[39]

Paul Ohm sparked an academic debate on the efficacy of anonymisation as a data protection practice that demarcates personal and non-personal data. As per Ohm "re-identification science disrupts the privacy policy landscape by undermining the faith we have placed in anonymisation".[40] He argues that the debates focussed on the distinction between personal and non-personal data are "missing the point" since an adversary can invade an individual's privacy irrespective of whether the data used for re-identification is classified as personal or non-personal data.[41] This distinction

---

[39] Ohm's arguments on the failure of anonymisation are built upon the advancements in re-identification science. One critique of Paul Ohm's article is the broad use of the term anonymisation even when data was pseudonymised.

[40] Paul Ohm, 'Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization' (2010) 57 UCLA Law Review 77.

between personal and non-personal data becomes immaterial since every re-identification of seemingly non-personal data aids further identification through a process of accretive re-identification that uses different data sources to identify individuals even in anonymised datasets.[42] A few early studies that demonstrate the re-identification of individuals from de-identified datasets have been influential in this discourse.[43] [44] [45]

2.2. The Proponents of Maintaining the Distinction Between Personal and Non-Personal Data based on Anonymisation

On the other hand, there are scholars who acknowledge the limits of anonymisation but argue that it provides a sufficient level of privacy and data protection in

---

[41] ibid.

[42] Ohm explains the accretion problem by relying on studies by Computer Scientists Narayanan and Shmatikov who have shown that "once any piece of data has been linked to a person's real identity, any association between this data and a virtual identity breaks the anonymity of the latter."

[43] Re-identification from de-identified medical records: Latanya Sweeney re-identified a medical-record dataset by linking quasi-identifiers like zip code, sex and date of birth in a de-identified medical record dataset with an easily accessible voter list. # # She found that through just three attributes including a 5-digit zip code, sex and date of birth, 87 per cent of the US population could be uniquely identified. Latanya Sweeney, 'Simple Demographics Often Identify People Uniquely' 3 Data Privacy Working Paper 34.

[44] Re-identification from Netflix anonymous movie ratings: Arvind Narayanan and Vitaly Shmatikov developed a de-anonymisation algorithm for high-dimensional micro data to re-identify individuals and reveal sensitive information like their political preferences.# They used the Netflix Prize dataset that contained anonymous movie ratings of 500,000 users to show that someone with background information like the film ratings given by users on the publicly available websites like the Internet Movie Database could help identify them in the Netflix dataset and infer sensitive information about them. Among other aspects, their research demonstrated that "k-anonymization completely fails on high-dimensional datasets, such as the Netflix Prize dataset and most real-world datasets of individual recommendations and purchases". Arvind Narayanan and Vitaly Shmatikov, 'Robust De-Anonymization of Large Sparse Datasets', *2008 IEEE Symposium on Security and Privacy* (2008).

[45] Re-identification from pseudonymised AOL search queries: The AOL research team published the search queries of 650,000 of its users with a unique identifier for each AOL account instead of user names. They had intended this data to be used for research, however the New York Times and others identified the grave privacy risk that such data posed to individuals whose search queries were released. For instance, they were able to reach out to several individuals whose queries led reporters to their address. Michael Barbaro and Tom Zeller Jr, 'A Face Is Exposed for AOL Searcher No. 4417749' *The New York Times* (9 August 2006) <https://www.nytimes.com/2006/08/09/technology/09aol.html> accessed 2 June 2022; Paul Boutin, 'You Are What You Search' [2006] *Slate* <https://slate.com/technology/2006/08/the-seven-ways-that-people-search-the-web.html> accessed 2 June 2022.

combination with process-based or administrative controls on data sharing.[46] Some scholars like Ann Cavoukian and Daniel Castro argue that de-identification works and questions about its effectiveness are greatly exaggerated by questioning the evidence on re-identification and its interpretation in legal literature.[47]

For instance, they acknowledge that while Latanya Sweeney's early research on re-identification of medical records helped advance the policy discussion on data protection, its claims of re-identification of 87 percent of the US population based on three data attributes would drop to re-identification of 0.2 percent of the US population if date of birth is replaced by month and year of birth and the ZIP code is replaced by county name.[48] Cavoukian and Castro propose that "one must assess the quality and appropriateness of the de-identification techniques used on the dataset before weighing in on the effectiveness of de-identification".[49] They also underline that since de-identification and re-identification are highly specialised fields in computer science, the real-world risk of re-identification may be significantly lower than what academic studies claim as technical expertise and access to additional datasets for re-identification may be hard to obtain.

Contrary to arguments by Paul Ohm, researchers across computer science and law are not in favour of abandoning the idea of a distinction between personal and non-personal data. For instance, scholars such as Schwartz and Solove consider Ohm's proposal concerning the elimination of the distinction between personal and non-personal data as "problematic" since the concept "establishes the boundaries of privacy regulation" without which the scope of privacy law would be limitless.[50]

Schwartz and Solove do not agree with approaches that hinge upon a neat division between personal and non-personal data or no distinction at all.[51] Instead, they see

---

[46] 'FTC Issues Final Commission Report on Protecting Consumer Privacy' (*Federal Trade Commission*, 26 March 2012) <http://www.ftc.gov/news-events/news/press-releases/2012/03/ftc-issues-final-commission-report-protecting-consumer-privacy> accessed 2 June 2022.

[47] Ann Cavoukian Castro Daniel, 'Setting the Record Straight: De-Identification Does Work' (2014) <https://itif.org/publications/2014/06/16/setting-record-straight-de-identification-does-work/> accessed 3 April 2024.

[48] ibid 4.

[49] ibid 10.

[50] Paul M Schwartz and Daniel J Solove, 'The PII Problem: Privacy and a New Concept of Personally Identifiable Information' (2011) 86 New York University Law Review 1814.

personal data as two categories – identified and identifiable – each with a distinct risk profile. This view of personal data as identified and identifiable lays the foundation for risk-based approaches to determining the efficacy of data protection practices.

Two aspects are clear from the discussion above. First, there is support for anonymisation as a means to protect privacy even as there is increasing recognition of the fact that it is not the silver bullet that it was considered to be in early literature. Second, the effectiveness of de-identification varies with a host of factors like the type of technique used, context, and adversary motivation among others. Most scholars agree that there is always a risk of re-identification. Therefore, adopting risk-based approaches to de-identification prescribes an appropriate level of technical-organisational control over de-identified data based on the risks of re-identification associated with the data.

These risk-based approaches have been used by national statistics offices for decades to disclose census data and other national surveys and the field has a rich academic literature to back its practices in statistical disclosure controls.[52] The next section discusses risk-based approaches to data protection and their effectiveness in data protection.

3. RISK-BASED APPROACHES TO DATA PROTECTION

---

[51] Schwartz and Solove identify the following weaknesses with the current conception of personal data: "First, many people believe in an "anonymity myth,"—that is, a belief that individuals remain anonymous if they have not formally used their name. This belief is especially prevalent for cyberspace activity. Yet, the growth of static IP addresses and other developments creates some built-in identifiability when one enters cyberspace. Second, information that is initially non- PII can be transformed into PII. Third, technology itself is constantly evolving, which means that the line between PII and non-PII is not fixed but rather depends upon changing technological developments. Fourth, the ability to distinguish PII from non-PII is frequently contextual. Many kinds of information are not inherently non- identifiable, or identifiable as an abstract matter".

[52] Khaled El Emam and others, 'The Seven States of Data: When Is Pseudonymous Data Not Personal Information?' [2016] Brussels Privacy Symposium 14. El Emam and Arbuckle emphasise that risk-based approaches to re-identification require computation of re-identification probabilities given certain assumptions about the re-identification adversary.

The debate on de-identification has established that the effectiveness of de-identification techniques like anonymisation or pseudonymisation depends upon a variety of factors including the type of technique used, the context, the sensitivity of the data in question and the motivation of the adversary trying to re-identify the data. In such a case, risk-based approaches to de-identification may be considered as an appropriate regulatory response which can balance the risk between privacy and utility. This section will present an overview of risk-based approaches discussed in the academic literature.

Paul Ohm proposes a re-identification risk-based approach that does away with the distinction between personal and non-personal data and focuses on "how close an adversary is to connecting a given fact to a given individual". This approach broadens the scope of data protection laws to cover all data. However, cognisant of the regulatory burden of an enlarged scope of data protection laws, he proposes to regulate "entities that amass massive databases containing so many links between so many disparate kinds of information" that carry a high risk of re-identification "even if they delete from their databases all particularly sensitive and directly linkable information".[53]

In Ohm's view, based on the level of risk assessed, regulators may choose to do nothing or impose "new restrictions on data collection, use, processing, or disclosure, and requiring specific data safe-handling procedures".[54] Therefore, based on the assessed risk, regulators may be required to perform a cost benefit analysis to justify any restrictions on data collection, use, processing, or disclosure.[55]

---

[53] Ohm (n 40).

[54] ibid.

[55] As per Ohm, once the entities to be regulated are identified, the risk of re-identification should be assessed based on five factors i) Data-Handling Techniques: Different de-identification techniques may be categorised per the probability of re-identification of data ii) Private Versus Public Release: The risk of re-identification is lower if the flow of data is between trusted parties instead of being made publicly available. Consequently, data which is being released between private 'trusted' parties may be at a lower risk of re-identification as compared to data which is released to the public as a whole. iii) Quantity: Large sized datasets provide more chances of re-identification because of a higher number of data attributes that can then be linked with other datasets. iv) Motive: The motive of data controllers influences their incentive for re-identification. For instance, academic researchers dealing with sensitive data rarely have the intention to re-identify individuals, however, the same may not be true of a private corporation looking to monetise personal data for product marketing. v) Trust: A

Schwartz and Solove critique Ohm's risk assessment framework that calls for a costs and benefits analysis for data release in advance as being too onerous and also being a detriment to open science – a concern shared by other researchers.[56] [57] Instead, they have proposed three categories where data can be about i) identified, ii) identifiable, or iii) non-identifiable person. The three categories are not differentiated by hard rules but by the use of standards where "a standard is an open-ended decision making yardstick, and a rule is a harder-edged decision making tool".[58] [59]

Ira Rubinstein and Woodrow Hartzog argue that the heavily contested de-identification domain is only part of "a larger approach to protecting the privacy and confidentiality of data subjects known as statistical disclosure limitation (SDL)".[60] To advance the debate, they propose a strong process oriented risk minimisation approach. Their process-based approach draws inspiration from the field of data security where "companies can be liable even in the absence of an actual breach because the law mandates procedures, not outputs".

They have proposed seven risk factors built upon the risk framework developed by

---

certain category of data controllers like academic researchers may be more trusted than government or third-party data controllers.

[56] Jane R Bambauer, 'Tragedy of the Data Commons' (2011) 25 Harvard Journal of Law and Technology <https://papers.ssrn.com/abstract=1789749> accessed 2 June 2022.

[57] Schwartz and Solove (n 50).

[58] ibid 1870. "First, standards are generally the superior choice for dealing with situations of rapid change, because, as mentioned, rules can become obsolete. Indeed, rules function best when an area of social and technological development has reached a fairly settled state. A second ground to prefer defining Personally Identifiable Information (PII) as a standard is the heterogeneous nature of the behavior to be regulated. As this Article has demonstrated, the means to track individuals and re-identify information are diverse. The third benefit, a standard for PII has the further merit of being capable of identifying discrete areas in which rules will be more useful for defining data as PII or non-PII. If developments in technology and society around a certain subcategory of data use become settled, it may become possible to formulate harder-edged rules to supplement a broad standard".

[59] Given the rapid technological changes, the reliance on standards instead of rules coded into law is a more agile approach to regulation as standards provide benchmarks that can be updated to take into account new technological developments without the need to amend the hard coded rules in law.

[60] Ira Rubinstein and Woodrow Hartzog, 'Anonymization and Risk' (2016) 91 Washington Law Review 703.

the National Institute of Standards and Technology (US) and Information Commissioner's Office (UK). These risk factors are:

I. Data Volume: The volume of data affects the risk of re-identification as large volume datasets have a higher degree of unicity or the number of attributes that can be used to re-identify data subjects.

II. Data Sensitivity: When the data contains sensitive information on health or finances, the incentive to launch re-identification attacks and hence the threat of privacy loss is higher.

III. Data Recipient: The risk of re-identification increases with the type of recipient i) internal recipient, ii) trusted recipient and iii) general public as a recipient.[61]

IV. Data Use: Some uses of data are riskier than others. For instance, Rubenstein and Hartzog ask “if the data be used for routine, administrative purposes like record keeping, website development, or

[61] As per Emam et al, the risk of re-identification depends further on seven states in which data may be released types of data states are: i) Public Release of Anonymized Data – In this release state, the data is publicly released with no restrictions on who can access the data and no contractual obligations on the use of data. The data is carefully de-identified to ensure high levels of privacy protection in the absence of other controls on data sharing and access. ii) Quasi-Public Release of Anonymized Data - In this release state, the data is publicly available to any recipient who shares their identity with the data holder and agrees to terms of use of data that prohibit re-identification. iii) Non-Public Data Release - In this release state, the data is made available to some recipients on request with contractual obligations concerning legitimate use of the data and the use of security and privacy measures at the end of the user. iv) Flexible Pseudonymised Data – In this release state, the data is accessible for secondary purposes and can be shared without requiring consent from the data subject. Despite this, it is not classified as non-personal data due to the continued high risk of re-identification. To mitigate this risk, three additional controls are proposed for this data release: a) no human processing, b) ensuring analysis results don't lead to re-identification, and c) exclusion of sensitive data. v) Protected Pseudonymous Data - In this release state, the data is made available under strict contractual arrangements concerning re-identification, privacy and security. Pseudonymous data only masks direct identifiers and any indirect identifiers may remain in the data. This category is not considered to be non-personal data. vi) “Vanilla” Pseudonymous Data - In this release state, direct identifiers are removed from the data but indirect identifiers remain. In addition, there are no contractual arrangements concerning data use, privacy and security. vii) Raw Personal Data - In this release state, the data is not modified or has been modified in a manner that re-identification is highly likely.

customer service? Or will it be used for commercial or discriminatory purposes? Will certain uses of data create a motivation for attackers to attempt reidentification?"[62]

V. Data Treatment Techniques: The risk of re-identification varies with the type of de-identification technique used to modify the data. Query-based methods like differential privacy offer a lower risk of re-identification as compared to techniques like suppression, generalisation or noise addition.

VI. Data Access Controls: Recipient controls can be complemented by data sharing controls to limit the risk of re-identification through contractual arrangements or through practices like downloading of data from dedicated portals instead of it being made openly available without any access controls.

VII. Consent: Data subject's consent when properly obtained for specific purposes in a transparent and easy to understand manner can enable processing of data without additional constraints.

Further, like Schwartz and Solove, Ira Rubinstein and Woodrow Hartzog advocate for the use of standards instead of hard rules for data release to ensure that data protection stays in sync with new technical challenges to privacy. Some researchers have proposed a co-regulatory approach where "regulatory agencies maintain a safe harbour list of data-sharing mechanisms appropriate for different contexts that can be maintained and regularly updated with the input of experts and stakeholders".[63]

The risk factors described above provide a useful framework for regulators to move beyond defining an anonymisation based distinction between personal and non-personal data in data protection laws. The risk-based approach is a more comprehensive way of protecting privacy by assessing the risk of re-identification along different factors discussed above.

---

[62] Rubinstein and Hartzog (n 60).
[63] ibid.

To summarise, there are three broad models of risk-based approaches that have been identified in academic literature.

- The first model, based on Ohm's scholarship, focuses on how close the adversary is to connecting facts to an individual. It focuses on regulating large entities which amass large quantities of data and suggests using a risk assessment framework which employs a cost benefit analysis to justify restrictions on data collection, use and processing.
- The second model, put forward by Schwartz and Solove, criticises the first model for being too onerous and instead proposed three categories depending on who the data is about, namely– identified, identifiable or non-identifiable person. They propose the use of standards rather than more inflexible rules to differentiate between the three categories of data.
- The third model, put forth by Rubenstein and Hartzog also advocates for standards instead of hard rules, but goes a step further to propose that de-identification should only be a part of a larger approach and advocates for process based administrative controls for risk minimisation.

Some challenges remain in defining standards for risk assessments, implementation practices and governance mechanisms for risk-based approaches to privacy and data protection. The next section discusses some examples of risk-based approaches adopted by regulators worldwide.

4. STATE PRACTICES AROUND THE USE OF ANONYMISATION

As discussed in Section 3, there are several risk factors to be considered while de-identifying data. Most legislation defines two broad tiers of risks based on the de-identification techniques applied to the data - these are pseudonymization and anonymisation. The reason for this distinction, as we have discussed above, is that in one category - anonymisation, a broad exemption from the data protection legislation applies. However, the practical process of anonymisation is complicated. There are a multitude of considerations to be taken into account to process the data and for the processed data to be considered "anonymised". The body of knowledge in this field is dynamic and entails bringing technical expertise to bear on what

constitutes anonymisation. The latest guidelines released by various regulators are extremely useful in providing necessary insight into the expectations from data controllers and processors regarding leading practices in the domain.

This also has an immense benefit for businesses and society at large as effective anonymisation permits sharing of data externally. This can include data sharing with researchers for societally beneficial research, governments for effective policy decision making or other businesses to boost competition. This concept of "open data" is being increasingly looked at for innovation and accountability.[64] For instance, France has recently enshrined into law that some categories of personal data can be published as open data in the interest of the public and society at large.[65]

In the age of AI, the importance of data cannot be overstated. Training of AI requires vast datasets with parameters running into the billions in order to achieve the level of utility we require to have AI fulfil its potential. The vast use of the data, however, contradicts a key principle of data protection - data minimisation. Given the often "black box" nature of AI systems, it is not always easy to predict how the data is processed and what is the outcome. This impacts two other key data protection principles - transparency and purpose limitation. These are just two of the several areas of concern in privacy on account of AI. By placing the data outside the ambit of the data protection laws, effective anonymisation can alleviate these concerns regarding the training of AI by providing de-identified datasets for the same,[66] while ensuring a high level of protection for the data. Regulators have also recognised the importance of such AI guidances like the ICO's anonymisation framework and specifically mention that the guide is designed for Data Controllers seeking to train AI models using vast quantities of data.[67]

---

[64] 'Anonymisation and Open Data: An Introduction to Managing the Risk of Re-Identification (Report)' (Open Data Institue 2019) <https://theodi.org/article/anonymisation-report/> accessed 2 June 2022.

[65] Laure Lucchesi, 'Le décret fixant les catégories de données diffusables et réutilisables sans anonymisation est paru' (*Etalab*) <https://www.etalab.gouv.fr/le-decret-fixant-les-categories-de-donnees-diffusables-et-reutilisables-sans-anonymisation-est-paru> accessed 2 June 2022.

[66] 'The Impact of the General Data Protection Regulation (GDPR) on Artificial Intelligence' (*European Parliament*, June 2020) <https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2020)641530> accessed 2 June 2022.

[67] 'Anonymisation: Managing Data Protection Risk Code of Practice' (Information Commissioner's Office) 4.

Anonymisation, hence, has seen a vast amount of regulatory literature written around its implementation with varying rules around the process across jurisdictions. Regulators worldwide have released detailed guidance in conjunction with their legislation to support the use of anonymised datasets by businesses trying to balance privacy considerations with business utility imperative. These documents help businesses as well as data subjects acquaint themselves with the nuances of complying with the legislation, and often in a level of detail that would not be possible to include in the legislation itself. These guidelines also elaborate on the specific techniques of how to achieve functional anonymisation.

The sections below provide a high-level overview of the legislative and regulatory landscape of key jurisdictions in Asia Pacific and the West followed by an analysis of the observable trends in this jurisprudence.

4.1. Europe

The General Data Protection Regulation ('GDPR') enacted by the European Union came into force in 2018 and has served as the benchmark for data protection regulations across jurisdictions.[68] The Data Protection Directive, which was the predecessor of the GDPR, excluded anonymised data from its ambit.[69] Like its predecessor directive, the GDPR also treats pseudonymised data as 'personal data' and follows a stringent standard of de-identification set up through Recital 26 of the GDPR.[70] The recital prescribes a 'reasonably likely' approach[71] towards the threat of re-identification stating that "an account of all means likely to be used" must be considered. For instance, the type of technology used at the time, costs, and amount of time taken to de-identify must be taken into account as factors in considering whether re-identification of the data subject is likely or otherwise.[72]

---

[68] Fernanda Nicola and Oreste Pollicino, 'The Balkanization of Data Privacy Regulation' (2020) 123 West Virginia Law Review <https://papers.ssrn.com/abstract=3824143> accessed 2 June 2022.

[69] Directive 95/46/EC 1995.

[70] Recital 26.- "The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable." 'General Data Protection Regulation' (n 23).

[71] Finck and Pallas (n 16).

The key source of guidance on anonymisation comes from the Article 29 Working Party (the predecessor of the European Data Protection Board (EDPB)), opinion on the subject.[73] The opinion by the Working Party, the guidance from other Data Protection Authorities[74] as well as the EDPB[75] have time and again differentiated between pseudonymisation and anonymisation. Pseudonymisation, whilst being de-identified and considered an additional safeguard within the meaning of Article 89(1), still carries an unacceptable risk of re-identification (hence being considered personal data within the meaning of Article 4(5)). Anonymisation, on the other hand, does not come with this risk and data put through proper anonymisation can be exempt from the applicability of the strict mandates imposed by the GDPR.

The Opinion of the Working Party recognizes that "anonymity" is interpreted differently across the EU. While some countries prefer a "computational anonymity" definition i.e., it should be computationally difficult to identify, directly or indirectly, the data subjects, others prefer an impossibility of re-identification standard, i.e it should be impossible to re-identify data subjects.[76] The Working Party's Opinion in 2014 holds "that the outcome of anonymisation as a technique applied to personal data should be, in the current state of technology, as permanent as erasure, i.e. making it impossible to process personal data."[77] [78] It is notable that the above opinion contradicts its earlier Opinion[79] on the subject in 2007 which still left room for a

---

[72] Recital 26, 'General Data Protection Regulation' (n 23).

[73]The Article 29 Working Party, set up under, article 29 of Directive 95/46/EC opinion 05/2014 (10 April 2014).

[74] 'Guidance on Anonymisation and Pseudonymisation' (n 24); '10 Misunderstandings Related to Anonymisation' (*Agencia Española de Protección de Datos*) <https://edps.europa.eu/system/files/2021-04/21-04-27_aepd-edps_anonymisation_en_5.pdf> accessed 2 June 2022.

[75] 'EDPB Documentation Response to the Request from the European Commission for Clarifications on the Consistent Application of the GDPR, Focusing on Health Research' (European Data Protection Board 2021) <https://edpb.europa.eu/sites/default/files/files/file1/edpb_replyec_questionnaireresearch_final.pdf> accessed 2 June 2022.

[76] 'Opinion 05/2014 on Anonymisation Techniques' (n 21) 27.

[77] ibid 6.

[78] This was also echoed by the country DPAs such as the Norwegian one. See A Guide to the Anonymization of Personal Data (2015), Datatilsynet. Available at: https://www.datatilsynet.no/en/regulations-and-tools/reports-on-specific-subjects/anonymisation/

measure of flexibility in what would be considered anonymised. This contradiction leaves room for what is termed as a zero-probability re-identification standard for anonymisation - one which treats anonymisation as a definitive and permanent process. This view has been criticised for deviating from a risk-based approach, causing confusion from a compliance perspective and setting unrealistic standards for anonymisation.[80]

Recent regulatory developments coming out of the EU, however, show that the approach to data anonymisation has recognized the difficulty in setting an irreversibility and zero-probability re-identification standard. The EDPB acknowledged that the "anonymisation of personal data can be difficult to achieve (and upheld) due also to ongoing advancements in available technological means, and progress made in the field of re-identification" warning controllers to approach the subject with caution.[81] A report by the European Parliament on the GDPR's impact on AI recognises the increasing methods of re-identification and proposes tackling the issue by encouraging regulators to issue clearer guidance on the "extent to which the abstract possibility of using more advanced AI techniques for re-identification may lead to de-identified data being considered as still personal".[82] The latest guidance from the Irish DPC also acknowledges the difficulty in conclusively saying the data cannot be re-identified and moves away from the zero-probability standard linking the "means reasonably likely to be used" standard of Recital 26 to a motivated intruder instead.[83]

United Kingdom

---

[79] 'Opinion 04/2017 on the Concept of Personal Data' (Article 29 Data Protection Working Party) <https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf>.

[80] Sophie Stalla-Bourdillon and Alison Knight, 'Anonymous Data v. Personal Data - False Debate: An EU Perspective on Anonymization, Pseudonymization and Personal Data' (2016) 34 Wisconsin International Law Journal 284; Finck and Pallas (n 16).

[81] 'EDPB Documentation Response to the Request from the European Commission for Clarifications on the Consistent Application of the GDPR, Focusing on Health Research' (n 75) 11.

[82] 'The Impact of the General Data Protection Regulation (GDPR) on Artificial Intelligence' (n 66).

[83] 'Guidance on Anonymisation and Pseudonymisation' (n 24).

The UK largely borrows its' privacy regime from the GDPR, as it was the prevailing law for it while it was a part of the European Union. Since then, the UK's version of the GDPR, known as the UK GDPR, has also been recognised as "adequate" by the EU.[84]

The UK Information Commissioner's Office (ICO) issued the Code of Practice on "Anonymisation: Managing Data Protection Risks" wherein it warned processors and controllers that methods of de-identification must take account of increasingly sophisticated re-identification methods.[85] Echoing a similar idea that the Irish Data Protection Authority addressed in its guidance note,[86] the ICO recognised that it is impossible to determine a technique that would be 100% effective in protecting the identifiability of data subjects. It mentions that the onus to classify a dataset as truly anonymous would lie with the controller.[87]

A draft version of the anonymisation guide that was released in May 2021 recognises the spectrum of identifiability and its application in data sharing.[88] It considers the re-identifiability potential of anonymised data and seeks a risk-based framework.[89] The United Kingdom Anonymisation Network (UKAN) has also released the Anonymisation Decision Making Framework (ADF) to address the need for a practical guide that is "primarily intended for those who have microdata that they

---

[84] On 28 June 2021, the EU Commission published two adequacy decisions in respect of the UK: one for transfers under the EU GDPR; and the other for transfers under the Law Enforcement Directive (LED). 'Commission Implementing Decision of 28.6.2021- C(2021) 4800 Final' (*European Commission*, 28 June 2021) <https://ec.europa.eu/info/sites/default/files/decision_on_the_adequate_protection_of_personal_data_by_the_united_kingdom_-_general_data_protection_regulation_en.pdf> accessed 6 June 2022; 'Commission Implementing Decision of 28.6.2021- C(2021) 4801 Final' (*European Commission*, 28 June 2021) <https://ec.europa.eu/info/sites/default/files/decision_on_the_adequate_protection_of_personal_data_by_the_united_kingdom_law_enforcement_directive_en.pdf> accessed 6 June 2022.

[85] 'Anonymisation: Managing Data Protection Risk Code of Practice' (n 67).

[86] 'Guidance on Anonymisation and Pseudonymisation' (n 25).

[87] 'Guidance on Anonymisation and Pseudonymisation' (n 24).

[88] 'Introduction to Anonymisation: Draft Anonymisation, Pseudonymisation, And Privacy Enhancing Technologies Guidelines' (n 24).

[89] ibid.

need to anonymise with confidence, typically in order to share it for some purpose in some form compliant with GDPR and the UK Data Protection Act (2018)."[90] The objective of the framework is to focus on the end state of anonymised data and the external environment in which such processes are operationalised, i.e., to see if the data has been sufficiently anonymised to meet the challenges it might face based on the release model and environment.[91]

Echoing the EU's thinking, a draft guidance on anonymisation, pseudonymisation and privacy enhancing technologies published by the ICO in May 2021, on the subject of anonymisation recognises the fact that one cannot completely rule out the possibility of re-identification and that the Controllers must proceed with considering "means reasonably likely to be used" and should be based on objective factors like the cost and time required to identify, and an assessment of the available technologies and the ones likely to come up over time.[92] The draft guidance exempts the Controllers from relying on a purely hypothetical or theoretical chance of re-identification. In its' general guidance on anonymisation[93] as well as based on a separate draft chapter released on ensuring the effectiveness of anonymisation,[94] as discussed above, the ICO also encourages the application of the 'motivated intruder test' as a starting point.[95]

The guidance, however, leaves much to the controller to decide and even leaves the frequency of the risk assessments and reviews to the controllers to determine based on context. In a boost to the risk-based processing of data for the purposes of de-identification, the ICO has released a dedicated chapter for pseudonymization as well. The chapter covers details around suitability of the process based on the

---

[90] Mark Elliot, Elaine Mackey and Kieron O'Hara, *The Anonymisation Decision-Making Framework 2nd Edition: European Practitioners' Guide* (2020).

[91] ibid.

[92] 'Introduction to Anonymisation: Draft Anonymisation, Pseudonymisation, And Privacy Enhancing Technologies Guidelines' (n 24).

[93] 'Anonymisation: Managing Data Protection Risk Code of Practice' (n 67) 22.

[94] 'Introduction to Anonymisation: Draft Anonymisation, Pseudonymisation, And Privacy Enhancing Technologies Guidelines' (n 24) 114.

[95] Treacy (n 32).

situation, the process to be adopted, when pseudonymization would be preferable to anonymisation, and guidance on choosing techniques.[96]

4.2. United States of America

The US does not have a federal data protection law and data protection regulation differs based on the state, industry and the sensitivity of the information. For instance, the US has separate data protection/privacy laws for the financial sector,[97] health,[98] credit reporting,[99] telecom,[100] student education records,[101] electronic communications,[102] children's privacy[103] and video privacy[104] amongst other sectors, and also has a divergence in privacy laws across the state and federal level. Most of the sectoral legislation in the US have adopted self-regulatory standards – an approach which has also received criticism from academics and practitioners with some describing the laws as ineffective in protecting privacy.[105]

The above legislation, hence, provide sectoral mandates and standards for de-identification of data in the US. Outside of the sectors mentioned above which have dedicated privacy laws, the data collected by a vast majority of products people use every day is not regulated, unless the state has its own privacy law. At present, only five states have introduced privacy laws although several states have bills in active consideration by respective legislatures.[106] The states with privacy laws are :–

I. California[107]
II. Colorado[108]

---

[96] 'Introduction to Anonymisation: Draft Anonymisation, Pseudonymisation, And Privacy Enhancing Technologies Guidelines' (n 24). Chapter 3.

[97] 15 U.S. Code § 6801 - Protection of nonpublic personal information.
[98] Health Insurance Portability and Accountability Act 1996.
[99] Fair and Accurate Credit Transactions Act 2003.
[100] Telephone Consumer Protection Act 1991.
[101] Family Education Rights and Privacy Act 1974.
[102] Electronic Communications Privacy Act 1986.
[103] Children's Online Privacy Protection Act 1998.
[104] Video Privacy Protection Act 1988.
[105] Ana Domingo and Nathalie Villar, 'Self-Regulation in Data Protection' [2018] BBVA Research 4.
[106] 'US State Privacy Legislation Tracker' (*IAPP*) <https://iapp.org/resources/article/us-state-privacy-legislation-tracker/> accessed 28 June 2022.
[107] California Consumer Privacy Act.
[108] Colorado Privacy Act 2021.

III. Connecticut[109]
IV. Virginia[110]
V. Utah[111]

Chief amongst these state legislation is the California Consumer Privacy Act of 2018. where the meaning of de-identified is understood to be "information that cannot reasonably identify, relate to, describe, be capable of being associated with, or be linked, directly or indirectly to a particular consumer" with the condition that a business that uses de-identified information must implement technical and process safeguards that prohibit reidentification of the consumer, has implemented processes to prevent inadvertent release of de-identified information, and makes no attempt to re-identify the information.[112] The definition above indicates the term is akin to anonymisation as understood under various other legislations.

It is important to note that these definitions recognise de-identification as a process which is subject to adequate technical safeguards, and prescribe a standard where information cannot be reasonably identified, as opposed to a strict rule where information must be irreversibly anonymised. This implies an understanding and acceptance of the fact that information may not be irreversibly de-identifiable. Additionally, the legislation also recognises two additional classes of information - *"Aggregate consumer information"*[113] *and "pseudonymised information".*[114] *It explicitly makes clear in the definition of aggregate consumer information that it is not the same as de-identified information. The above distinction along with the definition of pseudonymisation indicates a risk based tiered approach to the process allowing for more leeway to covered entities. However, the advantages of pseudonymisation*

[109] An Act Concerning Personal Data Privacy and Online Monitoring.
[110] Consumer Data Protection Act 2021.
[111] Utah Consumer Privacy Act 2022.
[112] California Consumer Privacy Act s 1798.140 (h).
[113] ibid 1798.140 (a).
"[M]eans information that relates to a group or category of consumers, from which individual consumer identities have been removed, that is not linked or reasonably linkable to any consumer or household, including via a device. 'Aggregate consumer information' does not mean one or more individual consumer records that have been deidentified."
[114] ibid 1798.140(r).
"[M]eans the processing of personal information in a manner that renders the personal information no longer attributable to a specific consumer without the use of additional information, provided that the additional information is kept separately and is subject to technical and organizational measures to ensure that the personal information is not attributed to an identified or identifiable consumer."

have been entirely restricted to the domain of research purposes under the law.[115] *A recent amendment to the CCPA also clarified a possible conflict with HIPAA and exempted HIPAA de-identified information from the scope of the CCPA as well, provided it was not re-identified.*[116]

The Department of Health and Human Services issued a guidance note on anonymisation and de-identification of Personal Health Information (PHI) to convert it into De-identified Health Information (DHI) under the Health Insurance Portability and Accountability Act (HIPAA) privacy rule.[117] As per the guidance note, the anonymised information must satisfy two standards to constitute DHI:

I. Expert Determination Standard[118] where an expert identifies the risk of re-identification is 'very small'. There is no specific professional qualification which has been designated to identify expertise in deidentification. The guidance clarifies that relevant expertise may be gained through various routes of education and experience.

II. Safe Harbour Standard[119] that entails identifiers that must be removed and the entity must not have information regarding the possibility of identification.[120] This regime provides a safe harbour to companies that

---

[115] Mitchell Noordyke, 'CCPA Offers Minimal Advantages for Deidentification, Pseudonymization, and Aggregation' (*IAPP*) <https://iapp.org/news/a/ccpa-offers-minimal-advantages-for-deidentification-pseudonymization-and-aggregation/> accessed 28 June 2022.

[116] 'Amendment to California Consumer Privacy Act of 2018.' <https://leginfo.legislature.ca.gov/faces/billCompareClient.xhtml?bill_id=201920200AB713&showamends=false> accessed 28 June 2022.

[117] The HIPAA Privacy Rule establishes national standards to protect individuals' medical records and other individually identifiable health information (collectively defined as "protected health information") and applies to health plans, health care clearinghouses, and those health care providers that conduct certain health care transactions electronically. Office for Civil Rights (OCR), 'The HIPAA Privacy Rule' (*HHS.gov*, 7 May 2008) <https://www.hhs.gov/hipaa/for-professionals/privacy/index.html> accessed 6 June 2022; Office for Civil Rights (OCR), 'Guidance Regarding Methods for De-Identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule' (*HHS.gov*, 7 September 2012) <https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html> accessed 6 June 2022.

[118] '45 CFR § 164.514 - Other Requirements Relating to Uses and Disclosures of Protected Health Information.' <https://www.law.cornell.edu/cfr/text/45/164.514> accessed 6 June 2022.

[119] ibid.

anonymise the data they collect.

The guidance note recognises the increasing adoption of health information technologies in the United States and their potential to facilitate beneficial studies that combine large, complex data sets from multiple sources. The process of de-identification is used to mitigate the privacy risks to individuals and thereby support the secondary use of data for comparative effectiveness studies, policy assessment, life sciences research, and other endeavours.

The US framework has been criticised as being ineffective as datasets once de-identified are unprotected and can be open-sourced. The guidance provides for two mechanisms of de-identification as discussed above. However, both methods, even when properly applied, yield de-identified data that retains some risk of identification and there is a possibility that de-identified data could be linked back to the identity of the patient to which it corresponds.[121]

Once de-identified these datasets are no longer considered protected under HIPAA. De-identified data is excluded from the scope of protected health information and there is no restriction upon the use or disclosure of this information. These datasets can subsequently be freely bought and sold by companies that can combine them with other information for various purposes.[122]

HIPAA specifies 18 identifiers which includes names, email addresses, social security numbers, or medical record numbers — that need to be removed for data to qualify as non-identifiable. Once these identifiers are removed using the mechanisms specified above, the data is no longer considered as protected health information'. This all or nothing approach leaves no scope of interpretation for deciding whether health data should be qualified as identifiable or not and does not

---

[120] Rights (OCR), 'Guidance Regarding Methods for De-Identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule' (n 117).

[121] ibid.

[122] Ram Sagar, 'Why De-Identifying Data Doesn't Ensure Privacy' (*Analytics India Magazine*, 2 August 2021) <https://analyticsindiamag.com/healthcare-data-privacy-deidentification-machine-learning/> accessed 6 June 2022.

take into account recent studies which show that the removal of these identifiers may still enable re-identification of individuals.[123]

4.3. Australia

The Privacy Act, 1988 governs the handling of personal information in the jurisdiction.[124] The Australian Privacy Principles (APP) contained in Schedule 1 of the legislation outline how most organisations in Australia (any organisation including government agencies, private sector and non-profits with over $3 million in revenue) must handle personal information.[125] APP 11.2 specifically mandates that when an entity under the Privacy Act no longer needs personal information for the purpose for which it was collected or may be used, the entity must take reasonable steps to destroy or de-identify the information. Section 6 of the Privacy Act defines what it means for personal information to be "de-identified", which is, when the information is no longer about an identifiable individual or an individual who is "reasonably" identifiable. Once de-identified as per the required threshold, which like most others is context dependent and detailed in guidance issued on De-Identification Decision-Making Framework,[126] the data is no longer subject to the requirements under the Privacy Act.

Re-identification became a major talking point in Australia in light of a few incidents showing the possibility and ease of re-identification - namely, when a group of Melbourne researchers revealed that sensitive patient health data could be re-identified without decryption, by linking the unencrypted parts of the dataset with known information[127] and another instance where the data of Victoria's public

---

[123] Kerstin N Vokinger, Daniel J Stekhoven and Michael Krauthammer, 'Lost in Anonymization — A Data Anonymization Reference Classification Merging Legal and Technical Considerations' (2020) 48 The Journal of Law, Medicine & Ethics 228.

[124] Privacy Act 1988.

[125] ibid 6; 'Chapter B: Key Concepts' (*Office of the Australian Information Commissioner*) <https://www.oaic.gov.au/privacy/australian-privacy-principles-guidelines/chapter-b-key-concepts> accessed 7 June 2022.

[126] 'De-Identification Decision-Making Framework' (*Office of the Australian Information Commissioner*) <https://www.oaic.gov.au/privacy/guidance-and-advice/de-identification-decision-making-framework> accessed 7 June 2022.

transport system was re-identified using only two data points.[128] This prompted the introduction of a bill in the Senate to criminalise the re-identification of the data,[129] but the bill eventually lapsed amidst concerns regarding the implications on cybersecurity research.[130] Reporting around the issue has suggested that the government is planning to introduce a fresh bill to criminalise re-identification.[131]

At present, two key pieces of literature guide the industry on anonymisation in the jurisdiction. The first is the De-identification Decision making Framework,[132] which provides a comprehensive framework for approaching de-identification in accordance with the Privacy Act. The second is the De-identification and Privacy Act guide,[133] also released by the Office of the Australian Information Commissioner, which gives "general advice about de-identification, to assist APP entities to protect privacy when using or sharing information containing personal information." The De-identification and Privacy Act guide provides guidance on when de-identification may

---

[127] Donna Kevey, 'Research Reveals De-Identified Patient Data Can Be Re-Identified' (*Newsroom - University of Melbourne*, 23 February 2022) <https://www.unimelb.edu.au/newsroom/news/2017/december/research-reveals-de-identified-patient-data-can-be-re-identified> accessed 7 June 2022.

[128] Culnane Chris, 'Two Data Points Enough to Spot You in Open Transport Records' (*Pursuit - University of Melbourne*, 15 August 2019) <https://pursuit.unimelb.edu.au/articles/two-data-points-enough-to-spot-you-in-open-transport-records> accessed 7 June 2022.

[129] 'Privacy Amendment (Re-Identification Offence) Bill 2016' (*Parliament of Australia*) <https://www.aph.gov.au/Parliamentary_Business/Bills_Legislation/Bills_Search_Results/Result?bId=s1047> accessed 7 June 2022.

[130] Allie Coyne, 'Govt Could Lose Its Battle to Criminalise Data Re-Identification' (*iTnews*) <https://www.itnews.com.au/news/govt-could-lose-its-battle-to-criminalise-data-re-identification-450272> accessed 8 June 2022.

[131] Denham Sadler, 'Govt Has Another Go at Criminalising Data Re-Identification' (*InnovationAus.com*, 27 October 2021) <https://www.innovationaus.com/govt-has-another-go-at-criminalising-data-re-identification/> accessed 8 June 2022.

[132] 'A Framework for Data De-Identification' (*Data61*) <https://data61.csiro.au/en/Our-Research/Our-Work/Safety-and-Security/Privacy-Preservation/De-identification-Decision-Making-Framework> accessed 8 June 2022.

[133] 'De-Identification and the Privacy Act' (*Office of the Australian Information Commissioner*) <https://www.oaic.gov.au/privacy/guidance-and-advice/de-identification-and-the-privacy-act#ftnref1> accessed 8 June 2022.

be appropriate, how to choose appropriate de-identification techniques, and how to assess the risk of re-identification.[134]

An interesting change in the Australian approach is that a second version of the APP, the "APP 2" provides data principals the option of dealing anonymously or pseudonymously with an APP entity.[135] The provision, which comes with certain exceptions,[136] means that the individual dealing with the APP entity cannot be identified and the entity does not collect personal information or identifiers. One key difference that emerges between the GDPR and the Privacy Act here, then, is that unlike the GDPR, the Privacy Act does not deal with pseudonymisation as an operation performed on the data enabling reduced compliance burden, but only refers to pseudonymisation in terms of the right of an individual to interact with an APP entity using a pseudonym.[137] The Australian Privacy Act does not explicitly exclude anonymised data. The meaning of de-identification within the Act is understood as when "information is no longer about an identifiable individual or an individual who is reasonably identifiable."[138] The de-identification framework released by the Office of Australian Privacy Commissioner describes it as a process that renders personal information which would otherwise be subject to the Privacy Act, into a form that is not identifiable.[139] The framework, however, does recognise that the risk cannot be completely eliminated and describes the process of de-identification as more of a risk management process than an exact science.[140]

---

[134] ibid.

[135] 'Chapter 2: APP 2 — Anonymity and Pseudonymity' (*Office of the Australian Information Commissioner*) <https://www.oaic.gov.au/privacy/australian-privacy-principles-guidelines/chapter-2-app-2-anonymity-and-pseudonymity> accessed 8 June 2022.

[136] ibid. APP 2.2(a) For instance where the entity is required or authorised by law or a court or tribunal order to deal with identified individuals, or it is impracticable for the entity to deal with individuals who have not identified themselves.

[137] ibid. APP 2.1 states that the "Individuals must have the option of not identifying themselves, or of using a pseudonym, when dealing with an APP entity in relation to a particular matter.". Also see comparison of the two privacy regimes at 'Comparing Privacy Laws: GDPR v. Australian Privacy Act' (Data Guidance 2020) <https://www.dataguidance.com/resource/comparing-privacy-laws-gdpr-v-australian-privacy> accessed 8 June 2022.

[138] Privacy Act 1988 6.

[139] 'De-Identification Decision-Making Framework' (n 126) 9.

On a provincial level, the Chief Data Officer of the state of Victoria has issued de-identification guidelines in February 2018 under section 33 of the Victorian Data Sharing Act, 2017.[141] The guideline covers various techniques of de-identification and assists industry decision makers in their choice of de-identification techniques and controls to apply in various situations by assessing the risk of re-identification by considering data sensitivity, destination organisation/persons, access environments and disclosure risks, and other factors.[142] Similarly, eHealth Queensland has also issued its own De-identification and Anonymisation of Data Guidelines.[143] Apart from general concerns in de-identification, the guidance also addresses the subject of de-identifying and anonymising medical images, which can present problems, especially when dealing with Digital Imaging and Communications in Medicine (DICOM) image files commonly used for computed tomography scans (CTs), magnetic resonance imaging (MRI) and positron emission tomography (PET).[144]

On the federal and provincial level, by and large, de-identification has received ample attention from regulators and lawmakers/government in Australia. The key difference so far from the benchmark set by the GDPR is the absence of pseudonymisation and accompanying exemptions as a middle tier from its legislative framework.

4.4. Singapore

---

[140] ibid 2.

[141] Victorian Data Sharing Act 2017. Section 18 of the Act requires the Chief Data Officer (CDO) and data analytics bodies to take reasonable steps to ensure that data is de-identified before use.

[142] 'De-Identification Guideline' (Chief Data Officer 2018) <https://www.vic.gov.au/sites/default/files/2019-03/Victorian-Data-Sharing-Act-2017-De-identification-guidelines.pdf> accessed 8 June 2022.

[143] 'De-Identification and Anonymisation of Data Guideline' (Queensland Government 2021) <https://metrosouth.health.qld.gov.au/sites/default/files/content/de-identification-and-anonymisation-of-data-guideline.pdf>.

[144] ibid 6.2.1.

In Singapore, the Personal Data Protection Act, 2012 governs the collection, use and disclosure of individuals' personal data by organisations.[145] Key amendments made to this legislation, which were enforced in 2021, expanded the uses that the data can be put to without the consent of the data subject. The amendments include the "exceptions to the consent" requirement, which now allows businesses to use, collect, and disclose data for "legitimate purposes", business improvement, and a wider scope of research and development without user consent.[146] Additionally, the amendment has introduced criminal penalties for "egregious mishandling of personal data" which includes "knowing or reckless unauthorised re-identification of anonymised data".[147]

To help organisations with compliance, the Personal Data Protection Commissioner issued advisory guidelines that describe conditions under which personal data that has been anonymised may no longer be considered personal data for the purposes of the Personal Data Protection Act. These guidelines suggest frequently used anonymisation techniques that organisations may adopt based on their operational context, for example techniques like aggregation, generalisation, data suppression and interestingly, pseudonymisation.[148] The above guidance from Singapore PDPC did not recognize pseudonymisation as a separate tier for re-identification and stated that while others use 'anonymisation' to refer to de-identification that is irreversible, it used the term 'anonymisation' to refer to "the process of converting personal data into data that cannot be used to identify any particular individual, and can be reversible or irreversible".[149]

---

[145] 'PDPA Overview' (*Personal Data Protection Commission Singapore*) <https://www.pdpc.gov.sg/Overview-of-PDPA/The-Legislation/Personal-Data-Protection-Act> accessed 6 June 2022.

[146]Personal Data Protection (Amendment) Act 2020; 'Advisory Guidelines on Key Concepts in the Personal Data Protection Act' (*Personal Data Protection Commission Singapore*) <https://www.pdpc.gov.sg/Guidelines-and-Consultation/2020/03/Advisory-Guidelines-on-Key-Concepts-in-the-Personal-Data-Protection-Act> accessed 6 June 2022.

[147] Personal Data Protection Act, 2012, Section 48F.

[148] 'Advisory Guidelines on the Personal Data Protection Act for Selected Topics' (*Personal Data Protection Commission Singapore*, 9 October 2019) <https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Advisory-Guidelines/AG-on-Selected-Topics/Advisory-Guidelines-on-PDPA-for-Selected-Topics-9-Oct-2019.pdf?la=en> accessed 6 June 2022.

Further, taking account of re-identification risks, the Personal Data Protection Commissioner suggests tests for identifying re-identification risks. It posits that the risk of re-identification of a dataset may be based upon various factors such as nature of use and extent of disclosure, public or private disclosure of the dataset, disclosure of singular or multiple datasets, the data recipient's ability and motivation to re-identify the dataset, and the changing and evolving nature of technology, which may make re-identification possible at a later stage. The guidance suggests the use of the 'motivated intruder' test[150] from the ICO's Code of Practice as the general test for assessing the risks of re-identification and the robustness of the anonymisation.[151]

A more recent publication from the Singapore PDPC shifted its stance on pseudonymisation. The Guide to Basic Anonymisation released on 31st March, 2022 speaks at length about the methodology and process of pseudonymisation (which it refers to as "de-identification") and anonymisation.[152] While the previous guidance was agnostic on anonymisation being reversible or irreversible, the newer guidance has defined de-identification separately and clearly distinguished it from anonymisation.[153] Discussing the techniques (for anonymisation as well as calculation of re-identification risk) and various steps in the process, the guide provides a comprehensive overview of the various considerations and controls to be

---

[149] 'Advisory Guidelines on Key Concepts in the Personal Data Protection Act' (n 146) 9.

[150] 'Advisory Guidelines on the Personal Data Protection Act for Selected Topics' (n 148). Para 3.27 explains - The 'motivated intruder' test considers whether individuals can be re-identified from anonymised data by someone who is motivated, reasonably competent, has access to standard resources (e.g. the Internet and published information such as public directories), and employs standard investigative techniques (e.g. making enquiries of people who may have additional knowledge of the identity of the data subject).

[151] ibid 3.26.

[152] 'Guide to Basic Anonymisation' (*Personal Data Protection Commission Singapore*, 31 March 2022) <https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Advisory-Guidelines/Guide-to-Basic-Anonymisation-31-March-2022.ashx> accessed 6 June 2022.

[153] De-identification, as per the guidance, refers to "the removal of identifiers (e.g. name, address, National Registration Identity Card (NRIC) number) that directly identify an individual. De-identification is sometimes mistakenly equated to anonymisation, however it is only the first step of anonymisation. A de-identified dataset may easily be re-identified when combined with data that is publicly or easily accessible."

borne in mind by a Controller looking to anonymise data. The guide lays down the applicable steps for the anonymisation of data, starting with –

a) Know your data
b) De-identify your data
c) Apply anonymisation techniques
d) Compute your risks
e) Manage your risks

4.5. China

There is no single comprehensive legislation in the People's Republic of China (PRC) which governs data privacy. Instead, data security and privacy are divided between several different legislations with their own domains of influence. On June 1, 2017, the Cybersecurity Law (CSL) came into effect and became the first national-level law to address cybersecurity and data privacy protection.[154] The Data Security Law (DSL) came into force on September 1, 2021, and focuses on data security across a broad category of data (not just personal information).[155] More recently and significantly, the Personal Information Protection Law (PIPL) came into effect on November 1, 2021.[156] The PIPL is the first comprehensive, national-level personal information protection law in the PRC. However, it is pertinent to note that the PIPL does not replace but only augments the privacy landscape already comprising the other laws.

---

[154] Rogier Creemers, Graham Webster and Paul Triolo, 'Translation: Cybersecurity Law of the People's Republic of China (Effective June 1, 2017)' (*DigiChina*, 29 June 2018) <https://digichina.stanford.edu/work/translation-cybersecurity-law-of-the-peoples-republic-of-china-effective-june-1-2017/> accessed 6 June 2022.

[155] 'Data Security Law of the People's Republic of China' <http://www.npc.gov.cn/englishnpc/c23934/202112/1abd8829788946ecab270e469b13c39c.shtml> accessed 6 June 2022.

[156] Rogier Creemers and Graham Webster, 'Translation: Personal Information Protection Law of the People's Republic of China - Effective Nov. 1, 2021' (*DigiChina*, 7 September 2021) <https://digichina.stanford.edu/work/translation-personal-information-protection-law-of-the-peoples-republic-of-china-effective-nov-1-2021/> accessed 6 June 2022.

The Personal Data Security Specification (PDSS)[157], which came into force in May 2018, governs the handling of personal data by organisations. The PDSS is voluntary and recommended, setting out best practices concerning protection of personal information. It is widely expected that companies of all sizes will comply with the Specification, with respect to auditing and enforcement of the Cybersecurity Law.[158] However, these specifications are only advisory and inconsistency with the PDSS would not lead to legal consequences.[159] It regulates the collection, use and retention of personal information, interests of individuals, etc.[160]

The PDSS notes that "Anonymized Personal Information is no longer deemed Personal Information".[161] It defines anonymisation as a process wherein the subject cannot be identified and the personal information cannot be restored.[162] Further, the PDSS suggests that organisations delete or 'anonymise' information post the retention period in two circumstances: If the personal information controllers stop operation or if the PI subject wishes to close an account.[163] It also proposes that organisations pay attention to risks (including risk of re-identification of anonymised or de-identified information) when sharing or transfer of data is involved and also perform PI security impact assessments.[164]

While the Chinese regime on data protection prescribes protections for personal information, the broader interpretations of privacy by Chinese courts' risk defeating the purpose of China's privacy regulations by providing a loophole to companies.[165]

---

[157] 'China Issues New Personal Information Security Specification' (*WilmerHale*) <https://www.wilmerhale.com/en/insights/client-alerts/20200324-china-issues-new-personal-information-security-specification> accessed 6 June 2022.

[158] ibid.

[159] 'Analysis of the Personal Information Security Specification from a Practical Perspective' (*KWM*) <https://www.kwm.com/content/kwm/cn/en/insights/latest-thinking/personal-information-security-specification.html> accessed 7 June 2022.

[160] 'Information Security Technology— Personal Information (PI) Security Specification' (State Administration for Market Supervision of the People's Republic of China 2020) <https://www.tc260.org.cn/upload/2020-09-18/1600432872689070371.pdf> accessed 7 June 2022.

[161] ibid 3.14.

[162] 'Information Security Technology— Personal Information (PI) Security Specification' (n 160).

[163] ibid.

[164] ibid 11.2.

This approach by the Courts of treating anonymisation as a remedy and treating rich datasets as "fully anonymised" while leaving the definition of "full anonymity" vague is reflective of a more callous approach to the subject. It allows companies to use a vague definition of anonymisation to avoid their obligations under the PDSS.

The newly passed Personal Information Protection Law (PIPL) is the first comprehensive national data protection regulation that took effect on November 1, 2021.[166] It defines de-identification and anonymisation separately[167] and refers to anonymisation as information that is 'impossible to identify and impossible to restore'.[168] The law also uses the term "de-identification" as distinguished from anonymisation with the understanding of the term "de-identification" being similar to pseudonymisation under the GDPR i.e. removal of direct identifiers.[169]

The PIPL, which is modelled largely after the GDPR gives anonymised information an explicit exemption from the application of the law.[170] Personal information which has been anonymised is excluded from the scope of the law, and third parties are prohibited to try and re-identify anonymised information they receive from handlers.[171]

Consequently, the current iteration of the law does not envision a risk-based approach towards the reidentification of anonymised data.

---

[165] Camille Boullenois, 'The Loophole in China's Privacy Regime: Anonymization · TechNode' (*TechNode*, 3 June 2020) <http://technode.com/2020/06/03/the-loophole-in-chinas-privacy-regime-anonymization/> accessed 7 June 2022.

[166] Natasha Lomas, 'China Passes Data Protection Law' (*TechCrunch*) <https://social.techcrunch.com/2021/08/20/china-passes-data-protection-law/> accessed 7 June 2022.

[167] Hunter Dorwart and Gabriela Zanfir-Fortuna, 'China's New Comprehensive Data Protection Law: Context, Stated Objectives, Key Provisions' (*Future of Privacy Forum*) <https://fpf.org/blog/chinas-new-comprehensive-data-protection-law-context-stated-objectives-key-provisions/> accessed 7 June 2022.

[168] Creemers and Webster (n 156) 73.

[169] Dorwart and Zanfir-Fortuna (n 167).

[170] ibid 4, 73; Xu Ke and others, 'Analyzing China's PIPL and How It Compares to the EU's GDPR' (*IAPP*) <https://iapp.org/news/a/analyzing-chinas-pipl-and-how-it-compares-to-the-eus-gdpr/> accessed 7 June 2022.

[171] Dorwart and Zanfir-Fortuna (n 167).

4.6. India

India adopted its data protection legislation, the Digital Personal Data Protection Act ('DPDP Act') in 2023.[172] This Act was adopted after almost seven years after the Supreme Court of India's landmark judgement in the Puttaswamy case where the Court highlighted the necessity of a personal data protection law in India.[173] In 2018, a government appointed expert committee presented a draft Data Protection Bill in India.[174] In 2019, the government introduced the Personal Data Protection Bill, 2019 (PDP Bill) in the Parliament, which was then referred to a Parliamentary Standing Committee for further examination.[175]

After almost two years of deliberation, the Joint Parliamentary Committee submitted its report on the PDP Bill in December 2021.[176] The first edition of the Digital Personal Data Protection was introduced for public comments in 2022. The current iteration of the DPDP Act was adopted by both houses of Parliament in August, 2023, however it has not come into effect as of date. The government is currently in the process of framing rules and regulations to enable the implementation of the law.[177]

Unlike previous proposed versions of the law, the DPDP Act specifically governs the use of personal data only, and does not cover the use of non-personal data within its ambit. It also does not mention anonymisation or pseudonymisation at all. It is unclear how the DPDP act would apply to data being used by AI systems at large. Unlike the 2019 edition of the bill, the Act does not provide for the mandatory sharing

---

[172] Digital Personal Data Protection Act 2023.
[173] *KS Puttaswamy v Union of India* (2017) 10 SCC 1.
[174] 'Draft Personal Data Protection Bill, 2018' (*PRS Legislative Research*) <https://prsindia.org/billtrack/draft-personal-data-protection-bill-2018> accessed 8 June 2022.

[175] 'The Personal Data Protection Bill, 2019' (*PRS Legislative Research*) <https://prsindia.org/billtrack/the-personal-data-protection-bill-2019> accessed 8 June 2022.

[176] Pallavi Bedi, 'What the JPC Report on the Data Protection Bill Gets Right and Wrong' (*The Wire*) <https://thewire.in/tech/what-the-jpc-report-on-the-data-protection-bill-gets-right-and-wrong> accessed 8 June 2022.

[177] Anirudh Burman, 'Understanding India's New Data Protection Law' (*Carnegie India*) <https://carnegieindia.org/2023/10/03/understanding-india-s-new-data-protection-law-pub-90624> accessed 17 April 2024.

of non-personal data 'to enable better targeting of delivery of services or formulation of evidence-based policies by the Central Government'.[178]

The DPDP Act also contains some key departures from global best practices on data protection for data used in AI systems by excluding all publicly available personal data from the protections contained within the Act.[179] This position also seems to contradict the privacy protections which were available under the Puttaswamy regime, which established that the protection of an individual's private sphere extends beyond physical spaces to encompass public settings as well.[180] This principle from the Puttaswamy judgement upholds an individual's right to privacy in a public space.

However, the DPDP Bill 2023 introduces a notable exemption by excluding publicly available personal data voluntarily shared by individuals from protection.[181] Under this provision, individuals would likely forfeit their rights, including the right to rectify or erase information, as outlined in the DPDP Bill, once their data is utilised by third parties from the public domain.

It cannot be assumed that individuals who share personal data in public domains consent to its unrestricted use by any third party. For instance, while users may anticipate that their vacation photos shared on platforms like Instagram could be viewed by numerous individuals, including strangers, they would not expect or desire their images to be exploited on unauthorised platforms for commercial gain.

In 2019, the Ministry of Electronics & Information Technology (MeitY) constituted a Committee of Experts to deliberate on a Data Governance Framework. The first significant development in India's journey into the regulation of non-personal data started in 2020, with the report by the committee of experts on Non-Personal Data Governance Framework.[182] The NPD report considers the proliferation of big data,

---

[178] 'The Personal Data Protection Bill, 2019' (n 175).Clause 91(2).

[179] 'Data Protection Bill: In Process and Practice, a Step Back' (*The Indian Express*, 19 August 2023) <https://indianexpress.com/article/opinion/columns/data-protection-bill-in-process-and-practice-a-step-back-8899924/> accessed 17 April 2024.

[180] *K.S. Puttaswamy v Union of India* (n 173).

[181] Digital Personal Data Protection Act s 3(c)(ii).

analytics and Artificial Intelligence systems and aims at creating data governance frameworks which can maximise the value generated by data. The report holds that one of primary drivers for AI technology is the availability of data, and discusses the economic benefits arising from monetising this data.

The Committee classified non-personal data into three types: public data, such as 'anonymised data of land records, public health information, vehicle registration data, community data, such as 'datasets collected by municipal corporations and public electric utilities, datasets comprising user-information collected even by private players like telecom, e-commerce, ride-hailing companies', and private data, like 'inferred or derived data/insights, involving application of algorithms, proprietary knowledge'. Non-personal data is therefore seen as data which was never related to natural persons, such as data on weather or supply chains; or data which was initially personal data, but has been made anonymous through use of certain techniques to ensure that individuals to whom the data relates cannot be identified.[183] Therefore de-identification is at the heart of the non-personal data framework suggested by the committee of experts.

The NPD report views anonymisation as a technique which allows data to be shared, whilst preserving privacy.[184] According to the NPD report, the process of anonymising data requires that identifiers (both direct identifiers like names and indirect identifiers like age or occupation) are changed in some way such as being removed, substituted, distorted, generalised or aggregated.[185] The NPD report notes that through anonymisation, previously personal data can also be shared and developed into datasets to be used for AI, whilst preserving privacy, considering that the process changes the data by removing, substituting, distorting, generalising or

---

[182] 'Report by the Committee of Experts on Non-Personal Data Governance Framework' (Ministry of Electronics and Information Technology 2020) <https://ourgovdotin.files.wordpress.com/2020/07/kris-gopalakrishnan-committee-report-on-non-personal-data-governance-framework.pdf> accessed 8 June 2022.

[183] ibid 13.

[184] ibid 55.

[185] 'Report by the Committee of Experts on Non-Personal Data Governance Framework' (n 182) Annex 3.

aggregating direct and indirect identifiers like names, age, or occupation from the relevant data sets.[186]

Following stakeholder consultations, the committee released a revised report in December 2020.[187] The revised report removed references to the three distinct types of non-personal data, but retained the definition of non-personal data provided in the previous report. The revised report also asked for the deletion of provisions relating to non-personal data from the 2019 PDP bill – clearly stating that the ambit of the personal data protection laws should be limited to personal data only, and that non-personal data should be governed under its own framework.

In the revised report, the Committee also evaluated what would happen in the case of re-identification from non-personal data. It held that non-personal data would continue to be regulated by the NPD framework for as long as it remains non-personal data. However, if the individuals whose data constituted the anonymised dataset were re-identified in any manner, either (a) as a result of a subsequent failure of the anonymisation technology, or (b) by virtue of the association of the anonymised dataset with other anonymised datasets that together result in re-identification, or (c) through any other means of conscious re-identification undertaken by the data fiduciary, such data would no longer be characterised as anonymised data to which the provisions of the data protection law would not apply. The dataset would be deemed to have been re-identified and would once again fall within the purview of the data protection law. The report also clarifies that the determination as to whether the PDP framework or the non-personal data framework applies to a specific kind of data would be determined by the identifiability of that data; All personally identifiable data (including anonymised data that has subsequently been re-identified) will be governed by the PDP Bill and all anonymised data that, at the time of evaluation, has not been re-identified will be governed by the non-personal data framework.

---

[186] ibid 55.

[187] 'Revised Report by the Committee of Experts on Non-Personal Data Governance Framework' (Ministry of Electronics and Information Technology 2020) <https://ourgovdotin.files.wordpress.com/2020/12/revised-report-kris-gopalakrishnan-committee-report-on-non-personal-data-governance-framework.pdf> accessed 3 April 2024.

In summary, anonymised data would remain subject to the non-personal data framework until it is re-identified, at which point it would become subject to the personal data protection regulations outlined in the PDP Bill.[188]

The revised report also recommended 'consent' for anonymised data. Keeping in mind the dangers of re-identification, and the non-application of the PDP law, the committee recommended that at the time of collecting personal data, data collectors should provide a notice and offer the data principal the option to opt out of data anonymisation.

Following this, in May 2022 MeitY released a draft of the 'National Data Governance Framework Policy' ("NDGFP").[189] The policy was applicable to all non-personal datasets and was aimed at creating large Indian datasets for the benefit of the research and start-up ecosystem. The policy also recommended setting up an India Data Management Office, which would prescribe rules and standards for anonymisation for organisations, including government and private entities, so that they can create large and diverse datasets.[190]

There is currently no concrete legal framework for the governance of non-personal data. While there is growing cognisance of the need to provide data governance frameworks for both personal and non-personal data in India, the issues related to non personal data are not addressed by the DPDP Act. Instead, it is possible that the governance of non-personal data which goes into AI models will be contained in the upcoming Digital India Bill.[191]

The upcoming Digital India Act, will replace the Information Technology Act of 2000. The upcoming Act is meant to "deal with all harms of AI", potentially including privacy harms arising from the use of data as well.[192] While there is clear recognition of the

---

[188] ibid 11.
[189] 'National Data Governance Framework Policy' (Ministry of Electronics and Information Technology 2022) <https://www.meity.gov.in/writereaddata/files/National-Data-Governance-Framework-Policy.pdf> accessed 3 April 2024.
[190] ibid 5.
[191] Deeksha Bhardwaj, 'Govt May Frame Rules for Sharing of Non-Personal Data under Digital India Law' *Hindustan Times* (26 February 2023) <https://www.hindustantimes.com/india-news/govt-may-frame-rules-for-the-sharing-of-non-personal-data-under-digital-india-law-101677355138634.html> accessed 3 April 2024.

dangers of using anonymisation as a silver bullet solution to privacy risks, it is unclear how these risks will be mitigated.[193] However the state remains cognizant of the developments around anonymisation and the blurring distinction between personal and non-personal data. In July 2022, MeitY released on the 'Guidelines for Anonymisation of Data', which provides standard operating procedures and discusses various mechanisms for implementing anonymisation techniques.[194] The draft also clearly states that while anonymisation is an effective technique, it must not be used as a silver bullet to tackle data privacy risks and that anonymisation should only be a part of the organisations overall data governance strategy.[195] With the passing of the DPDP Act it has become clear that India plans to maintain a separate legal mechanism for the governance of personal and non-personal data, but it is yet to take any concrete steps towards governing the latter.

Jurisdiction	**Legal Framework**	**Risk-based approach**
EU	There is a distinction between Anonymised and Pseudonymised data. Anonymised data is exempted from the application of the GDPR. Pseudonymised data is treated as personal data. There is a legislative recognition of risk-based tiers, and the risks to re-identification are dependent upon the mechanism used to de-identify data.	
US	California's Consumer Privacy Act requires businesses that use anonymised data to implement process safeguards to prevent and prohibit	This is similar to the second and third models of the risk-based approach. The law in the US does not lay down

[192] 'Digital India Act Wil Deal with Ill Effects of AI: MoS Chandrasekhar' (*Business Standard*, 24 November 2023) <https://www.business-standard.com/india-news/digital-india-act-will-deal-with-ill-effects-of-ai-mos-chandrasekhar-123112401136_1.html> accessed 3 April 2024.
[193] 'Revised Report by the Committee of Experts on Non-Personal Data Governance Framework' (n 187) 16.
[194] MeitY, Draft Guidelines on Data Anonymization; Aarathi Ganesan, 'Draft Guidelines on Data Anonymization Released by MeitY for Public Comments' (*MediaNama*, 1 September 2022) <https://www.medianama.com/2022/09/223-meity-draft-guidelines-data-anonymization-public-consultation/> accessed 6 September 2022.
[195] Ganesan (n 194).

	reidentification of data. The US standard requires anonymisation to be to a standard where no one can 'reasonably identify' the information- impliedly recognising that anonymisation can be reversed.	specific rules, but rather uses standards which can better adapt to developing technologies. It also prescribes process and administrative controls to prevent re-identification, in line with Rubenstein and Hartzog's approach.
Singapore	The Singaporean approach considers multiple factors which can affect risk of re-identification of anonymised data. It also lays down process-based controls for risk assessment including the motivated intruder test to judge the likelihood of reidentification of anonymised data.	This is similar to the model put forward by Rubenstein and Hartzog, and considers factors such as extent of disclosure, nature of use, disclosure of single or multiple datasets, etc.
China	The PIPL defines anonymisation as an irreversible process. It proposes that organisations should pay attention to risks, including the risk of re-identification while sharing or transferring data.	There is no concrete risk-based mechanism to prevent the risk of re-identification of anonymised data.
Australia	The APP applies to organisations with over $3 million in revenue. The APP does not completely exempt anonymised data, and it recognises that privacy risks cannot be completely eliminated through the use of de-identification.	The APP framework shows some similarity to the model suggested by Ohm where the regulation is focused on large entities which are demarcated by revenue.
India	The DPDP Act only governs personal data and does not mention anonymisation or non-personal data.	The current legal framework does not use a risk-based approach to the use of anonymised data.

5. ANALYSIS AND CONCLUSION

Our review of the legislative and regulatory framework surrounding anonymisation reveals that while there are differing viewpoints on the effectiveness of de-identification among scholars, there is a broad recognition amongst academics that the distinction between personal and non-personal data is getting blurred and de-identification techniques are not without limitations.

Another aspect that is emerging is the clear distinction between different de-identification processes, such as anonymisation and pseudonymisation, in the regulation of data protection. We can also see that most States are now moving towards a model of data protection which includes a risk-based approach to de-identification instead of exclusively relying only on anonymisation, which may be reversed and can therefore result in risks to privacy.

Drawing inspiration from the field of data security, many academics propose a sustainable process-based approach for risk-assessment (a combination of de-identification with legal and administrative tools) which suggest that companies can be held accountable even if there is no actual breach since the law mandates procedure, not outputs and outcomes. They maintain that a full range of data controls must be used by companies and propose different risk factors that a company must examine before releasing a dataset.[196]

In this paper, we have examined key western jurisdictions i.e, - Europe, USA and Australia, and compared their approach to Asian jurisdictions – China, Singapore and India. While data protection was not a creation of the GDPR or its preceding data protection directive, the central role of the “Brussels effect” in setting the agenda in the domain cannot be overlooked. Various jurisdictions espouse the core principles that the Data Protection Directive and subsequently the GDPR set into a legislative framework. The GDPR, hence, forms a trendline around which we can measure the deviations that most jurisdictions have taken around the trend. Specifically, within the domain of de-identification, this deviation can be measured along the following factors:

[196] Data controls are not limited to contracts, qualified investigators, or enclaves but to include data treatment techniques, organizational support, and mindful framing to establish a sound deidentification regime.

I. Legislative recognition of risk-based tiers: Personal Data, Pseudonymised data and Anonymised data

Along the above metrics, we see that jurisdictions like the EU and the UK have implemented a framework which includes various tiers of de-identification based on the risk of re-identification - anonymisation and pseudonymisation, and bringing with these categorisations of data, exemptions from certain provisions of the law. Provisions around restricting re-identification

The standards and definition of anonymisation adopted by most States, including Australia and certain States in the US have been cognisant of the realistic risk of re-identification and hence many jurisdictions have moved towards accepting a certain risk as long as the cost and effort of re-identification remains high enough to keep the possibility of re-identification remote. Some states like the UK and Australia specifically guard against the possibility of re-identification by barring sharing of information on methods of anonymisation and in some instances even criminalising the re-identification of de-identified datasets. However, as seen above, many jurisdictions such as the EU still prescribe a non-reversible standard of anonymisation, which is difficult to implement.

II. Issuance of regulatory guidance to implement anonymisation processes including general guidance on techniques

Most States have also released specific guidance on anonymisation to help clarify the bare provisions and mandate of the law and shed some light on expected standards for compliance from a regulatory standpoint. They provide various techniques which can be employed for de-identifying the data, how to decide when to de-identify the data, and the standard of anonymisation which must be achieved. These guidance elaborate on vague elements of the legislation on the subject of anonymisation. They address the standard to which data should be anonymised, such as "means reasonably likely to be used" to deidentify anonymised data, which must be protected against. Additionally, they discuss "all objective factors, such as the costs of and the amount of time required for identification," thereby giving industry a practical insight into the process.

Additionally, most States are either updating their data protection legislation to incorporate the contemporary debate surrounding anonymisation and pseudonymisation or are in the process of releasing guidelines to help clarify provisions that govern them. As the technology surrounding anonymisation continues to evolve, we can expect more changes in the legal and regulatory landscape to occur. Currently, regulations and laws surrounding anonymisation depend heavily upon technological processes, such as assessing the ease of reidentification of data, and as such these processes may require closer scrutiny to ensure that they adhere to the legal and regulatory standards, even as the process of anonymisation continues to evolve.

Finally, it can be seen that while Asian jurisdictions are in the process of consolidating and updating their data protection jurisprudence, mostly in line with the GDPR guidance, jurisdictions like Singapore have also started exploring technological interventions which may stand up to the evolution of anonymisation, such as the use of synthetic data which can provide guidance to other States.[197]

At a time when several countries are grappling with the application of law and regulatory frameworks to personal and non-personal data, the debates surrounding techniques like anonymisation become vitally important in shaping the discourse surrounding the regulation of data protection. The rise in the value of personal data, especially in relation to its contribution to training machine learning and artificial intelligence models and its increasing social and economic impact on all sectors has forced States to re-evaluate their data protections models. While the use of data and data analytics is at the forefront for economic and technological growth, the security and privacy implications of this widespread use of personal data is a growing cause of concern amongst States and academics. At times like these, anonymisation, which was initially the ultimate standard of protection that would allow for personal data to be converted into non-personal data without loss of privacy, becomes a key factor in the conversation surrounding data protection.

[197] 'Guide to Basic Anonymisation' (n 152).

Legal standards governing personal and non-personal data often consider the type of de-identification techniques used. For example, while anonymised data may be outside the scope of personal data protection laws like the GDPR, pseudonymised still falls within its ambit. In this way, anonymisation can often be used as a silver bullet to convert data into anonymised data, which can be removed from the ambit of personal data protection laws. However, States like Australia recognise the developments in technology which make re-identification possible and incorporate ways to deal with this in their legislation.

Scholars now argue anonymisation has failed in protecting privacy against data accretion and combination, which is made possible by AI and Big Data methods. Consequently, individuals can now be re-identified by combining various data sets, rendering the distinction between personal and non-personal data moot. On the other hand, academics also argue that the concept of non-personal data by itself may be outdated, and that the distinction between personal and non-personal data should be removed.

To this end, States have now started including various legal and regulatory safeguards, which look beyond the mere prescription of anonymisation to preserve privacy. From a brief review of the state practice of various States, we can summarise that most jurisdictions have accepted that there are various levels of de-identification, and that anonymisation can no longer be treated as irreversible in all circumstances. However, the economic benefits of data analytics still play a vital role in determining the balance of privacy and its utility in many States.

States have consequently tried to employ several measures to deal with the risks posed by re-identification including the introduction of risk-based frameworks, and guidance to clarify the scope of the law surrounding anonymisation. Most States balance the risk of re-identification with the economic advantages which the access to de-identified personal data would provide and consequently apply a risk-based framework which depends upon the risk of re-identification based upon the time and resources involved. With such a broad acceptance and move towards risk-based approaches, it is also important for States to evaluate the short-comings of risk-based approaches and work towards mitigating those as well.

A comparative study of the data protection regimes in Asian countries- China, Singapore and India, shows that to a large extent, the models for data protection in these States are based on European models such as the GDPR. However, each of these countries has contextualised data protection laws to their own context. The anonymisation regime in the three countries- China, Singapore and India are also vastly different. At this juncture, where countries are in the process of developing their norms surrounding personal and non-personal data, they must actively engage with the scientific and technological developments happening in this space and try to contextualise the deployment of techniques like anonymisation to their own legal framework. States should be cognisant of the developments happening in this sector and their regulatory guidance must stay up to date with the developments in technology.

Based upon the review of literature and state practice we have undertaken for this paper; we can see that there is no 'silver-bullet' solution to the current challenges with anonymisation. States are working to maintain a balance between the privacy and security of personal data, while also being able to leverage this data for its economic potential. Our brief discussion highlights that while there is no clear path available yet, there are a few general recommendations to be considered

I. States must recognize that the process of de-identification has different levels, which offer differential levels of privacy and security. While anonymisation may offer more protection, it cannot be treated as being irreversible, and this must be recognised in the state's policy. Consequently, anonymised data can no longer be given a blanket exemption from the ambit of data protection laws and there must be additional safeguards prescribed for data protection.

II. Additional legal and regulatory protections for re-identified data must be considered. With technology outpacing legal and regulatory guidance, States must have in place guidelines and penalties to discourage re-identification. Additionally, data subjects must also be provided some degree of control over their information, even after it has been anonymised, given the possibility of its re-identification.

III. Administrative safeguards for data protection must also be supported by robust mechanisms which can be used to identify the risk of re-identification of data. This would have to consider factors like the sensitivity of the data and the ease of re-identification based upon technical considerations of the time.

With many countries around the world working towards developing their AI capabilities, States can also learn from the experiences of the others, to evaluate and adopt mechanisms which are more contextualised to their needs.

For instance, India, which has just adopted its data protection law, and is currently working towards creating non-personal data frameworks, and a new law to govern the harms from emerging technologies like AI, would have to consider how to best balance the economic benefits which can be derived from the use of large datasets with potential privacy violations and the significant harms this can cause to individuals. From our previous discussions of the academic approaches and state practice as covered in the paper we can see that there are several avenues which can be considered as this new law is developed.

It remains to be seen how other countries will tackle the challenges posed to deidentification of data as the advances in technology continues to pose greater challenges to the privacy and security of personal data. As States develop their guidelines surrounding anonymisation and re-identification through the use of personal data protection laws or other guidelines, it is also important to address overall concerns regarding the use of personal data by the AI industry.

Centre for Communication Governance

National Law University Delhi

Sector 14, Dwarka

New Delhi – 110078

011- 28031265

ccgdelhi.org

privacylibrary.ccgnlud.org

twitter.com/CCGNLUD

ccg@nludelhi.ac.in