Policy Brief

Response to the Expert Committee Report on NON-PERSONAL DATA GOVERNANCE FRAMEWORK

TANDEM RESEARCH

September 2020

Response to the Expert Committee Report on

NON-PERSONAL DATA GOVERNANCE FRAMEWORK

BY Tandem Research

13 September 2020

TANDEM
RESEARCH

Executive Summary

We recognise and concur with the motivation for the proposed framework - to better distribute gains in a digital economy, to tap the social and public value of data, and to identify appropriate institutional and regulatory structures for a well-functioning data society.

However, in our view, the proposed framework is unlikely to achieve its intended objective. We strongly caution against the proposed distinction between personal and non personal data and recommend pursuing alternative measures to redistribute value in a data economy. Moreover, the framework risks undermining individual rights and liberties, and will have a negative impact on innovation, business, and markets.

1. The distinction between personal and non-personal data is not tenable. Whether data relates to an individual or not cannot be derived solely from the source of the data. It is also dependent on the purposes for which the data is processed and the context in which it is used. The inadequacy of anonymisation techniques also discredits the argument that anonymised personal data can be considered non-personal data. Further, even anonymised data can have harmful or exclusionary effects on communities.

2. Identifying something as a commons requires specifying clear boundary conditions. The lack of a clear boundary between personal and non-personal data challenges claims for data commons and for data as a national resource. Similar to other commons, data commons are also likely to face appropriation and sustainability dilemmas.

3. Addressing the market dominance of Big Tech firms and redistributing value in the data economy can be better achieved through other means. We recommend three strategies:(a) update competition law to include control over data and network effects; (b) platform neutrality, so that Big Tech platforms cannot unfairly discriminate against other businesses using their platform; and, (c) platform interoperability, to enable consumer choice and reduce the weight of network effects.

4. Computer scientists are seeking ways to reduce the reliance on Big Data for developing future AI capabilities. This challenges one of the core motivators for this framework. Advances in the field of AI are likely to need better computing power and talent and skill, both of which are lacking in India.

5. Mandating the sharing of metadata, and raw data under certain conditions, is likely to stifle businesses, innovation, and healthy markets. It could also stifle competition and benefit larger players in the domestic ecosystem. The ambiguity around thresholds for data businesses and data sharing further compounds this problem. The proposed framework is likely to fall short in advancing the goal of redistributing value in the data economy.

6. Data stewardship models like data trusts can help balance privacy and innovation. But, many issues around community representation, accountability, and power dynamics would need to be addressed before these models can become a basis for new data governance frameworks .

7. The proposed framework could enable overreach by the state, infringing on individual rights and liberties, and cause a chilling effect on the functioning of democracy and markets. We do not endorse the idea that data is a national resource and its use must pass the tests of proportionality and legitimacy.

8. While the committee has recommended separate collection of consent for anonymisation and further use of the anonymised data as a safeguard, if the definition of processing is construed to include anonymisation, as it has been in the EU, then users merely consenting to processing will also be consenting to anonymisation.

1. The framework rests on a number of problematic assumptions

Characteristics of Data

Data is created through the processes of identification, quantification and measurement, and these processes are shaped by societal priorities, values and culture. Data does not simply represent the reality of the world; rather, it is a construction of the world.

This implies that there is no such thing as 'raw data' or 'factual data'. Data harvested through measurement is always a selection from the total sum of all possible data available - data is thus inherently partial, selective and representative.[1] Data is not a natural resource, but a cultural one - it needs to be collected and the process of data collection has its own agenda.[2]

Data also does not hold any intrinsic value. This value is assigned by societal processes. The document suggests that data is an economic good. This is true of the current digital economy, but that is not because of the nature of data itself, but because of business models that assign value to data.

The report also argues that data is non-rivalrous. This idea needs further interrogation in the context of a digital economy. The over extraction or overuse of data can result in a loss of individual agency, and in this way data can be considered rivalrous . Companies are competing to capture consumer data, by offering integrated products and services, to create 'walled gardens' through network effects. Companies are battling for the ability to exclusively harvest personal data even more intensively as the value of personal data grows.[3]

<u>Distinction between Personal and Non-Personal Data</u>

The proposed framework suggests that non-personal data is any information which is not related to an identified or identifiable natural person. However, the question of the ways in which data can relate to natural persons is not identified - the only way data is assumed to relate to someone is in terms of its origin or source.

However, information can be said to relate to someone in terms of content (being about a person - eg. name, health status); purpose (when the purpose of processing is to evaluate, treat, or influence the status of a behaviour or person) or in terms of the result of impact (when the processing of data is likely to have an impact on a person's rights and interests).[4]

This implies that it is not necessary that the data focuses on someone for it to relate to him/her. Data that relates to a person is broader than the data about the purpose; the intended and unintended impact or the likelihood of impact, of data also needs to be taken into consideration.

Further, whether data relates to an individual or not is context-dependent. It could depend on a number of factors, such as the reason for collecting and processing that data, or the entity in possession of the data. It is near impossible to distinguish a-priori between data that does or does not impact people.[5]

The proposed framework also says that data which was initially personal data, but which was later anonymised, can be considered non-personal data. This is conceptually and operationally flawed. Even anonymised data can impact individuals - anonymised aggregated data is used to categorise individuals and make decisions about them as members of a group. This can have a direct impact on individuals' access to opportunities and services, and contribute to discriminatory practices, not all of which are directly visible.

Even if identifying features are scrubbed from the metadata, it would still contain data points on which generalizations could be made for data principal communities. For example, if anonymised personal data for an underdeveloped neighbourhood suggests that the demographic mostly comprises individuals below the age of 18 who would be unable to cast votes, it could lead to less attention and priority in improving civic facilities in the area.

Moreover, there is now plenty of evidence to show that anonymization techniques are inadequate in protecting individual privacy - re-identification is increasingly possible.[6] The assumption of successful anonymization is central to the proposed framework, but the question of how such re-identification can be addressed is only referred to in the annex, and that too in the form of a list of anonymization techniques, with no discussion of their efficacy or adoption.

The report further breaks down non-personal data into public non-personal, community non-personal, and private non-personal. The examples provided under each reinforce our argument above - that whether data is personal or non-personal depends on the context and purpose of processing. For example, the report argues that a university collecting pollution levels in a city through a publicly funded project is an example of a public non-personal data. However, if such data is used to determine health insurance premiums, then this data is directly related to an individual.

Data Commons

Identifying something as a commons requires specifying clear boundary conditions that clearly demarcate what belongs in the commons. As argued above, the boundary between personal and non-personal data is fluid and context-dependent. Hence, it is untenable to think of a data commons. To argue for a data commons in the absence of clear boundaries would then also bring personal data into the fold of the commons.

Data as a commons is also likely to experience problems associated with other commons - i) how to sustain the common resource or prevent against its depletion and ii) prevent against the unfair or harmful appropriation of the resource. As noted above, we do not agree with the proposition that data is non-rivalrous. The sustainability dilemma is thus not about the exhaustion of a physical resource, but the long term effects of the commodification of personal data and modern data

processing practices. Unsustainable data practices can lead to the disempowerment of people.[7] The appropriation dilemma arises from the competition for data, and the likelihood of the greatest benefits accruing to the most powerful, resulting in the enclosure of data by a few. Data commons in other words are not immune from power struggles.

Data as a National Resource

The epistemic and operational difficulty in distinguishing between personal and non personal data also makes it untenable to speak of data as a national resource, as proposed in the framework. It would directly contravene the fundamental right to privacy and individual autonomy .

Community

The framework lacks a clear definition of 'community'. Community can refer to a collection of individuals but it could also include new communities or social groups that are produced as a result of big data analytics. People can belong to multiple communities; membership can change; power struggles can ensue within communities; and communities can also be exclusionary. The framework also does not consider situations in which certain members of a community may wish to be excluded from data intelligence. Further, any conversation about community data needs to be accompanied by a conversation around community rights - we do not yet have adequate frameworks or rules to assign or protect community rights.

2. There are better ways to redistribute value in a data economy : Three alternatives.

There are alternative ways to check the dominance of a few large technology firms and redistribute value in the data economy. We propose the following policy pathways. [8]

Update Competition Policy

First, update competition policy to include control over data, network effects and the impact of mergers and acquisitions on market competition. This could prevent the dominance of a select few firms and enable a more competitive marketplace. Similarly, mergers and acquisitions need to be assessed differently. Evaluation of market power of the merged entity needs to incorporate control over data, possibilities of vertical integration, and whether the merger poses a barrier to entry for new firms.

Germany, for example, is considering a new approach to competition law, which considers, among other things—(a) direct and indirect network effects; (b) switching costs for users; (c) the economies of scale that may arise as a result of network effects and (d) how access to data affects competition and innovation in the market.[9]

Europe will soon be presented with a 'choice screen' to the default search engine on Android phones/tablets, allowing Android users to choose a search engine of their preference. This is a result of the current tussle between the European Commission and Google on the latter's breaching

of antitrust rules through its dominance in the market. For companies to offer their search engine to Android users, they have to participate in an auction led by Google, held every three months.[10]

Platform Neutrality

Second, platform neutrality should be mandated to prevent dominant platforms from unfairly discriminating against other businesses using their platform.[11] This would mean, for example, that Amazon-branded products compete on equal terms with those of other retailers over the Amazon platform. Traditional utilities like electricity and railways need to follow certain principles because they provide indispensable infrastructure. Similarly, Big Tech firms provide essential utilities for a digital economy and should adhere to similar principles. Just like traditional utilities, Big Tech companies should not be allowed to differentiate between consumers that use it, be it individuals or businesses. Such differentiation could take many forms—special access to data, prioritization of search results and allocation of space on feeds or pages. These issues are magnified for Big Tech firms, because they are often both the platform, as well as a player on the platform.

France, in 2016, passed a law on platform fairness which mandates that platforms:
a) Do not change rankings in a way that goes against user interests and do not give preference to their services over those of their competitors;
b) Inform users how algorithms work and flag results that are sponsored;
c) Publish rules around removal of lawful content and not discriminate;
d Communicate any changes in algorithms/ content policies to suppliers in advance.[12]

Platform Interoperability

Third, since data is a key resource for Big Tech firms, democratizing access to data will improve the competitive health of digital markets. One such solution is data portability, i.e. allowing users to transfer their data from one platform to another. This feature is now part of most data protection laws, including India's draft bill. However, data portability alone will be insufficient because the individual's network of friends, family, buyers, sellers or followers will still be on the original platform.It could also pose a challenge for privacy as it may result in a person carrying over data belonging to their social network while moving to a different platform.

A better solution might be platform interoperability. Just like a Gmail user can send an email to a user of any other service, platform interoperability would allow WhatsApp users to message Signal users or allow Amazon users to combine their order with items from Flipkart.By doing so, platform interoperability allows consumers to choose whichever platform they like, thereby reducing the weight of network effects. It will also be a boon for other businesses, which will be less dependent on Big Tech firms. It could fuel innovation as follow-on innovators will leverage pre-existing tools to create better products, building off of existing platforms' strengths and allowing users to interact with multiple services at the same time.

A bipartisan coalition in the US has introduced the Augmenting Compatibility and Competition by Enabling Service Switching (ACCESS) Act, as a response to the current dominance that Big Tech firms have over data in the United States. Like mobile number portability, it enables users of social

media platforms to move their data and even communicate across platforms. The proposal intends to provide start-ups with a level playing field by supplementing data portability with interoperability. It requires that platforms allow users to download their data and transfer it to another service, if desired. It also permits users to authorize access by third-party apps, subject to safeguards.[13]

3. Advances in Artificial Intelligence are likely to be less reliant on Big Data

One of the motivations underlying the report is to allow India to compete in the global AI race. Access to data is perceived to be essential to becoming an AI superpower. This assumption however requires re-examination in light of recent advances in machine learning and artificial intelligence. New advances in computer science are making AI less reliant on bottom-up big data, and more on top-down reasoning.[14] Companies are already experimenting and developing such capacities.

Computer scientists are recognising the limits of deep learning models to carry out even basic reasoning. Current machine learning models are based on identifying and reproducing patterns, but easily miss the 'edge cases', and reproduce existing societal biases. Computer scientists are trying to figure out ways to train systems on less data and AI experts expect the "data" variable in the AI growth equation to become less important, with small datasets overtaking big data as drivers of new AI innovation.[15]

A multitude of emerging AI-powered tools also rely not only on 'big data', but instead entirely on small data or a combination of both big and small data with the help of techniques such as collective learning, transfer learning, and meta-learning.[16] Wilson and Daugherty argue that small data is what could help organisations make advances in what they term the big-data arms race.[17]

Beyond the question of data, AI progress requires significant amounts of computing power. In 2018, OpenAI's research found that the amount of computational power required to train the largest AI models at the time was doubling every month after 2012.[18] After adding new data to their analysis, they found that between the years 1959 and 2012, the amount of computational power required doubled every two years. This indicates that computing power needed to train AI models today is increasing seven times faster than before, implying that AI is not only dependent on data but also resource intensive.[19]

Talent and skill are other critical variables .A survey by Great Learning found that even though India almost doubled it's AI workforce between 2018 and 2019, a large number of positions remained vacant, indicating that the field faces a shortage of employable workers.[20] This indicates that strengthening AI-related education and skilling in India warrants special focus.

4. No clarity on threshold for Data Business

The report introduces a form of entity categorised as a 'data business'. Organizations which derive economic value from data - by collecting, storing, processing and managing data - will have to mandatorily register as a data business after having reached a certain data-related threshold. A detailed disclosure of information including metadata about the data user and community will be required by businesses over the threshold as part of this registration process and this metadata will be made available through open access to citizens and organisations. The disclosure is aimed at enabling users to identify potential opportunities to combine data from multiple Data Businesses, governments through the metadata.

The data-related threshold concept put forth by the report remains unclear. There is no specificity or indication of the factors that will be considered in developing the threshold for determining Data Businesses. Further, the threshold intends to distinguish small from big businesses, but does not indicate what besides the threshold would determine this distinction. For example, if a firm falls above the threshold, what determines it as either small or big? Such questions arise in light of the report indicating that small businesses within the threshold will not be mandated by the data sharing requirements.

5. Detrimental for Businesses and Markets

The report posits that data businesses will provide open access to metadata and regulated access to the underlying data. Data is core to the business offering of firms, and such a move could undermine competition and healthy markets, and even enable new domestic monopolies.

The report states that the mandated data requirements may be voluntary when the Data Business is small but still above the data threshold. However, due to the lack of clarity on how this is determined, a measure that is intended to challenge the dominance of large firms, could inadvertently affect smaller businesses as well. In this way, the proposed framework could undermine its objective of promoting a vibrant domestic ecosystem.

Further, because raw data could be available for a price, albeit under certain conditions, this could give an unfair advantage to larger players in the domestic ecosystem.Once again, this could be harmful for smaller businesses. Such meta and raw data does not only pertain to the businesses customer base, but can also contain ancillary data that is core to the unique business offering of the firm. Access to such data will enable firms to imitate these otherwise exclusive models or services.

If businesses are aware that any data that plays a vital role in their business offering, will be shared with their competitors albeit at a price, it could lead to market wide dis-incentivisation. This not only throws into disarray existing market dynamics but will fail to achieve the very intent of promoting fair competition and innovation through non-personal data.

6. Data trusts are experimental ideas and should not yet be the basis of new data governance frameworks

In theory, data stewardship models such as data trusts allow users to maintain control over their data and have a collective say in how data is used, while allowing data to also be used for public good.[21]

However, to make these models succeed, we need frameworks to establish trust and accountability and engage the community in determining what constitutes a fair exchange of data.[22] Moreover, these models need to address issues arising from unequal power and representation that other such 'common resource' models are also prone to. Another problem with data stewardship models is that the range of possible harms is not obvious or even visible to the affected people in many cases.[23]

The report does not consider these issues. Data trusts are still an experimental concept. Far more attention needs to be paid to the design of these trusts, before resting crucial data governance frameworks on the assumption of their efficacy and legitimacy.

Further, the committee's suggestion of entrusting public authorities and industry bodies with the management of data trusts creates a classic principal - agent dilemma. While the data trusts are supposed to act in the interest of the principal, these could run contrary to the interests of the government or industry bodies from which they derive their membership.

Tensions could arise between the data trustee's mandate of representing the best interests of data principals and meeting the needs of the government or the market, which could interfere with the impartiality expected from data trusts, creating a classical principal - agency dilemm[24] This is likely to be further exacerbated since the committee has also recommended another layer between data trustees and data principals - that of data custodians.

7. Proposal risks undue overreach by the State

The report suggests that non-personal data can be accessed for sovereign purposes, such as national security, law enforcement, legal or regulatory purposes, without establishing any checks and balances against overreach by the state.

We have argued earlier in this response that the distinction between personal and non-personal data is dependent on the context, and cannot exist as absolute categories. We have also therefore noted that data cannot be thought of as a national resource - the category of public non personal data is untenable. Accordingly, we posit that the state's access to non-personal data for sovereign purposes must similarly pass the tests of proportionality and legitimacy, as laid out by Puttaswamy judgement.

The clause that the state can mandate such data sharing for preventive purposes is of particular concern. This can result in the profiling and targeting of individuals and communities, impinging on the fundamental right to privacy. It can also have a chilling effect on democracy, if it leads to self-censorship.[25]

8. Possible inconsistencies with the provisions of the PDP

The report suggests that data principals (users) provide consent for anonymisation and use of the anonymised data, at the same time that they provide consent for collection and use of their personal data, acknowledging that consent given for personal data doesn't automatically apply to non-personal data.

This recommendation, however, may be moot (along with the protection it hopes to offer). Clause 3(31) of the PDP Bill that defines 'processing' includes the terms 'adaptation' and 'alteration'. These terms (borrowed from the GDPR's definition), could be construed to include de-identification (anonymisation) as has been the case in the EU.[26] This clause in the PDP can render void the protection the non-personal data report aims to afford through additional consent. Consenting to processing under the PDP might already incorporate such consent to anonymisation.

Further, even if consent for anonymisation and use is separately sought when a user is consenting to the collection and use of their data, explicit consent would require the user be told exactly the purpose or use for which their data is being anonymised. This throws up the question of how the purpose of anonymisation and further use will be known at the time the data is collected.

We do not believe consent provides adequate protection to the user. Informed consent only works when the request is infrequent and where the harms are imaginable and severe. These are the conditions under which informed consent works in a medical context, for example. However, none of these conditions apply in the context of the data economy. Studies demonstrate that consent obtained seldom truly is free, informed, specific, and clear, as it is required to be under the PDP bill (Clause 11).[27]

Tandem Research is an interdisciplinary research collective that generates policy insights at the interface of technology, society, and sustainability. We believe that evidence-based policy, supported by broad-based public engagement, must steer technology and sustainability trajectories in India. Our work seeks to ensure that India's technology transitions are just and equitable.

www.tandemresearch.org
hello@tandemresearch.org

343 Coimavaddo Quitla, Aldona
Bardez - 403508, Goa

Endnotes:

[1] Kitchin, R. (2014). The data revolution: Big data, open data, data infrastructures and their consequences. Sage.
[2] Gitelman, L. (2013). *Raw data is an oxymoron*. MIT press.
[3] Purtova, N. (2017). Health data for common good: defining the boundaries and social dilemmas of data commons. In *Under observation: The interplay between eHealth and surveillance* (pp. 177-210). Springer, Cham.
[4] Graef, I., Gellert, R., Purtova, N., & Husovec, M. (2018). Feedback to the Commission's Proposal on a Framework for the Free Flow of Non-Personal Data. *Available at SSRN 3106791*.
[5] Ibid.
[6] Ohm, P. (2009). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA l. Rev.*, *57*, 1701.
[7] Purtova, N. (2017). Health data for common good: defining the boundaries and social dilemmas of data commons. In *Under observation: The interplay between eHealth and surveillance* (pp. 177-210). Springer, Cham.
[8] Aneja, U., & Chamuah, A. (2020, July 28). A Balancing Act - The Promise and Peril of Big Tech in India. Retrieved from https://tandemresearch.org/assets/Tandem-Research-Big_Tech_report.pdf

[9] Weck, T. (2020, April 14). The New Abuse Rules in the German Competition Act – What's in it for the EU? Retrieved from https://www.competitionpolicyinternational.com/the-new-abuse-rules-in-the-german-competition-act-whats-in-it-for-the-eu/
[10] Bradshaw, T. (2020, January 10). Google's Android auction puts tiny search rivals in spotlight. Retrieved from https://www.ft.com/content/3869bed2-330d-11ea-9703-eea0cae3f0de
[11] Aneja, U., & Chamuah, A. (2020, July 28). A Balancing Act - The Promise and Peril of Big Tech in India. Retrieved from https://tandemresearch.org/assets/Tandem-Research-Big_Tech_report.pdf
[12] Umhoefer, C. A. (2016, November 02). France's Law for a Digital Republic expands transparency rules – significant changes for platforms, telecoms, online providers: Insights: DLA Piper Global Law Firm. Retrieved from https://www.dlapiper.com/en/us/insights/publications/2016/11/france-expand-transparency/

[13] Robertson, A. (2019, October 23). How would opening up Facebook change the internet? Retrieved from https://www.theverge.com/2019/10/23/20926792/facebook-access-act-interoperability-data-portability-warner-hawley-bill-explainer
[14] Wilson, H. J., Daugherty, P. R., & Davenport, C. (2019, January 14). The Future of AI Will Be About Less Data, Not More. Retrieved from https://hbr.org/2019/01/the-future-of-ai-will-be-about-less-data-not-more
[15]The new AI innovation equation. (n.d.). Retrieved from https://www.ibm.com/watson/advantage-reports/future-of-artificial-intelligence/ai-innovation-equation.html
[16] Nivargi, V. (2020, May 19). The Small Data Revolution: AI Isn't Just For The Big Guys Anymore. Retrieved from https://www.forbes.com/sites/forbestechcouncil/2020/05/19/the-small-data-revolution-ai-isnt-just-for-the-big-guys-anymore/
[17] Wilson, H. J., & Daugherty, P. R. (2020, February 17). Small Data Can Play a Big Role in AI. Retrieved from https://hbr.org/2020/02/small-data-can-play-a-big-role-in-ai
[18] Amodei, D., & Hernandez, D. (2018, May 16). AI and Compute. Retrieved from https://openai.com/blog/ai-and-compute/
[19] Sastry, G., Clark, J., Brockman, G., & Sutskever, I. (2019, November 07). AI and Compute. Retrieved from https://openai.com/blog/ai-and-compute/
[20] India doubles its AI workforce in 2019, but faces talent shortage: Great Learning. (2019, December 27). Retrieved from https://economictimes.indiatimes.com/tech/ites/india-doubles-its-ai-workforce-in-2019-but-faces-talent-shortage-great-learning/articleshow/72997071.cms
[21] Aneja, U., & Chamuah, A. (2020, July 28). A Balancing Act - The Promise and Peril of Big Tech in India. Retrieved from https://tandemresearch.org/assets/Tandem-Research-Big_Tech_report.pdf
[22] Ibid.

[23] Ibid.

[24] Gailmard, S. (2009). Multiple Principals and Oversight of Bureaucratic Policy-Making. *Journal of Theoretical Politics*, *21*(2), 161–186. https://doi.org/10.1177/0951629808100762

[25] Murray, D. (n.d.). Live facial recognition and human rights: University of Essex. Retrieved from https://www.essex.ac.uk/centres-and-institutes/public-engagement/facial-recognition

[26] El Emam, K., & Hintze, M. (2019, January 29). Does anonymization or de-identification require consent under the GDPR? Retrieved from https://iapp.org/news/a/does-anonymization-or-de-identification-require-consent-under-the-gdpr/

[27] Bailey, R., Parsheera, S., Rahman, F., & Sane, R. (2018). Disclosures in Privacy Policies: Does ' Notice and Consent' Work?; Sinha, A., & Mason, S. (2016, January 11). A Critique of Consent in Information Privacy. Retrieved from https://cis-india.org/internet-governance/blog/a-critique-of-consent-in-information-privacy