BJHS Themes (2023), **8**, 205–220
doi:10.1017/bjt.2023.11

RESEARCH ARTICLE

Biometric data's colonial imaginaries continue in Aadhaar's minimal data

Sananda Sahoo

Faculty of Information and Media Studies, The University of Western Ontario, London, Ontario, Canada
Email: Ssahoo3@uwo.ca

Abstract

This paper considers three moments in the treatment of data about race and identity in India. Many elements go into the development of data imaginaries as these change over time. A complete history is beyond the scope of this paper, but I develop three key episodes to explore critical but changing features of interrelations between race, identity and statistical arguments historically. One aim is to explore key features of the argument developed by two significant individuals – Thomas Nelson Annadale and P.C. Mahalanobis – as they sought to develop databases that could answer questions about race formation and, in the case of Mahalanobis, might also be used to develop statistical methods on the one hand and aid governance on the other hand. A second aim is to use this historically based but highly selective investigation to uncover key features of the ideology with which the government of India has presented Aadhaar, its vast biometric identification system powered by authentication technologies afforded by artificial intelligence. This enables me to identify different forms of racial or ethnic identity that could be – and in one or two cases actually have been – implicated in the way Aadhaar has been used in practice.

Histories of artificial intelligence are intertwined with the history of data imaginaries. That is because narratives of AI centred on efficient, accurate and speedy data interpolation enable new understandings of how data can and should be used. This paper will focus on the various assumptions associated with biometric databases and data analysis methods in India that, in their present form, promote narratives of trust and empowerment associated with the database of individual identity numbers established by Aadhaar, now the principal form of identity in the country, and increasingly associated with diverse service provisions. The data analysis methods referred to in this paper involve physical measurements (as in anthropometric studies), statistics and AI-based fingerprint authentication (as in Aadhaar). AI in the Aadhaar ecosystem also comes into play whenever other government departments and agencies collate their databases with Aadhaar and authenticate individual identities.

The government of India reiterates that the Aadhaar system captures only what it calls minimum demographic data besides biometric data to provide only functions of issuance and authentication related to identity verification. It frames Aadhaar's design philosophy in the language of respect for residents' privacy that obligates it not to hold or receive non-essential data, such as religion, caste, ethnicity and geography. While retaining name, address, gender, date of birth and biometrics – that is, ten fingerprints and an

iris scan, along with a photograph – the central Aadhaar database does not link to existing systems or applications that use Aadhaar. The government's appeal for citizen confidence is based on the assurances that minimal data are retained and that authentication of such data is AI-based, and on the formation of what one could describe as data islands; that is, data stores with no or limited external linkages. These assurances promote a vision that Aadhaar's minimal identity data, databases and AI-based data analysis methods can confer a foolproof identity effectively isolated from social values. This paper will argue that such a view obfuscates the thick history of biometric data in India. It will show that biometric databases and their methods of analysis, specifically anthropometry and statistics, were initially engaged with studies of race and identity, and although ideals of statistical objectivity have both asserted and unlinked different forms of social and biological identification, biometric data, databases and data analysis methods have always carried social values and have been based on assumptions to help perform certain functions where the individual has little control over the usage of one's own biometric data.

The rapid distributed and networked introduction of Aadhaar in daily public and private-sector services has received much attention from scholars, especially for its implications for questions of state governance and identity. This attention, however, has primarily been through ethnographic studies – as one sees, for example, in the valuable articles on Aadhaar in a special issue of the journal *South Asia: JSAS* in 2019.[1] In the rare cases in which the introduction of Aadhaar has attracted historical attention, this has primarily been in terms of understanding it as a new form of identity among existing residency authentication documents such as the multi-purpose national identity card, or as generating new conceptions of society.[2] This paper complements the concerns of these primarily ethnographic studies by developing a strategic historical account that links individual identity in the Aadhaar ecosystem to earlier biometric and statistical studies – which were used to understand peoples and races rather than identity and citizenship, but which also approached questions of governance, sometimes obliquely. Such questions have often been studied through the lens of changing understandings of objectivity, such as through the work of Lorraine Daston and Theodore Porter, amongst others, and this remains an important approach. However, the form in which they have been developed and marketed in Aadhaar puts a new emphasis on databases, both for the state and for entrepreneurial start-up commercial projects that use some form of identification while providing citizen services.[3] Complementing a history of objectivity that examines contexts of observation, I suggest here that a history of what I call 'data imaginaries' can successfully account for colonial as well as Aadhaar's biometric data in the context of identity construction, and help tease apart some of the subtle implications that Aadhaar has held for the politics of race and identity.

An 'imaginary' is a mental concept or thought, an animating ideal – both more intimate and less specific than a recipe or protocol – that is projected onto the experiences around us, to sort and make sense of them. An imaginary is based on reality but is not controlled by it. Hence the imaginary varies depending on who is doing the imagining, when it is being done, and how that imagining sorts out experiences and motivations for doing something. As a result, there is no one imaginary associated with a

1 Ursula Rao and Vijayanka Nair, 'Aadhaar: governing with biometrics', *South Asia: Journal of South Asian Studies* (4 May 2019) 45(3), pp. 469–81.

2 Itty Abraham, 'Prehistory of Aadhaar: body, law, and technology as postcolonial assemblage', *East Asian Science, Technology and Society: An International Journal* (1 December 2018) 12(4), pp. 377–92; Bidisha Chaudhuri and Lion König, 'The Aadhaar scheme: a cornerstone of a new citizenship regime in India?', *Contemporary South Asia* (June 2018) 26(2), pp. 127–42.

3 Kavita Dattani, '"Governtrepreneurism' for good governance: the case of Aadhaar and the India Stack', *Area* (2020) 52(2), pp. 411–19; Ranjit Singh, 'Give me a database and I will raise the nation-state', *South Asia: Journal of South Asian Studies* (4 May 2019) 42(3), pp. 501–18.

https://doi.org/10.1017/bjt.2023.11 Published online by Cambridge University Press

phenomenon. Scholars have drawn on different features in considering imaginaries associated with technological innovations and projects. Paul Dourish and Genevieve Bell call ubiquitous computing 'technological imaginaries', arguing they invite 'new sorts of speculation about what information technology might and could be'.[4] For Sheila Jasanoff, collectively imagined forms of social life and order as reflected in technological or scientific projects are 'sociotechnical imaginaries', the plurality indicating their existence in 'tension or in a productive dialectical relationship'.[5] Ben Williamson uses the term 'big data imaginaries' to discuss the ways data are interpreted and their social implications.[6] danah boyd defines a statistical imaginary as a collective vision 'of what data are and what they could be'.[7] My use of the term also draws on ethnographers' articulation of the social imaginaries implied by Aadhaar, and their discussion of the way the Indian nation has been configured as a database through its introduction. For instance, Lawrence Cohen discusses the design of Aadhaar in the context of what he calls 'the social-yet-to-come', a promised network of deliverables that necessitates the production of 'a political subject outside of biography'.[8] It identifies that the individual as a member of the database is a unique individual and yet devoid of the typical social markers such as caste or religion. Ursula Rao identifies the tension with an Aadhaar individual, who is at once unique and has a specific status.[9]

By data imaginaries, I refer to what data are imagined to be, and then what they can do. For example, when the Unique Identification Authority of India (UIDAI) talks about all the benefits of using Aadhaar, such as authenticating an identity from anywhere in the country, it is producing a vision.[10] As boyd says, imaginaries can be 'grounded in practice, rooted in pragmatic goals, and realized through technical systems', but can also come unmoored from reality.[11] Aadhaar's database is a useful tool to validate identity to access benefits and services and eliminate fake accounts, but to suggest that its database has the 'tremendous potential to bring transformation as it empowers people in numerous ways' while preserving the privacy of data is an exaggeration.[12] Media reports suggest that data breaches have occurred and Aadhaar's empowerment of people is not uniform across the social hierarchy, with poor and other marginalized communities such as the Dalits suffering as a result.[13]

4 Paul Dourish and Genevieve Bell, *Divining a Digital Future: Mess and Mythology in Ubiquitous Computing*, Cambridge, MA: MIT Press, 2011, p. 161.

5 Sheila Jasanoff, 'Future imperfect: science, technology, and the imaginations of modernity', in Sheila Jasanoff and Sang-Hyun Kim (eds.), *Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power*, Chicago: The University of Chicago Press, 2015, pp. 1–33, 4.

6 Ben Williamson, 'Smarter learning software: education and the big data imaginary', Faculty of Social Sciences Conference Papers and Proceedings, Big Data – Social Data, Warwick, UK (2015), at http://dspace.stir.ac.uk/handle/1893/22743, p. 2.

7 danah boyd, 'Statistical imaginaries', Substack newsletter, *Data: Made Not Found (by danah)* (blog), 1 December 2021, at https://zephoria.substack.com/p/statistical-imaginaries.

8 Lawrence Cohen, 'The "social" de-duplicated: on the Aadhaar platform and the engineering of service', *South Asia: Journal of South Asian Studies* (4 May 2019) 42(3), pp. 482–500, 483, 500.

9 Ursula Rao, 'Population meets database: aligning personal, documentary and digital identity in Aadhaar-enabled India', *South Asia: Journal of South Asian Studies* (4 May 2019) 42(3), pp. 537–53.

10 UIDAI, *What Are the Features and Benefits of Aadhaar?*, Unique Identification Authority of India, Government of India, 9 June 2023, at https://uidai.gov.in/en/286-faqs/your-aadhaar/aadhaar-features,-eligibility/1934-what-are-the-features-and-benefits-of-aadhaar.html.

11 boyd, op. cit. (7).

12 UIDAI, *Annual Report 2021–22*, New Delhi, 2023, p. 1.

13 Hemant Gairola, 'How an Aadhaar fraud forces the poor into paying for welfare schemes they do not want', Article 14 (22 February 2023), at https://article-14.com/post/how-an-aadhaar-fraud-forces-the-poor-into-paying-for-welfare-schemes-they-do-not-want-63f57eb9e8d15; Elizabeth Donger and Ayesha Mehrotra, 'Aadhar and child protection in India: access for the poorest remains elusive', FXB Center for Health & Human Rights, Harvard University (1 May 2017), at https://fxb.harvard.edu/2017/05/01/aadhar-and-child-protection-in-india-access-for-the-poorest-remains-elusive; Chandan Shantaram Haygunde, 'Bhima Koregaon violence: police probe fake

https://doi.org/10.1017/bjt.2023.11 Published online by Cambridge University Press

Many elements go into the development of data imaginaries as these change over time. A complete history of them is beyond the scope of this paper, but I develop three key episodes to explore critical but changing features of interrelations between race, identity and statistical arguments historically. One aim is to explore key features of the argument developed by two significant individuals as they sought to develop databases that could answer questions about racial formation, and, in the case of Mahalanobis, might also be used to develop statistical methods on the one hand or aid governance on the other hand. A second aim is to use this historically based but highly selective investigation to uncover key features of the ideology with which Aadhaar has been presented by the government and offered as a key element in service provisions (they present it as being shorn of just these kinds of social information), and the different forms of racial or ethnic identity that could be – and in one or two cases actually have been – implicated in the way it has been used in practice.

This paper attempts to contribute to the literature on data imaginaries by historicizing biometric data imaginaries, showing how changing notions of objectivity associated with data and data analysis methods enable significant understanding but may also hide critical gaps in arguments relating race, identity and statistical methods. To do so, the paper considers bureaucrat Thomas Nelson Annandale's Anglo-Indian data set based on anthropometric data collected in 1916 in Calcutta, statistician Prashanta Chandra Mahalanobis's statistical treatment of the same data set in a series of papers published between 1922 and 1936, and biometric data collected since 2010 as part of India's Aadhaar identification project.

Human data during colonial rule

The first two sections of this paper will show that Annandale's and Mahalanobis's work incorporated many racial and caste assumptions of the period, underscoring the fact that treatments of biometric data and data sets typically present significant relations between individual scientists and their social context because they articulate and sometimes challenge the social order of the period. Examining their work will also show important respects in which understandings of data changed, in particular concerning scientists' readiness to trust different kinds of measurement and the number of cases required to draw reliable inferences about different groups of people. In keeping with the interests of contemporary race scientists such as Francis Galton and Karl Pearson, Annandale turned to anthropometry to study the evolution of races through fusion.[14] For instance, in the 1892 monograph *Finger Prints*, Galton pursued questions of personal description and identification with race in mind. He alternated between dejection that fingerprints failed to assign racial categories and his vision that racial typologies exist.[15] In *Decipherment of Blurred Finger Prints* (1893), Galton professed to study indistinct impressions from a group of men in India. Having failed to empirically establish fingerprinting as a method to typify races, his racial prejudices link the indistinct fingerprints to generic characteristics of the Indian race.[16] Fingerprints as a tool of individual identity developed from their ancient origins in 300 BC in China, to the increasingly precise study of the

aadhaar card with Delhi address', *Indian Express*, 24 January 2018, at https://indianexpress.com/article/india/bhima-koregaon-police-probe-fake-aadhaar-card-with-delhi-address-5036742.

14 Karl Pearson, '(I.) Editorial: the scope of Biometrika', *Biometrika* (1901) 1(1), pp. 1–2; Pearson, '(II.) Editorial: the spirit of Biometrika', *Biometrika* (1901) 1(1), pp. 3–6.

15 Francis Galton, *Finger Prints*, London: Macmillan and Co., 1892, p. 193; S.M. Stigler, 'Galton and identification by fingerprints', *Genetics* (July 1995) 140(3), pp. 857–60.

16 Francis Galton, *Supplementary Chapter to 'Finger Prints': Decipherment of Blurred Finger Prints*, London: Macmillan and Co., 1893.

https://doi.org/10.1017/bjt.2023.11 Published online by Cambridge University Press

ridges from the 1800s onwards in Europe, as well as in post-independence India, for example, with Salil Kumar Chatterjee, which led governments to see it as a trusted source of individualization. Not surprisingly, identity documents such as voter ID cards and passports in India allow fingerprints in lieu of signature. One can see here the clear links with Aadhaar's reliance on fingerprint as a key identifier.

Annandale's view of anthropometric data is situated within these broader contemporary assumptions associated with anthropometric data sets and anthropometry's goals without pursuing questions of individual identification. While still an undergraduate student at Oxford's Balliol College, Annandale started his first anthropometric measurements on Faroe Island and Iceland in 1896 in the context of a general ethnological study. Annandale chose the two islands due to the isolation of their people. Based on the data collected, he concluded that the residents of the two islands evolved differently, including in physical appearance, despite similar ancestry and geographic environment. In 1903, as a research fellow in anthropology at Edinburgh University, Annandale co-authored *Fasciculi Malayenses: Anthropological and Zoological Results of an Expedition to Perak and the Siamese Malay states, 1901–1902*. He noted similarities and divergences in physical appearance – regardless of similarities in geography and climate – within and among various population groups that were then categorized into different 'races' living in proximity in the Malay peninsula. Here, Annandale's data imaginary assumes that human anthropometric data can categorize the human population into races and trace the mixture of races based on morphological measurements. Looking back on these studies in 1922, however, Annandale voiced new doubts about the possibility of answering larger questions about differences between 'races' given Pearson's argument that a 'mixed race' would show a more significant variation in physical measurements as a result of interbreeding across different races than purer races that have not interbred.[17] His Malay peninsula data failed to show the differences he expected in head measurements between what he called 'Negrito' jungle tribes and 'Indonesian' jungle tribes – even though data from the same region showed a major difference in head measurements between 'civilized' and 'uncivilized' tribes that he did anticipate.[18] For Annandale, it was frustrating that anthropometric data from his previous studies did not provide conclusive answers about different pathways in the evolution of people.

Annandale's 1922 observations were informed by the anthropometric data he began collecting in 1916, some ten years after becoming director of the Indian Museum and after taking over responsibility for the newly formed Zoological Survey of India (a position he held until his death in 1924). In this case, he collected data from a sample of two hundred individuals belonging to the Anglo-Indian community in Calcutta. Variously regarded as a group or race, the 1911 census officials first classified this community as a separate 'race', which numbered approximately 101,657 in India at that time.[19] In this census, the officials defined 'Anglo-Indians' as people of mixed European and Asiatic descent, a term intended to replace 'Eurasians'. The Anglo-Indian community was a subject of

17 Karl Pearson, 'Craniological notes: remarks on Dr. C.S. Myers' note', *Biometrika* (1903) 2(4), pp. 506–8.

18 Nelson Annandale, 'Introduction' to P.C. Mahalanobis, 'Anthropological observations on the Anglo-Indians of Calcutta. Part 1: analysis of male stature', *Records of the Indian Museum* (April 1922) 23(1), pp. 1–4, at https://archive.org/details/dli.zoological.records.023.01/mode/2up; P.C. Mahalanobis, 'Anthropological observations on the Anglo-Indians of Calcutta. Part 1: analysis of male stature', *Records of the Indian Museum* (April 1922) 23(1), pp. 5–94, at https://archive.org/details/dli.zoological.records.023.01/mode/2up. By 'Indonesian jungle tribes', Annandale most probably refers to another Malay peninsula Negrito tribe, the Sakai, related to the 'Indonesian' tribes of French Indo-China, as suggested by a later anthropometrist, Ivor H.N. Evans, in his book *The Negritos of Malaya*, New York: Routledge, 1937, p. 44.

19 E.A. Gait, *Census of India, 1911. Part 2: Tables*, vol. 1, Calcutta: Superintendent of Government Printing, 1913, p. 378.

https://doi.org/10.1017/bjt.2023.11 Published online by Cambridge University Press

particular interest to Annandale because, in them, he saw a recent mix of European and Indian blood through which he could examine race formation, and ultimately the origin of 'human races by fusion'.[20] However, the anthropometric method to study race formation based on physical measurements led to an intellectual dead end that failed to explain for Annandale the relation between the origins and the recent admixture of races. Scientists and statisticians furthered the essentialist idea of race by treating human data, such as morphological data and later genomic data, to validate race as an analytical frame to study human evolution.[21] Such 'scientific racism', as Lisa Gannett calls it, is used to legitimize human data to carry out projects such as Annandale's usage of anthropometric data to study the formation of races and Mahalanobis's treatment of similar data in his biometric papers to question the colonial hierarchy of races that eventually furthered the prevailing notions of caste hierarchy.[22] The notion of data imaginary in these contexts is mediated through the notion of objectivity. With Annandale, the assumption is that human morphological data can be bestowed with a mechanical objectivity, and can produce valid answers about the history of human races and relations among them when framed in what Elazar Barkan calls 'ostensibly scientific terms'.[23] As we shall see later, with Mahalanobis, this objectivity of data relates to statistical objectivity associated with the nature of scientific claims that data can make, the procedure of data analysis, and the qualification of the researcher as to their objective intentions.[24] Mahalanobis shifted focus from the typological concept of race that animated Annandale's database to population-based accounts afforded by larger databases mediated through the notion of statistical objectivity. While Annandale did not use the term 'objectivity', Mahalanobis used it in terms of a value, as a principle of scientific validity, where the results of the statistical analysis are open to 'objective checks'.[25] This sense of statistical objectivity associated with the data analysis method and large volumes of collective data about a population group also drives the project of Aadhaar, evident in the government's consistent assertions that its minimal database does not amount to profiling. In response to a question in the Lok Sabha of the Parliament in 2016, the minister of electronics and information technology said that Aadhaar's demographic information is not linked to information on caste because caste data are not collected during the process, even though citizens can furnish caste or domicile certificate as one of the documents to prove address.[26]

Between 1916 and 1919, Annandale and his collaborators collected data on stature, head length, head breadth, nasal length, nasal breadth, zygomatic or face breadth and

20 Annandale, op. cit. (18), p. 4.

21 Lisa Gannett, 'Racism and human genome diversity research: the ethical limits of "population thinking"', *Philosophy of Science* (2001) 68(S3), pp. S479–92; Jenny Reardon, 'Decoding race and human difference in a genomic age', *Differences: A Journal of Feminist Cultural Studies* (2004) 15(3), pp. 38–65; Veronika Lipphardt, 'Traditions and innovations: visualizations of human variation, *c.*1900–38', *History of the Human Sciences* (1 December 2015) 28(5), pp. 49–79.

22 Gannett, op. cit. (21), p. S485.

23 Elazar Barkan, 'Race and the social sciences' , in Theodore M. Porter and Dorothy Ross (eds.), *The Cambridge History of Science*, 1st edn, vol. 7, Cambridge: Cambridge University Press, 2003, pp. 693–707, 693.

24 Lorraine Daston, 'How probabilities came to be objective and subjective', *Historia Mathematica* (1 August 1994) 21(3), pp. 330–44; Daston, 'Objectivity and the escape from perspective', *Social Studies of Science* (1 November 1992) 22(4), pp. 597–618, ; Daston, 'The ideal and reality of the republic of letters in the enlightenment', *Science in Context* (1991) 4(2), pp. 367–86; Theodore M. Porter, *The Rise of Statistical Thinking, 1820–1900*, Princeton, NJ: Princeton University Press, 1986; Porter, *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*, Princeton, NJ: Princeton University Press, 1995.

25 P.C. Mahalanobis, 'Analysis of race-mixture in Bengal', *Journal and Proceedings of Asiatic Society of Bengal* (new series) (1927) 23(3), pp. 301–33, 323.

26 Ninong Ering, 'Unstarred question no. 4460: linking of documents with Aadhaar', Lok Sabha, 14 December 2016, at https://eparlib.nic.in/bitstream/123456789/693568/1/45344.pdf.

https://doi.org/10.1017/bjt.2023.11 Published online by Cambridge University Press

upper face length of two hundred Anglo-Indians in Calcutta. Publishing in the journal he founded to record the work of the museum, Annandale aimed to argue that the Eurasian population of contemporary Calcutta resulted from a recent admixture of races and, based on that data, shed some light on 'the origin of human races through fusion'.[27] However, he came across the same problem of unsatisfactory measurements as in his earlier data sets. Annandale said in his 1922 introduction to Mahalanobis's paper, 'Anthropological observations on the Anglo-Indians of Calcutta, part 1: analysis of male stature', that the measurements were largely unsatisfactory – so much so that he had decided to reject them. While he did not explain why the data were unsatisfactory to him, he did say that he was 'fortunate' to get Mahalanobis's offer to study the data statistically.[28]

The data set and method of analysis also reflected particular social assumptions that Annandale referred to only obliquely. Annandale based his decision about whom to include in his measurements on visual inspection, choosing individuals who did not look as if they had any recent admixture of Negro or Mongoloid blood – that is, within the last two generations. He also took the self-declared rigid maintenance of social distinction within this group as a valid reason to assume that they had not interbred with Negro or Mongoloid races recently. However, he acknowledged there were likely to be a few outliers inadvertently included from these different races, and 'probably in many individuals' we might see the presence of other blood, but '[a]s to old Negro blood, no definite information was obtained'.[29] Annandale said it was important to eliminate 'Negro' or 'Mongoloid' blood to avoid the complexity of including more variables in the data when he did not have the mathematical numbers for proper treatment. Yet, while Annandale acknowledged these uncertainties regarding the homogeneity of the data, he did not explain what made the data set valid, especially when it was as small as two hundred people.

Despite these doubts over its homogeneity and lack of a standard head measure for this group, he still thought the Anglo-Indian data set would be able to answer significant questions about race fusion. His understanding of these data was based on the social imaginary that the Anglo-Indians of the data set strictly adhered to social distinctions that prevented them from intermixing with Negroid, Mongoloid or other population groups. Such an imaginary is carried forward through stories and anecdotes, such as what Herbert A. Stark, the principal of the Armenian College, Calcutta, himself a prominent member of the local Anglo-Indian community, told Annandale about the rigid social distinctions maintained by Anglo-Indians.[30] These social distinctions presumably rooted Anglo-Indians as an educated, middle-class community that would not intermarry with lower classes such as formerly enslaved people or Kintalis.

Annandale did not explain race as a category for his study and did not state which so-called 'pre-existing' races he was allowing into the sample under investigation. However, it is evident that he meant a broad category of Indian and European population groups. He did not define the Indian race. Geographic classification of population groups thus got entangled with racial classification, which primarily relies on Pearson's anthropometric studies. This idea of race assumed the existence of a 'pure' race and that bodily measurements could point to a racial identity. The data imaginary is, therefore, also based on the racial imaginary that people can be separated into races according to morphological features, and that these features form the basis and existence of a pure race,

27 Annandale, op. cit. (18), p. 4.

28 Annandale, op. cit. (18), p. 4.

29 Annandale, op. cit. (18), p. 3.

30 C.J. Hawes, *Poor Relations: The Making of a Eurasian Community in British India, 1773–1833*, Richmond: Curzon Press, 1996); Annandale, op. cit. (18).

https://doi.org/10.1017/bjt.2023.11 Published online by Cambridge University Press

and on the assumption that pure racial types exist and that fusion of such races can be studied through physical features. With Annandale, we see early efforts to relate anthropometric data to one's larger group identity though emphasizing racial identity through such features as head measurements. When situated in the context of anthropometric studies such as that of Herbert Hope Risley's census of the Indian population in 1901, the aims of Annandale's project to categorize the Indian population get aligned with that of the colonial bureaucratic state. The historian C.J. Fuller has noted that the importance of anthropometric projects diminished vis-à-vis the colonial state from the early 1900s as these failed to provide concrete strategies to counter the rising tide of nationalism.[31] Nonetheless, these projects show a similarity with the Aadhaar system in their data imaginaries of biometric data's ability to construct an identity for larger purposes associated with large population groups.

Human data with Mahalanobis

P.C. Mahalanobis, a student of mathematics and later physics at Cambridge University for two years until 1915, came across the fields of statistics and anthropometry through Karl Pearson's *Biometrika*, and was influenced by Pearson's use of statistics.[32] Pearson himself was influenced by Galton's usage of the term 'biometry' to mean the application of statistical methods to the analysis of biological variation.[33] In 1921, Mahalanobis met Annandale at the Nagpur session of the Indian Science Congress, where he asked Annandale if he could use the Anglo-Indian data set. Mahalanobis, however, used biometric studies not to make claims of racial superiority but rather to claim the ground for statistics and legitimize sampling and grouping in such a way as to make it possible to study populations statistically.[34] His treatment of the Anglo-Indian data project aimed to show the importance of the application of 'accurate statistical methods' to the 'crude' anthropological measurements for a scientific study of data where facts can be objectively studied and interpreted.[35] Mahalanobis exhibited a move to non-essentialize a racialized data set and the struggle to shear off sociological, anthropological and even historical claims attached to it, in keeping with the prevailing belief that the statistical method was value-free. As I will show, Annandale's anthropometric data provided an opportunity to demonstrate the need for accuracy in data to support any subsequent conclusions from the data set; the need to follow rigid standards among data categories, in the process revealing the biases that inevitably creep into data sets; and the caution that needs to be exercised while coming to conclusions from the data. At the same time, as he stated in the 1922 paper published in the *Records of the Indian Museum* on the same biometric data set, one way to ensure a that data set is fairly representative and heterogeneous statistically was to work with massive amounts of data.[36] Here again, the assumption in operation is that large volumes and 'correct' data can be revelatory, and that correct statistical group analysis can solve problems. Correct data for him meant data that are collected and recorded error-free. Mahalanobis argued that ideal data sets are those with error-free data and greater data points to facilitate more correlations and varied interpretations. In his early biometric studies, the problem he tried to solve was whether one could be

31 C.J. Fuller, 'Colonial anthropology and the decline of the Raj: caste, religion and political change in India in the early twentieth century', *Journal of the Royal Asiatic Society* (2016) 26(3), pp. 463–86.

32 Ashok Rudra, *Prasanta Chandra Mahalanobis: A Biography*, New Delhi: Oxford University Press, 1996.

33 Robert R. Sokal and F. James Rohlf, *Biometry: The Principles and Practice of Statistics in Biological Research*, 2nd edn, San Francisco: W.H. Freeman, 1981, p. 4.

34 Prabhat Patnaik, 'The Nehru–Mahalanobis strategy', *Social Scientist* (New Delhi) (2015) 43(3–4), pp. 3–10.

35 Mahalanobis, op. cit. (18), p. 7.

36 Mahalanobis, op. cit. (18), p. 33.

https://doi.org/10.1017/bjt.2023.11 Published online by Cambridge University Press

definite about the mechanism of race formation.[37] Between 1922 and 1936, he wrote fifteen papers based on anthropometric and biometric data from Annandale and Risley and his data on Indian population groups that led to the formulation of the generalized distance formula in 1936. Known as the Mahalanobis distance, it is now widely used in AI-powered tools for pattern recognition.

Dealing with Annandale's incomplete and messy data on the Anglo-Indian community, Mahalanobis clarified that determining variability in Anglo-Indians' morphological features constituted a statistical study for him. In mathematically accounting for biases and contradictions in data – that is, observed facts, in this case – Mahalanobis was closely following the contemporary and historical enthusiasm surrounding the statistical method that assumed that rigorous and explainable quantification methods rendered data oblivious to its social processes.[38] He concluded that 'Anglo-Indian Variability is just about the same as the Variability of European (in a geographical sense only) races.'[39] He corrected the errors in recording the measurements by imposing statistical standards, and made these valid for biometric study. Mahalanobis concluded that Annandale's Anglo-Indian data set was statistically homogeneous.[40] At the same time, he was aware that '[o]ur Anglo-Indian data cannot be assumed to be homogenous in character'.[41] In his view, a lack of homogeneity colours all data sets in the face of human variability, and the only way to study it analytically is to impose a standard that will 'smooth out' the discrepancies in data.[42] Following the logic of statistics, this standard is at once 'arbitrary and conventional', but, once set, such standards must be rigidly followed to enable correct analysis.[43] Even though claims to objectivity and accuracy were made in earlier anthropometric studies, the results were often clouded by the researchers' imaginaries of British racial superiority, as was evident in the writings of Galton and Pearson.[44] Mahalanobis did not partake in this specific imaginary, maintaining that his biometric studies of race were a statistical endeavour. Unlike the studies of Annandale or Galton, Pearson and Fisher, Mahalanobis refused to make anthropological deductions from the statistical analysis, instead drawing disciplinary boundaries.[45] However, he did use race as a category, and did not question the hierarchy among castes. He also explicitly used and revealed his acceptance of a binary standard for civilized races, and hence a hierarchy among races.[46]

By naturalizing the most frequently occurring association among data – or the dominant relationship, establishing these associations as facts, and flattening the variability of data from individuals to make correlations in large data sets, the data imaginary based on the statistical method typified the dominant pattern in the data sample. Once established as the legitimate relationship, this typical formula was then applied to the entire population through 'logical induction', thereby naturalizing the relationship through logic.[47] Mahalanobis applied this principle to the Anglo-Indian data set and to his later biometric

37 Mahalanobis, op. cit. (18).

38 Victor L. Hilts, 'Aliis exterendum, or, the origins of the Statistical Society of London', *Isis* (1978) 69(1), pp. 21–46; Porter, opera cit. (24).

39 Mahalanobis, op. cit. (18), p. 63.

40 Mahalanobis, op. cit. (18), p. 89.

41 Mahalanobis, op. cit. (18), p. 8.

42 Mahalanobis, op. cit. (18), p. 10.

43 Mahalanobis, op. cit. (18), p. 31.

44 Galton, op. cit. (15); Galton, op. cit. (16); Darcie A.P. Delzell and Cathy D. Poliak, 'Karl Pearson and eugenics: personal opinions and scientific rigor', *Science and Engineering Ethics* (2013) 19(3), pp. 1057–70.

45 Mahalanobis, op. cit. (25), p. 323.

46 Mahalanobis, op. cit. (18), p. 67.

47 Mahalanobis, op. cit. (18), p. 10.

https://doi.org/10.1017/bjt.2023.11 Published online by Cambridge University Press

studies. In his first paper in 1922, he concluded that the variability among Anglo-Indians in the data set is greater than that shown by Indian castes but is close to that of modern European races. In a second paper on the same data set in 1927, Mahalanobis showed that, contrary to Annandale's belief, the Anglo-Indian population group in the data set had a close resemblance to higher Hindu castes.[48] In a 1931 paper, he concluded that the variability in the head length of Anglo-Indians is more significant than in other caste groups and showed that they are a recent intermixture.[49]

Despite his claims of disciplinary boundaries, Mahalanobis did continue to study the Anglo-Indian data set vis-à-vis Risley's caste data. In his 'Analysis of race-mixture in Bengal' (published in 1927), Mahalanobis sought to 'provide statistical answers to anthropological questions such as: do the Anglo-Indians show a greater affinity with the higher castes of Bengal or with the lower castes, and is there any appreciable admixture with the aboriginal tribes?'[50] He did not explain in his 1922 paper what made him compare the Anglo-Indian data with that of Risley's caste data, where Risley assumed that the Indian population group comprised several races and that castes were sites of manifestation of the racial stocks.[51]

Mahalanobis went to great lengths to establish the legitimacy of the Anglo-Indian data, explaining factors such as the effect of grouping and the problem of establishing a type and homogeneity despite inherent variability. While he said that such efforts were necessary for the Anglo-Indian data set, given its messiness, he added that such explanations were needed for any data set under study. This emphasis on the explainability of the method, which has gained prominence in current critical AI studies, is used to legitimize data and the method.

In 'Analysis of race-mixture in Bengal', Mahalanobis argued that broad classifications can win over difficulties of 'indefiniteness' of whom to include in a category, especially for cultural categories that Mahalanobis himself admitted are complicated because of the absence of a 'regular hierarchy of social order [in the Hindu community] in which every caste can be placed in a definite intermediate position between any two other castes'.[52] In his view, broad categories also to a great extent solve the problem of how to 'compare and fix the relative position of castes belonging to different provinces'.[53] Despite this awareness, Mahalanobis proceeded to classify caste based on the assumption of homogeneity of its composition on the one hand while accepting prevailing popular assumptions relating to caste hierarchy on the other and statistically treating the Muslim population as a separate category within the Indian population. In this, he showed the contemporary biases that Indian researchers faced in working with the analytical frames and methodologies borrowed from Western epistemologies, such as the racial framework applied to caste and tribes and statistics applied to the studies of the Indian population.[54] As dichotomization between Hindus and Muslims was institutionalized during the colonized years, Muslims were categorized as a separate electorate with the

48 Mahalanobis, op. cit. (25), pp. 321–2.

49 P.C. Mahalanobis, 'Anthropological observations on the Anglo-Indians of Calcutta. Part II: analysis of Anglo-Indian head length', *Records of the Indian Museum* (February 1931) 23, pp. 97–149.

50 Mahalanobis, op. cit. (25), p. 301.

51 C.J. Fuller, 'Ethnographic inquiry in colonial India: Herbert Risley, William Crooke, and the study of tribes and castes', *Journal of the Royal Anthropological Institute* (2017) 23(3), pp. 603–21; Projit Bihari Mukharji, 'Profiling the profiloscope: facialization of race technologies and the rise of biometric nationalism in inter-war British India', *History & Technology* (December 2015) 31(4), pp. 376–96.

52 Mahalanobis, op. cit. (25), p. 302.

53 Mahalanobis, op. cit. (25), pp. 301–3.

54 Thiago P. Barbosa, 'Indian sociology and anthropology between a decolonising quest and the west', *Revue d'histoire des sciences humaines* (15 December 2022) 41, pp. 181–211.

https://doi.org/10.1017/bjt.2023.11 Published online by Cambridge University Press

Morley–Minto reforms in 1909 and were given increasing representation based on religious identity after the 1921 Montagu–Chelmsford reforms.[55] In his 'Analysis of race-mixture in Bengal', Mahalanobis considered Risley's anthropometric data published in 1891 belonging to people from thirty typical castes of northern India.[56] Mahalanobis took Risley's data and categorized the data across six geographical divisions of northern India – Bengal, Chota Nagpur, Bihar, North-Western Provinces and Oudh, Punjab, and Eastern Districts – and five broad cultural strata – high castes, low castes, aboriginal tribes, Eastern tribes and Mohammedans.

Mahalanobis allowed Risley's usage of the value-prone data as such: 'We must conclude, therefore, that the real defect in Risley's data creeped in during the calculation of the average values, and that his primary data of individual measurements can be used with safety, especially after applying the corrections discussed in the present paper.'[57] This turned his study into what he was trying to avoid; that is, value-prone. By accepting Risley's data imaginaries based on their social and racial assumptions, he agreed with social categories such as the hierarchy of caste, the broad classification of tribes into eastern and aboriginal tribes, and Muslims in Bengal as a separate population group for the study. His findings reiterate the caste composition of Bengal as per the ethnic and social make-up in his time; in this respect, his non-critique of the data treated caste and religious segregation as an important given reality.

Mahalanobis's biometric studies thus furthered new data imaginaries where uncertain data in terms of homogeneity were transformed into representative facts through rigorous modes of enquiry. It also supported the narrative of data's ability to represent a wider population. Mahalanobis later used the sampling procedure, which he termed 'operational research', to address the needs of the population as part of the Planning Commission that he helped set up after India's independence in 1947. The sampling procedure fuelled development planning in India's post-independence years, responding to the needs of refugees, shortages of food and rising inflation. The current biometric data usage under the Aadhaar programme, too, is described as a part of the overall planning efforts to address the population's needs. The current government says,

> Aadhaar number will help the residents to avail various services provided by banking, mobile phone connections and other Govt and Non-Govt services in due course ... [It is] unique and robust enough to eliminate the large number of duplicate and fake identities in government and private databases. [It is] a random number generated, devoid of any classification based on caste, creed, religion and geography.[58]

The Aadhaar system relies upon data imaginaries that transform minimal biometric data into value-free data that must be owned by the government–corporate entanglement for the greatest interoperability among sectors such as welfare distribution and commerce. In their treatment of big data, Elena Aronova and colleagues point to the importance of situating data in the *longue durée*.[59] By historicizing the imaginary associated with biometric data sets, I next attempt to explore critical features of what changed and what remains

55 Fuller, op. cit. (31).

56 H.H. Risley, *The Tribes and Castes of Bengal: Ethnographic Glossary*, 4 vols., vol. 1, Calcutta, India: Bengal Secretariat Press, 1892.

57 P.C. Mahalanobis, 'A revision of Risley's anthropometric data relating to the tribes and castes of Bengal', *Sankhyā: The Indian Journal of Statistics* (June 1933) 1(1), pp. 76–105, 104.

58 UIDAI, 'What is Aadhaar?', Government of India, at www.uidai.gov.in/en/16-english-uk/aapka-aadhaar/14-what-is-aadhaar.html (accessed 7 June 2023).

59 Elena Aronova, Christine von Oertzen and David Sepkoski, 'Introduction: historicizing big data', *Osiris* (September 2017) 32(1), pp. 1–17.

https://doi.org/10.1017/bjt.2023.11 Published online by Cambridge University Press

unchanged with such data imaginaries in India. As Ursula Rao and Vijayanka Nair suggested, the assumptions about data and data sets made in Aadhaar help to depoliticize the state and how the government sees the population.[60]

Human data in the democratic state

After independence, the Planning Commission set up in 1950 was tasked with drawing up five-year plans to solve the basic needs of the new country, such as poverty, famines and agricultural output. In the First Five-Year Plan, data sampling aimed to address the problem of food shortage. Another response was to establish a public distribution system of food grain, a system first introduced after the Bengal famine of the 1940s and prioritized in the 1960s. Under the scheme, families were eligible for subsidized food grain based on income. While food grain reached millions of families in need, wastage was also common, partly due to fake accounts. By 1992, the World Bank estimated that wastage amounted to 35 per cent of the total amount of distributed grain.[61] The Aadhaar system was born out of the need to plug these leakages.

In 2009, the Indian government introduced the Aadhaar scheme, under which biometric and demographic data were collected to assign a unique identification number to residents. The definition of biometrics has also changed to mean the technology of identifying an individual based on a person's physical attributes.[62] The assumption is that iris and fingerprint patterns, along with facial recognition, are accurate to a high degree. The individual self is then established as equivalent to their physical attributes, conferring an empirically valid relationship between them. The ownership of one's data, however, at the same time is not restricted to the self but is extended to the state during the early days of Aadhaar and then increasingly to a state–corporate entity. Questions placed in the Lok Sabha in 2019 over the proposed changes to the Aadhaar Act voiced this concern over private players' usage of personal data belonging to citizens of the country, and the government's intention, as mentioned in the *Economic Survey 2018–2019*, to monetize the Aadhaar data.[63] As the subsequent amendments in 2019 show, any public or corporate entity such as telecom companies and banks can now authenticate Aadhaar numbers if an individual voluntarily uses their Aadhaar number to establish their identity. The amendments also introduce the Aadhaar ecosystem, which includes enrolling agencies, requesting agencies and offline verification-seeking entities that further make room for a private-sector role. The amendments also changed the language in the Aadhaar Act to say that instead of the core biometric data, no demographic information or photographs shall be published, displayed or posted publicly except for the purposes specified by regulations. The implication is that the core biometric data are free from similar compulsions of publication. The argument here is that fingerprints and iris scans shorn of caste, ethnicity, religion and geography are not profiling data; that residential identity does not lead to profiling; and that ownership of the biometric data must belong to a state–corporate entanglement for most efficient governance. The government's claim that biometric and residential data in the Aadhaar regime are isolated

60 Rao and Nair, op. cit. (1).

61 S. Guhan, 'Social security options for developing countries', *International Labour Review* (1994), 133(1), pp. 35–54.

62 Anil K. Jain, Patrick Flynn and Arun A. Ross, eds., *Handbook of Biometrics*, New York: Springer, 2008.

63 Lok Sabha, 'Seventeenth series, Vol. II, first session, 2019 written answers to questions: starred question nos. 188 to 191 and 193 to 200', 4 July 2019, pp. 140, 150–8, at https://eparlib.nic.in/bitstream/123456789/786385/1/lsd_17_01_04-07-2019.pdf; Ministry of Law and Justice, 'The Aadhaar and Other Laws (Amendment) Act, 2019', 2019, at https://egazette.gov.in/WriteReadData/2019/207897.pdf; Ministry of Finance, *Economic Survey 2018–19*, vol 1, New Delhi: Government of India, July 2019, https://www.im4change.org/docs/21134Economic_Survey_2018-19_Volume-1_Ministry_of_Finance.pdf, p. 94.

https://doi.org/10.1017/bjt.2023.11 Published online by Cambridge University Press

from the notion of profiling is contested in some north-eastern states in India, such as Tripura, where, as Nafis Hasan has shown, local mediations about immigrant groups reflect the questions of race and ethnicity that are raised in the work of Annandale and Mahalanobis.[64] Security concerns have led the government to push Aadhaar as a residency verification tool in some of these border states.[65] This, however, has inflamed old concerns of ethnic minorities who fear that illegal immigrants from across the border could gain legal residency through Aadhaar, and subsequent homogenization of state policies at the expense of the rights of native ethnic minorities, further linking the biometric database not only to issues of ethnicity but also to identity and citizenship.[66] Moreover, in 2018, cybersecurity researcher Srinivas Kodali tweeted that the digital dashboard of the Andhra Pradesh State Housing Corporation in India displayed the Aadhaar identification number of thousands of individuals.[67] The ID numbers were also accompanied by caste information, which is not collected by the UIDAI. It was evident that other government agencies that did collect such data could link it with the Aadhaar data for their purposes.

The increasing expansion of Aadhaar to encompass several sectors, such as security, banking and transport, underscores the political confidence in governing through population as a database, Itty Abraham points out.[68] As with Annandale and Mahalanobis, the assumption then involves a common belief about what human data sets can achieve regarding identity and, at times, their ability to answer significant questions about governance. If Aadhaar has a face, it would be that of technology entrepreneur Nandan Nilekani, who spearheaded its implementation from 2009 until 2014, when he resigned. It can be argued that data assumptions under Aadhaar are his, just as Mahalanobis's data assumptions regarding population sampling methods had roots in his anthropometric studies and can be regarded as his own. However, with Aadhaar, the data assumptions go back to 2006 with government approval for a unique ID for Below Poverty Line families. Since then, Aadhaar has undergone increasing expansion of services and technologies, the latest being the application of image-readable AI to authenticate fingerprints to rule out duplication of fingerprints in Aadhaar accounts.[69] Under the Aadhaar regime, human data can enable residential proof and access to services, target welfare distribution, minimize wastage, and eventually expand the role of private enterprises in the public sector, as the NITI Aayog's 2021–2 report recognized. In 2015 the NITI (National Institution for Transforming India) Aayog, also in charge of drawing up India's AI strategy, replaced the now defunct Planning Commission, which had similar functions. The NITI Aayog operates within the new paradigm of governance through data that the current government prioritizes.

After the current government came to power in 2014, it expanded the Aadhaar system to include every resident of the country. The Aadhaar number is now required to perform

64 Nafis Aziz Hasan, 'Bureaucratic mediations for biometric governance in India's Northeast: Aadhaar in Tripura', *South Asia: Journal of South Asian Studies* (4 May 2022) 45(3), pp. 560–76; Anuj Srivas, 'After lull, momentum gathers for frenzied Aadhaar push in India's Northeast', *The Wire*, 5 September 2017, at https://thewire.in/government/uidai-kicks-off-reinvigorated-aadhaar-push-indias-northeast.

65 Hasan, op. cit. (64); Srivas, op. cit. (64).

66 Ipsita Chakravarty, 'Aadhaar has run into pockets of resistance in three states of the North East', *Scroll.in*, 15 November 2017, at https://scroll.in/article/857074/aadhaar-has-opened-up-pockets-of-resistance-in-three-states-of-the-north-east.

67 Scroll.in staff writer, 'Government website leaked 1.3 lakh Aadhaar numbers, linked them with caste, religion: researcher', *Scroll.in*, 24 April 2018, at https://scroll.in/latest/876775/government-website-leaked-1-3-lakh-aadhaar-numbers-linked-them-with-caste-religion-researcher.

68 Abraham, op. cit. (2).

69 'UIDAI upgrades Aadhaar fingerprint authentication technology with artificial intelligence', *The Hindu*, 27 February 2023, at www.thehindu.com/news/national/uidai-upgrades-aadhaar-fingerprint-authentication-technology-with-artificial-intelligence/article66560107.ece.

https://doi.org/10.1017/bjt.2023.11 Published online by Cambridge University Press

everyday transactions, such as getting a telephone connection, even though the government maintains that Aadhaar data are not required to enroll in such services. The government narrative linking Aadhaar with various services is to prevent identity fraud and check money laundering. According to the 2021–2 UIDAI annual report, 315 schemes of the central government were linked with Aadhaar-based Direct Benefits Transfer.

The sense of planning implicit in the current use of the Aadhaar regime is based on the data imaginary that an enormous volume of biometric data from citizens must be collected for accurate residential identification, that digital identity must be authenticated in real time, and that these data now in government possession must be used and collated for future services for the distribution of benefits, subsidies and services. This idea of planning contrasts with Mahalanobis's sense of planning based on population sampling, without technologies to collect and analyse such enormous volumes of biometric data across almost the entire country – 92.8 per cent of the 1.4 billion population are enrolled in Aadhaar as of 31 March 2022 according to the 2021–2 UIDAI annual report. He believed that the standardization of data from population samples through statistical methods could formulate policies for the masses. At the heart of this difference in approach to planning is the difference in associated data imaginaries that dictate particular ways of using data and the state's role in governance. While during the early post-independence years, the state was the only player in welfare distribution, from data collection to data analysis, interpretation and usage, under the Aadhaar regime the state aspires to be an enabler rather than the sole player, making room for a larger role for the private sector, especially in data usage.[70] While for both the Aadhaar regime and Mahalanobis, data imaginaries are built around massive amounts of data, the difference lies in what data can enable. For Mahalanobis, data based on population samples enabled specific policies to solve large problems, such as unemployment or food shortages. In the Aadhaar regime, data can digitize every citizen with an identity that can be authenticated in real time for purposes beyond access to government services, including private-sector banking and telecom services, with an implication that scholars have likened to biopolitics.[71]

As economist Jean Dreze has pointed out, the Aadhaar Act's broad language enables the usage of Aadhaar data minus biological data and fits in with the business opportunities of such data.[72] According to the Aadhaar (Targeted Delivery of Financial and Other Subsidies, Benefits and Services) Act, passed in 2016, no core biometric information – that is, fingerprint, iris scan, or other biological attribute – will be shared with any entity. It is the collation of datasets that is of concern. While the UIDAI does not do the collation, various government agencies and private entities authorized as requesting entities can collate and categorize data, and the number of requesting entities is only rising. The NITI Aayog lays down its aim as strengthening the role of the private sector to improve the 'overall planning capacity in the country' through 'procuring technical consultancy services, strengthening project structuring and management skills in the public sector, and the empanelment of private sector consultancies'.[73] As such, various government agencies and private entities with access to data not covered by the privacy safeguards of the Aadhaar Act can build digital profiles of citizens based on personal information such as travel, education and income tax.[74] In 2023, the government has proposed that

70 UIDAI, *UIDAI Annual Report 2019–20*, New Delhi, 2020; NITI Aayog, *NITI Aayog Annual Report 2021–2022*, New Delhi, 2022.

71 Joseph Pugliese, *Biometrics: Bodies, Technologies, Biopolitics*, New York: Routledge, 2010; Btihaj Ajana, *Governing through Biometrics: The Biopolitics of Identity*, 1st edn, London: Palgrave Macmillan, 2013.

72 Jean Drèze, 'All that data that Aadhaar captures', *The Hindu*, 9 September 2017, at www.thehindu.com/opinion/lead/all-that-data-that-aadhaar-captures/article19646150.ece.

73 NITI Aayog, op. cit. (70), p. 25.

74 Drèze, op. cit. (72).

https://doi.org/10.1017/bjt.2023.11 Published online by Cambridge University Press

private entities be allowed to access Aadhaar data for ease of service delivery.[75] This is an amendment to the Aadhaar Authentication for Good Governance (Social Welfare, Innovation, Knowledge) Rules, 2020, which authorize requesting entities to access Aadhaar for identity verification for the purposes of good governance, prevention of leakage in welfare benefits or technological innovation. Besides the Union government, several states, such as Tamil Nadu, are already collating biometric data with their agencies under the State Resident Data Hub (SRDH) project.[76] The Tamil Nadu SRDH project aimed to unify its data repository with biometry-enabled data from the National Population Registry (NPR) and covers commercial tax, education schemes for underprivileged castes, health insurance and unorganized labour welfare boards, among other things. The Union government had begun seeding the Aadhaar number in the NPR database to enable government departments to select beneficiaries for their schemes in 2015 but halted it in 2016, according to the 2021–2 UIDAI annual report.

The 2016 Act also allows UIDAI to sign deals with the government, the states or Union Territories, or any other agencies, to collect, store, secure or process information; deliver Aadhaar numbers; or perform authentication. As such, future entrepreneurial opportunities are enormous. The official narratives about Aadhaar – data privacy; non-collection of social data such as caste, religion, and income by UIDAI; and data not being used by private entities – therefore take on a layered meaning when the various authentication services are considered. These narratives reveal the bureaucratic rationality of data usage where human data, including biometric data, are controlled by the state and the private sector. Moreover, the government can waive all these protection provisions and act on the Aadhaar information as it deems fit for matters of national security.

The size, scale and diversity of data available in India, and therefore to the government and corporate sectors due to the entanglement of the two, reveal the difficulty of understanding how Aadhaar and its biometric database would work or be used in the future. The Aadhaar Act keeps the door open for the future usage of biometric data. It states that it would promote research and development for 'advancement in biometrics and related areas, including usage of Aadhaar numbers through appropriate mechanisms'.[77] Again, this is a potential subversion of the provision protecting core biometric information, which the Act said would not be shared with anyone for reasons other than generation of Aadhaar numbers and authentication. The future looks even more complicated as one tries to understand the usage of human data, including biometric, identity and personal data, augmented by AI to make social policies for the population. The government justifies its inclination towards not only an AI-enabled but also an AI-enabling policy environment through data imaginaries that promise a range of possible services and goods, such as cancer research and agricultural productivity. The current data imaginaries involve India-specific data and data sets. To this end, Nasscom, the apex body for India's technology industry, recommended unlocking public-sector data sets to enable entreprenuers to provide customized AI applications in its 2021 report *Data Annotation: Billion Dollar Potential Driving the AI Revolution*. In the same year as Nasscom's report, NITI Aayog, which released India's National Strategy for Artificial Intelligence (NSAI) in 2018, reiteratedthat its focus area must now move 'from being data rich to data intelligent by making available clean, structured and annotated data'.[78]

75 The Wire staff, 'Govt's plan to let private companies access Aadhaar data has potential for mischief: SC ex-judge', *The Wire*,18 May 2023, at https://thewire.in/rights/aadhaar-private-companies-services-bn-srikrishna.

76 Anand Venkatanarayanan, 'The 360 degree database', *Medium*, 9 December 2017, at https://medium.com/karana/the-360-degree-database-17a0f91e6a33.

77 UIDAI, op. cit. (70), p. 8.

78 NITI Aayog, *Aim to Innovate* newsletter, New Delhi, April 2021, p. 9, at https://aim.gov.in/pdf/AIM_Newsletter.pdf; NITI Aayog, *National Strategy for Artificial Intelligence: #AIFORALL*, New Delhi, June 2018; Nasscom, *Data Annotation: Billion Dollar Potential Driving the AI Revolution*, Noida, February 2021, p. 43.

https://doi.org/10.1017/bjt.2023.11 Published online by Cambridge University Press

Conclusion

In the Indian government's insistence on Aadhaar's minimal biometric database and functions of identity issuance and authentication, the aim has been to establish this biometric data and AI-based data processing method as the most trusted ID document in India. Indeed, all other ID documents, such as passports, voter's ID and driving licences, have to deal with the problem of the existence of fake accounts and gaps in verifiability. This data imaginary associated with Aadhaar, even though rooted in practical benefits, creates a vision of 'tremendous potential to bring transformation as it empowers people in myriad ways so that a sense of enhanced security and trust prevails in the life of people at large'.[79] It enables the government to assume that biometric data owned by the individual belong to the state–corporate entanglement. Yet in the distance between the minimal data collection logic of the UIDAI and the distributed and networked possibilities of data mining through the Aadhaar Act, the emerging data imaginaries of the Aadhaar narratives commonly overlook the profiling aspect of current biometric data through the collation of various residents' data sets available to government agencies. The application of AI to this vast data set affords such narratives.

The history of biometric data and data sets that I have offered with Annandale and Mahalanobis, however, shows that biometric data are never shorn of profiling aspects. Annandale's data practices show that data imaginaries associated with physical and biometric data were historically associated with a racial imaginary. For Mahalanobis, the data imaginary perpetuated some colonial hierarchies while claiming the ground for population sampling for developmental planning in the post-independence years. In the Aadhaar regime, the government's motivation to use citizens' biometric data to distribute welfare and other services, driven by the ideology that digital data are an efficient governance tool, leads to the argument that minimal, essential biometric data are value-free, and something that the state–corporate entanglement can legitimately own for indefinite current and future purposes. This data imaginary feeds into the neo-liberal logic of diminishing state roles in social lives and developmental planning for maximum efficiency, a change from the socialist, Nehruvian planning model that Mahalanobis followed.[80] Historicizing data imaginaries through these three vignettes reveals how biometric data remain laden with assumptions, enabling them to perform certain functions of producing and using identities where the actual individual owner of the biometric data has little control over their usage.

Acknowledgements. Earlier versions of this article were presented at the Winter Symposium of the Mellon Sawyer Seminar Series on Histories of AI: A Genealogy of Power, University of Cambridge, at the University of Western Ontario, and the Annual Conference on South Asia, University of Wisconsin–Madison. Many thanks to my fellow speakers and panelists at these conferences for their timely input. I'm grateful to the seminar series organizers, particularly Syed Mustafa Ali and Matthew Jones, as well as Aaron Mendon-Plasek, who were early reviewers of this work. I also want to thank Stephanie Dick, Sarah Dillon, and Jonnie Penn for their feedback. This paper would not have been possible without Prof. Richard Staley's detailed interrogations and interventions during its development. I received excellent suggestions from two anonymous reviewers, for which I am very grateful. My sincere thanks to the editors of this journal and to Dr Rohan Deb Roy, the *Themes* editor for the 2023 issue of this journal, for their feedback and patience.

79 UIDAI, *UIDAI Annual Report 2017–18*, New Delhi, 2018, p. 1.

80 Benjamin Zachariah, *Developing India: An Intellectual and Social History, c. 1930–50*, New Delhi: Oxford University Press, 2012; Nikhil Menon, *Planning Democracy: Modern India's Quest for Development,* Cambridge: Cambridge University Press, 2022; Partha Chatterjee, 'Development planning and the Indian State', in Chatterjee, *Empire and Nation: Selected Essays*, New York: Columbia University Press, 2010, pp. 241–66.

Cite this article: Sahoo S (2023). Biometric data's colonial imaginaries continue in Aadhaar's minimal data. *BJHS Themes* **8**, 205–220. https://doi.org/10.1017/bjt.2023.11