spheres
JOURNAL FOR DIGITAL CULTURES

www.spheres-journal.org
ISSN 2363-8621
#5 *Spectres of AI*

NOOPUR RAVAL

AN AGENDA FOR DECOLONIZING DATA SCIENCE

In his essay "Technologies of Power: From Area Studies to Big Data", Manan Asif charts a fascinating continuity between early 20th century philological projects that were funded by the United States through a range of state and private entities and resulted in the field of 'area studies' and its echoes in the project of (big) data science studies. As Asif points out, in the aftermath of the Second World War, with the dawn of "the universal age" and the centring of the United States as the new global hegemon, the question of *knowing* the world arose – its peoples, areas, their ideological leanings and the way in which they could be managed. The Cold War era's anxieties about identifying, 'sniffing out' and converting countries that were supposedly in the danger of, or on the brink of falling prey to communism is also striking to me to the extent that this political ideology appeared as a certain predisposition that threatened the US-led new empire and needed to be quelled at all costs.

In this response piece, that I hope will be read as a companion to Asif's essay, I offer *other* histories, older and newer, to point to the fundamentally violent heart of technoscience as a historical, colonial enterprise. This is not a new point but bears repeating because it brings into question whether repurposing historically violent disciplines, knowledge projects and technologies might realise the decolonial futures we want. Further, thinking with Asif who asks for "a seat at the table" (for historians and other thinkers of coloniality) I offer a short phenomenology of data infrastructure as a toolkit for the historian to attend to institutional discourses but also, to the technics and logics of datafication. At the end, returning to Asif's call for an "algorithmic modality", I offer provocations for historians and others to adopt a radical commitment to a kind of decolonial futurity that might be necessary if we aspire to reconfigure the techniques and modalities of technoscience away from any present or future imperial project.

The continuities that Asif traces between the earlier project of area studies and the post-2008 rise of data science initiatives are indeed very revealing of the endurance of a knowledge/power project through which the US has sought to know, order and manage the world *for* its own purposes of domination. To say that most of the world's telecom and information infrastructure is made to advance the US hegemony would not be an exaggeration.[1] In that sense, to echo Asif's point, the area studies project very much lives on in the information and technology projects worldwide. If earlier philanthropic exercises were about collecting samples of flora and fauna, castes and tribes of the world to produce a handy diorama of what to expect 'out there', for at least two decades now, ICT for development (ICT4D) ventures supported by various US and European agencies have been trying to 'empower' the global South through technological interventions that impart tech-literacy at the minimum and seek to make behavioural changes at their extremes. Most recently, the push for Big Data for Development initiatives has advanced a peculiar methodological and ideological imperative – that we cannot empower the gendered, racial and geographic Other without rendering them completely knowable *and* that this must be achieved by harnessing the power of quantitative data and predictive analytics.[2] This is the developmental other to Asif's drone example. Both seek to render non-white bodies (within and outside of the US) completely transparent and knowable to manage and contain difference through a kind of potentiality – mapping bodies in space as 'activity' and producing enumerated and predictive narratives of how they might cause harm.

As we know from the disciplinary histories of anthropology, ethnology and development studies among others, these modalities of mapping, survey, enumeration etc. were developed much earlier in the service of various empires that actively commissioned seafarers, civil administrators, botanists, medical doctors and others to not only go *out there* and collect knowledge but also use colonies as laboratories of experimentation.[3] The starkest example is that of the colonial invention

---

1 Cp. Aouragh, Miriyam, and Paula Chakravartty, "Infrastructures of Empire: Towards a Critical Geopolitics of Media and Information Studies", *Media, Culture & Society*, 38 (4), 2016, pp. 559–575.

2 See the agenda-setting text of Sustainable Development Goals here that later materialized into specific data-for-development projects. Cp. United Nations (n.d.), "Big Data for Sustainable Development". Available at: https://www.un.org/en/sections/issues-depth/big-data-sustainable-development/index.html [accessed October 25, 2019].

3 See Rohan Deb Roy's article in the Smithsonian Magazine for a succinct summary of the British Empire and its relationship to imperial technoscience. Cp. Rohan Deb Roy, "Science Still Bears the Fingerprints of Colonialism", Smithsonian, April 9, 2018. Available at: https://www.smithsonianmag.com/science-nature/science-bears-fingerprints-colonialism-180968709/ [accessed October 25, 2019].

of fingerprinting technology. As Chandak Sengoopta demonstrates in his book *The Imprint of the Raj*,[4] while the routine identification of civilians would have been unthinkable within Britain, identifying potential criminal natives and keeping track of them, especially after what is now known as India's first war of Independence (in 1857), gained urgent imperative. Colonial administrators like Herbert Hope Risley who doubled up as anthropologists for the Empire produced regular census surveys and detailed handbooks that essentially decomposed and classified colonial subjects by castes, tribes, occupations and racio-ethnic descriptions along with sample photographs. These early databases were meant to equip colonial administrators with handy, retrievable information for the purposes of making the colony more predictable and hence governable. Similar to modern-day risk assessment models, the aim of colonial enumeration techniques was not so much to capture the entirety and richness of a native subject as to reconstruct the native *as a risky subject*,[5] (identifying and foregrounding personal, communal histories of vulnerability, poverty, disease, violence to gauge the potentiality of threat they pose to the colonial state).

There are countless other examples in postcolonial and colonial scholarship on technoscience that get at the heart of what I earlier described as the "violent heart of technoscience", but I am trying to get at something else here. In his forthcoming book *Distributed Blackness*, André Brock Jr. argues that, "there's nothing niche or subcultural about expressions of blackness on social media: internet use and practice now set the terms for what constitutes normative participation."[6] Similarly, I argue, both colonial and postcolonial bodies are inseparable from the past and contemporary technoscientific innovation and production. Being able to fingerprint the brown native, developing photographic technology that could only represent light skin and more recently, refining facial recognition technology to best capture the minority Uighur Muslims in China – are all forms of attunements built through experimentation on actual postcolonial bodies. Exploitative or violent use, then, is not incidental to technological enterprise (including data science) and, it both extracts the vitality of black and brown bodies but also enrols them as the data-labourers to assemble the global

---

4 Cp. Chandak Sengoopta, *Imprint of the Raj: How Fingerprinting was Born in Colonial India*, London, Macmillan, 2003.

5 Cp. Geeta Patel, "Risky Subjects: Insurance, Sexuality, and Capital, *Social Text*, 24 (4 (89)), 2006, pp. 25–65; David Arnold, *Colonizing the Body: State Medicine and Epidemic Disease in Nineteenth-century India*, Berkeley/Los Angeles/London, University of California Press, 1993.

6 André Brock Jr., *Distributed Blackness: African American Cybercultures*, Vol. 9, New York, NYU Press, 2020 (forthcoming).

assemblages of datafication. Following Asif's title ("technologies of power"), getting decolonial and anti-colonial historians a seat at the AI/ML/data science discussions is definitely of urgent imperative, not only to remind, invoke and infuse the technological with the historical but to also, consequently, replace general and post-hoc formulations of tech-ethics with more meaningful and deliberate ones. Attaching and visualising institutional histories, funding networks and most importantly, retaining colonial violence as contemporary memory and constantly mapping internationalist histories as a way of producing responsibility and solidarity could all be part of the reorientation of the algorithmic modality that Asif proposes.

However, another problem remains at hand. As many scholars of social computing have noted, there is also a barrier of technicality when it comes to investigating, critiquing and reorienting big data and algorithmic technologies.[7] Not going into the details of how that might be overcome (and certainly not suggesting the resort to digital humanities as a 'catchup' move), I propose that our 'seat at the table' might benefit from formal tactical engagement with information infrastructure and attunement to non-human agents as increasingly contributing to moral and political decisions. In his germinal essay titled "Databases as Discourse"[8], Mark Poster argues that a database is a discursive form for how it constitutes and "objectifies" a subject by disintegrating and reclassifying a subject into "grids of specification". In how a database holds a subject/object, the subject is not doubled in the postmodern sense but is "multiplied and decentered". The big shift that Poster and multiple other scholars after him have highlighted is how the discursive purpose of a database is not to store, mimic and represent an a priori rational, autonomous subject but rather, databases (as "perfect writing machines") operate primarily for *retrievability* – whereby signs referring to individuals and relationalities can be 'called' in myriad ways for myriad fleeting purposes. In that sense, who you *really* might be as a person and how inauthentic your (one) data doppelganger might be, are both irrelevant concerns because 'data doubles' (the databased representations) are not about representing the truth.

Furthermore, it is somewhat futile to demand accuracy or post-hoc

---

7 Cp. Paul Dourish, "Algorithms and Their Others: Algorithmic Culture in Context", *Big Data & Society*, 3 (2), 2016. Available at: https://doi.org/10.1177%2F2053951716665128 [accessed October 25, 2019]; Jenna Burrell, "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms", *Big Data & Society*, 3 (1), 2016. Available at: https://doi.org/10.1177%2F2053951715622512 [accessed October 25, 2019].

8 Cp. Mark Poster, "Databases as Discourse; or, Electronic Interpellations Mark Poster", *Computers, Surveillance, and Privacy*, 175, 1996.

accountability of datafied decisions, not only because they are largely opaque and dynamic assemblages but also because the databased decisions are only traces of dynamic 'interpolations'[9] or 'algorithmic pulls' that, when unpacked, only provide a regime of signs and techniques that could be assembled in numerous possible ways. The important political effect of databases, when one thinks about democratic processes and the public sphere is that they are now ready, permanent and malleable repositories of knowledge/power about populations agnostic to *who* is pulling the data *why*.[10]

As Sandra Robinson and others have argued building on Poster's original provocation, if the foundational form of the database is divorced from relationships of ideological signification (or fixed purposes of time and intention), since Poster's writing, data infrastructure (datasets and algorithms, their co-constitutive others) has only gotten more dynamic and complex. Not only does the datafied surveillance apparatus act on information *about* bodies and minds, but also datafied decisions are increasingly being produced automatically through machine-machine interactions.[11] Circling back to Asif's essay, there is a longer temporal shift that datafication introduces to the discourses of technological responsibility and accountability that still very much depend on a truth discourse – of signs (of activity) that must be fixed indexically in time and space. This is the kind of stability that the big data assemblage does not allow us as our multiple data selves are both situated and ephemeral but also suspended in a state of potentiality. As Robinson explains, there is also a bind within the construction of data proxies that are created through past activity but for short and long-term futurities. The big data regime, then, is not just performative but also spectral in scope.

To summarise, as both the original essay and my response illustrate, the longue durée of datafication suggests that knowledge-making, especially about the Other, is never innocent or incidental. Rather, when historicised, it reveals careful ways of constructing and maintaining power over the datafied. These historical imperial and colonial legacies are global in scope and as much sedimented in institutions as in the technics and logics of data science. If so, merely repurposing the master's tools may not dismantle the master's house.

---

9 Cp. John Cheney-Lippold, "A New Algorithmic Identity: Soft Biopolitics and the Modulation of Control", *Theory, Culture & Society*, 28 (6), 2011, pp. 164–181.

10 For a comprehensive discussion on signification and the politics of datafication, see Sandra Robinson, "Databases and Doppelgängers: New articulations of power", *Configurations*, 26 (4), 2018, pp. 411–440.

11 Cp. Simon Bart, "The Return of Panopticism: Supervision, Subjection and the New Surveillance", *Surveillance & Society*, 3 (1), 2005. Available at: https://doi.org/10.24908/ss.v3i1.3317 [accessed October 25, 2019].

The second part of my response takes up the questions of non-human as well as distributed agency in complex data infrastructure and the problems it may pose if we want to bring databases to the table as social agents and locate interests, ownership and responsibility within data science work. Asif advocates for an "algorithmic modality" of the kind that "[…] would disallow any triumphal narrative about the data sciences and would make clear that the historians and thinkers of coloniality deserve to be seated at the very tables where automation of our present is being considered." The points I have offered in my response essay take this proposition seriously and, in some ways, try to hint at *where* we (historians and others) could intervene in regimes assembled through human, non-human and distributed work. To this end, then, I suggest that the historian might have to adopt a radical commitment to decolonial futures, one that involves a politics of refusal (such as #abolishbigdata[12]) and a pedagogical move that includes visualising, politicising and organising the global South data workers who *materialise* the imperial data science visions through their labour, often with little knowledge of where their labour eventually gets plugged into.

[12] Y. Milner,. "Abolish Big Data", *Medium*, July 8, 2019. Available at: https://medium.com/@YESHICAN/abolish-big-data-ad0871579a41 [accessed November 7, 2019].