AI and Ethics (2022) 2:595–597
https://doi.org/10.1007/s43681-021-00130-8

OPINION PAPER

'Data dregs' and its implications for AI ethics: Revelations from the pandemic

Sun Sun Lim[1] · **Roland Bouffanais**[2]

Received: 12 December 2021 / Accepted: 24 December 2021 / Published online: 25 January 2022

Abstract

Technology giants today preside over vast troves of user data that are heavily mined for profit. The concentration of such valuable data in private hands to serve mainly commercial interests must be questioned. In this article, we argue that if data is the new oil, Big Tech companies possess extensive, encompassing and granular data that is tantamount to premium oil. In contrast, governments, universities and think tanks undertake data collection efforts that are comparatively modest in scale, scope, duration and resolution and must contend with 'data dregs'. Viewed against the backdrop of the COVID-19 pandemic, this sharp data asymmetry is unfortunate because the data Big Tech monopolizes is invaluable for boosting epidemiological control, formulating government policies, enhancing social services, improving urban planning and refining public education. We explain why this state of extreme data inequity undermines societal benefit and subverts our quest for ethical AI. We also propose how it should be addressed through data sharing and Open Data initiatives.

Keywords Big data · Big tech · Ethical AI · Data sharing · Academia · Government · Pandemic · Surveillance

If data is indeed the new oil that makes Big Tech companies our latter day oil barons, where does that leave states, academia and civil society that also thirst for data, albeit for non-commercial purposes? In our digitalising world, technology giants preside over highly detailed, identifiable data capturing individuals' physical, commercial, financial and even social activity, all of which is mined for profit. The extensive, encompassing and granular data these behemoths liberally tap is tantamount to the most prized grade of pure, unadulterated oil. In contrast, governments, universities and think tanks can only undertake data collection efforts that are comparatively modest in scale, scope, duration and resolution. Grossly inferior to premium grade oil, the data these institutions can muster veers closer to the dregs that are expunged after purifying oil. This sharp asymmetry between premium data being held in private hands while public institutions must contend with 'data dregs' is both unequal and damaging, with adverse implications for the future of AI and AI ethics. The prevailing pandemic offers an illuminating frame for analysing this egregious state of affairs.

Aside from claiming scores of victims, COVID-19 has also exposed deep cleavages between the digital haves and digital have-nots in almost every society. At the outset, with half of humanity subjected to lockdowns that saw business, education and healthcare services migrating online, digital connectivity was the clear game changer. Communities with inadequate or no digital access were significantly poorer off, unable to avail of online education, telework, telemedicine and e-government services [1]. But even among people and entities that enjoyed digital access, another divide emerged—that between the data haves and data have-nots [2]. Notably, companies with extensive online operations could instantly zero in on services and products in greatest demand to respond accordingly, with Amazon and Alibaba being prime examples. In contrast, firms without an online presence were data starved and less able to monitor and anticipate consumer needs.

Governments worldwide also leveraged data to contain COVID-19, mobilising diverse sources for insights into the population's physical movement and social interactions.

✉ Sun Sun Lim
Sunsun_lim@sutd.edu.sg

Roland Bouffanais
Roland.Bouffanais@uottawa.ca

[1] Singapore University of Technology and Design, Singapore, Singapore

[2] University of Ottawa, Ottawa, Canada

Many had to resort to 'coarse' proxy data from public transport, healthcare, security and public utility services, with countries, like Singapore and South Korea, initially utilising mobile phone GPS data for contact tracing[1] and identifying super spreader events [3]. Economic and financial data using anonymised sources from several private firms—credit card issuers, job posting aggregators and financial services firms—also offer governments near-real-time economic compasses for monitoring and adapting to rapidly evolving circumstances. While these data sources are not without value, their immediacy, extensiveness and multidimensionality are a pale shadow of the data held by the data haves, the victorious Big Tech companies.

These companies systematically engage in 'surveillance capitalism' to capture our 'behavioural surplus', gathering data on human activity, mobility, physiology, emotions and sentiments in astonishing detail [4]. Indeed, the likes of Apple, Facebook, Google and WeChat command vast troves of information on users that can give epidemiological control a veritable shot in the arm. Unfortunately—and rightfully—data privacy regulations restrict these companies from sharing such data, despite the legitimate imperative to contain the pandemic. Even so, the data privacy justification appears to be moot given the countless studies published in 2020 and 2021 using anonymised data sources originating from various private companies. For instance, researchers were able to identify social inequalities in human mobility during the early lockdowns of 2020 by using highly detailed mobile phone data from the operator Orange [5]. More recently, using anonymised mobile phone data from the same private operator, researchers also assessed the impact of mobility on epidemic spread, and more importantly, the impact of policies, such as mass quarantines and selective re-openings [6].

Beyond the pandemic, the data Big Tech has a stranglehold over is invaluable for formulating government policies, enhancing social services, improving urban planning and refining public education. Besides governments, academia, think tanks and civil society are also *not* privy to such data. This concentration of such valuable data in private hands to serve exclusively commercial interests must therefore be questioned, especially in light of humanity's bruising experience with COVID-19. Unless we change the status quo, the current state of data inequity that privileges private gains over public good will significantly hobble societally beneficial research for the foreseeable future [2]. Consider how, previously, academic research could draw on phone companies' anonymised records of phone calls to uncover the patterns underlying social exchange of information. Information exchange is a key tranche of urban interaction and provides illuminating insights into social governance and urban planning. With more phone calls and text messaging shifting to proprietary platforms, such as WhatsApp and Telegram, our ability to understand such social exchange of information is significantly constrained unless data sharing by Big Tech companies is mandated.

As a testament to their resourcefulness, academic researchers have developed techniques to gather data on social activity in the absence of access to privately held data. These include using apps to survey and interview individuals via their mobile devices, as well as to collect mobile trace data stored in individual mobile devices, including calling and texting logs, location data tagged to photographs, and app usage records. Similarly, research geared towards tracking human mobility patterns have deployed custom sensors that are provided to respondents, thereby necessitating considerable logistics [3]. These efforts, while laudable, will simply not yield data comparable in scale, granularity, comprehensiveness and quality to that collected by technology companies with both ease and regularity.

Ultimately, in a technologising world undergirded by Big Data, we must address the pressing question—how does the prevailing data asymmetry subvert our quest for ethical AI? First, the commercial exploitation of data for algorithms that automate everything from online advertisements to social media feeds and insurance premiums is an opaque exercise. In our 'black box society', these critical processes evade regulatory scrutiny through secrecy and active obfuscation [7]. Big Tech companies' data mining and algorithmic design processes are so complex that they have become incomprehensible to regulators, rendering hollow any requirements for transparency and accountability. Even if such pernicious trends are increasingly condoned, accountability of AI algorithms must not be forsaken as a lost cause [8].

Second, academic research is held to more rigorous ethical standards than that conducted in corporations [9]. Research-intensive universities have multidisciplinary ethical review boards that have oversight of detailed research protocols. Peer-review processes for academic publications also routinely require evidence of ethical research procedures. Such safeguards, even if not entirely failsafe, create a commendable culture of accountability. If academics are granted access to Big Data, they can help raise professional standards around its management, treatment and analysis to enhance fairness and explainability. These efforts can then help translate AI ethics from lofty principles to concrete practices.

[1] Nevertheless, there are concerns that the collection and publication by states of location data of Covid-19 victims have led to discrimination against marginalised groups, such as LGBTQ individuals in Seoul and Africans living in Guangzhou. See Benedetta Brevini & Frank Pasquale (2020). Revisiting the Black Box Society by rethinking the political economy of big data. Big Data & Society, October 2020, 10.1177/2053951720935146.

To be sure, technology companies are not immune to such criticisms and in a bid to boost their corporate social responsibility efforts, have sought to share some of their data through collaborations with research institutions. The Partnership on AI created in 2016 by several Big Tech companies is one such effort, although some partners have complained about the lack of achievements and progress [10]. Big Tech companies are also heavily involved in funding and participating in AI research conferences, where transparency norms and peer-review processes help lift the veil over some of their Big Data projects. However, such arrangements and initiatives are piecemeal and undertaken on terms that weigh decidedly in favour of the companies' interests.

Finally, all commodities in our societies are regulated and taxed for good reason. We must therefore ask ourselves why in our current Digital Gilded Age, one of the most valuable commodities of all—data—is effectively not regulated beyond individual privacy. Mandating some levels of data sharing could be achieved through the concept of 'Open Data', which borrows some of its tenets from the open-source software, open design, open knowledge and open access movements. Some governments have also recognised the societal benefits of making data available through national online portals. Initiatives by the open-source culture movement aim to make freely available a range of innovations, including software source code and hardware designs, to promote wider adoption and further refinement. Thanks to the collective ingenuity of developers, numerous hardware and software developments have achieved exemplary outcomes. For instance, the Linux computer operating system is widely recognised as the most successful and secure ever programmed and is widely used by commercial firms in data centres and to power the Internet of Things.

In totality therefore, when we regard the shifting contours of our Big Data society, private entities continue to gorge on data of the highest quality, while states and research institutions that seek data for the collective good must settle for vastly inferior 'data dregs'. As the amount of data society generates grows exponentially, we must reckon with the current data asymmetry becoming even more lopsided. If the existing quasi-monopolistic and proprietary model for Big Data persists, substantial societal benefits will fail to materialise. Regrettably, so will our quest for ethical AI.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. United Nations Conference on Trade and Development. The COVID-19 crisis: accentuating the need to bridge digital divides. https://unctad.org/system/files/official-document/dtlinf2020d1_en.pdf. (2020)
2. Boyd, D., Crawford, K.: Critical questions for big data. Inf Commun Soc **15**(5), 662–679 (2012). https://doi.org/10.1080/1369118X.2012.678878
3. Bouffanais, R., Lim, S.S.: Cities—try to predict superspreading hotspots for COVID-19. Nature **583**(7816), 352–355 (2020). https://doi.org/10.1038/d41586-020-02072-3
4. Zuboff, S.: The age of surveillance capitalism: the fight for a human future at the new frontier of power. Profile Books, London, UK (2019)
5. Hernando, A., Mateo, D., Barrios, I., Plastino, A.: Social inequalities in human mobility during the Spanish lockdown and post-lockdown in the Covid-19 pandemic of 2020. medRxiv. (2020). https://doi.org/10.1101/2020.10.26.20219709
6. Mazzoli, M., Pepe, E., Mateo, D., Cattuto, C., Gauvin, L., Bajardi, P., Tizzoni, M., Hernando, A., Meloni, S., Ramasco, J.J.: Interplay between mobility, multi-seeding and lockdowns shapes COVID-19 local impact. PLoS Comput. Biol, **17**(10), e1009326 (2021)
7. De Cremer, D., Kasparov, G.: The ethical AI—paradox: why better technology needs more and not less human responsibility. AI Ethics (2021). https://doi.org/10.1007/s43681-021-00075-y
8. Pasquale, F.: The black box society: the secret algorithms that control money and information. Harvard University Press, Cambridge, MA (2015)
9. Bell, E., Wray-Bliss, E.: Research ethics: regulations and responsibilities. In: Bryman, A., Buchanan, D. (eds.) Sage handbook of organizational research methods, pp. 78–92. Sage, London (2009)
10. Leenders, G.: The Regulation of Artificial Intelligence—a case study of the partnership on AI. https://becominghuman.ai/the-regulation-of-artificial-intelligence-a-case-study-of-the-partnership-on-ai-c1c22526c19f. (2019)