CENG SHOU YI/NURPHOTO VIA GETTY

Taiwan's innovative civic data culture shaped its rapid and effective pandemic response.

Everyone should decide how their digital data are used — not just tech companies

Jathan Sadowski, Salomé Viljoen & Meredith Whittaker

Smartphones, sensors and consumer habits reveal much about society. Too few people have a say in how these data are created and used.

A few decades ago, if a researcher wanted to ask how bad weather affected commuting patterns — the transport modes people use, the routes they take, the times they travel — they might have surveyed hundreds of people and counted cars, buses and bikes at major junctions.

Today, it is possible to access data on the movements of millions of people, taken from location trackers in phones or vehicles, sometimes in real time. These data can be combined with analyses of COVID-19 vaccinations to investigate the effects of commuters returning to the office. And weather data can be incorporated to determine whether more people are now more likely to work from home when heavy rain hits than they were a few years ago.

In theory. Reality often falls far short of this rosy vision.

Most of the data available to — or sought by — computational social scientists today are generated to answer questions that have nothing to do with their research. The data instead bear the mark of their original purpose, be that targeted advertisements or personalized insurance premiums. These data can be cautiously repurposed to answer other questions — wearable fitness trackers could inform studies of obesity, for example — but crucial gaps usually remain. As a result, scientists often use workarounds to glean meaning from what they can get[1].

For instance, analysts trying to answer questions about transportation patterns for a city government in Greater Sydney, Australia, had to make use of the low-quality spatial and temporal data created when mobile phones ping cell towers[2]. What's more, they had to purchase these data at high cost from a telecommunications provider.

In our view, the current model, in which the digital traces of our lives are monopolized by corporations, threatens the ability of society to produce the rigorous, independent research needed to tackle pressing issues. It also restricts what information can be accessed and the questions that can be asked. This limits progress in understanding complex phenomena, from how vaccine coverage alters behaviour to how algorithms influence the spread of misinformation.

Instead, we call for the creation, management and curation of behavioural data in public data trusts.

Access denied

This political economy of data puts social scientists in a difficult position. Access comes with conditions: companies have an active interest in the questions researchers ask (or don't) as well as the data they can access and how it is analysed. And it is rarely possible for scientists to determine what information was not included when the gatekeepers do grant access, or how the data were generated in the first place.

At best, this can have a chilling effect on scholarship. Some studies won't be done if they could threaten the data provider's reputation or bottom line. At worst, researchers can feel pressure to align their studies and results with the values and priorities of technology companies. Unflattering findings could see data access revoked, imperilling the continuity of a researcher's work, and potentially also their standing in their institution and with their peers.

In March, for example, a report on the Responsible AI team at Facebook showed that researchers were restricted in the types of problems they could study and the solutions they could propose. Rather than being able to root out the disinformation and hate speech that contributes to engagement, their work had to focus on technical changes to bias in systems (see go.nature.com/2t5kudw).

A reliance on the largesse of private companies also challenges tenets of scientific rigour and responsibility. Contractual restrictions can prevent researchers from reproducing and validating others' results. In 2019, health researchers reported that 'significant racial bias' in the training data for a proprietary commercial algorithm meant that US$1,800 less per year was spent on the treatment of Black patients compared with white patients with the same level of health[3]. The bias — which is disputed by the company — was revealed only when researchers did an independent audit of the records of a large university hospital.

The status quo poses serious problems. Increasingly, approaches that amass demographic data, study behaviours and predict risk factors are equated with big tech's unscrupulous practices — diminishing the reputation and credibility of the techniques in the long-term. And the dominance of a few walled gardens of data is shaping computational social science. PhDs and tenure are often granted on the basis of the funding, data, publications and prestige secured through industry partnerships.

Data pipeline

What we face is not simply limited access to proprietary data, but fundamental questions regarding the entire pipeline of how those data arise and where they go.

What companies deem valuable can distort the kinds of data available for analysis. Tech giants place great importance on behavioural information about people as a new asset class[4]. This influences research agendas, in part because the data are available in large amounts. Computational social scientists often use social-media data, for instance, as an imperfect proxy for many other factors such as mobility or health, even when they are far from ideal for answering their questions[5].

Moreover, insights can be tainted by data that were, often unknowingly, constructed using inappropriate assumptions and harmful biases. For instance, AI researchers have uncovered how large data sets such as ImageNet, used to train and assess machine-learning systems for more than a decade, are encoded with sexist and racist stereotypes, which then carry forward into software[6,7].

Democratic governance

There's little recourse to address these fundamental issues without transformative change to the private monopolization of data. Systems need to be established that are more suitable for the analysis of social phenomena, and in ways that are ethical, equitable and scientifically sound. Just as patented knowledge enters the public domain when intellectual-property rights expire, so, too, should behavioural data gathered by companies come under democratic control after some time.

A model with better control would involve collective stewardship of the data pipeline, in public trusts that are subjected to scientific oversight and democratic accountability. Existing work paves the way for such instruments. For example, a report by Element AI and Nesta outlines how trusts are an attractive policy tool for pooling the rights of data subjects and setting terms of use (see go.nature.com/3decirk; S.V. was a participant in the workshop on which the report is based).

Barcelona in Spain has piloted a promising approach. In 2017, it created a 'city data commons', giving residents control over how data about them and their communities were produced, as well as the power to participate in governance decisions. Its Open Data portal currently contains 503 data sets about the municipality, including real-time information on the use of the city's bicycle-sharing system.

Such democratic control helps to protect the people these data purport to be about. Public governance confers extra rights and rules — including anti-discrimination, due process and greater accountability. In most cases, these protections extend much further

MATTHIAS BALK/DPA/ALAMY

Data sovereignty champion Francesca Bria helped to found Barcelona's Smart City initiative to give residents control over their data.

than do private obligations, although there is variation between countries and regions.

Collective stewardship can emphasize the socially valuable aspect of information — not what is known about a person, but what that reveals about how people are alike and connected[8]. Instead of focusing only on the rights of individuals, a public trust can and should also represent the interests and values of groups affected by downstream uses of data products. For instance, when photos were scraped from cloud-storage websites and used by the company ClearView AI in New York to train powerful facial recognition software, the people photographed weren't aware this was happening. What of the businesses and police departments who bought the package?

Of course, the owning of data by public institutions comes with its own challenges. Governments sometimes use data to inflict serious harms, such as by targeting marginalized populations, and they can escape accountability through authoritarian measures. This is why public trusts must be designed for democratic governance from the outset. They must be representative of and responsive to the communities that the data are created about.

Strict siloes must be put in place so that the public data pipeline is not accessible to or influenced by other government organizations, such as the police or military. Singapore, for example, has used GPS data from mobile phones for contact-tracing during the COVID-19 pandemic. But citizens' trust was eroded when it was revealed that police used these same data during murder investigations.

Three steps

We recommend three steps that policymakers and scientific institutions should take towards the safeguarding of behavioural data as a public good.

Build public infrastructure. Measurement, computational and storage systems should be funded and maintained to support the construction of large data sets that are appropriate for quantitative and qualitative research. Resources should go to communities and organizations that are already engaging in these practices, including Indigenous peoples working to govern, classify and control their knowledge using principles of 'data sovereignty'. The infrastructure must be buttressed with robust participation mechanisms, so that those whom the data are 'about' are able to set the collection agenda as well as challenge and remedy inaccurate or harmful use.

Take control. Policies are needed that transfer data created and controlled by private entities to public institutions. They must also cover details of the underlying measurement methodologies, collection processes and storage environment.

There is already a legal precedent for granting private companies restricted rights to non-tangible assets that eventually revert to the public domain. For instance, the Hatch–Waxman Act governs intellectual property in generic drug production[9]. We propose a policy in which companies have a limited monopoly over the data they create and own. After a set period of time — say 3 years — these data either become a public resource or are eliminated.

Such a policy can also apply to any models that the data have been used to train or inform, because they could pose an undue risk to people if retained. There is precedent for this, too: in May, the US Federal Trade Commission required the destruction of facial-recognition algorithms trained on photos that were obtained deceptively. Provisions could also be tied to existing data-privacy regulations by giving companies favourable terms and incentives if they turn over data sets and metadata to universities, archives or other public institutions to manage.

> **"What companies deem valuable can distort the kinds of data available for analysis."**

Expand governance. Dedicated institutions should be created with the capacities to steward data in the public interest. There is no need to start from scratch. In the United States, the Library of Congress, National Science Foundation and National Institutes of Health could all serve as institutional models — and all have representatives on public data trusts.

Such institutions would be staffed by database managers who are trained in the ethical standards of library science, which balance knowledge curation for the public good against the risks that arise from sharing information. Experts in measurement and quantitative and qualitative methods could develop new efforts to generate data, working closely with researchers and communities to determine what socially minded questions to ask.

Computational social scientists, following the model of sworn statistical officers in the US Census Bureau, would evaluate the sensitivity of the source. Data from low-sensitivity sources could be published — as aggregated, anonymized information. And access to high-sensitivity data would be strictly safeguarded — including individual, identifiable information. A public data trust could also invite community groups and advocacy organizations to help shape protocols for consent and dispute; agendas for data construction and research goals; and requirements for accessing and using data.

Demand change

We are not alone in this fight. We need only look at the sweeping investigations into antitrust actions against platforms such as Google, Facebook, Amazon and Alibaba in the United States, European Union, Australia and China. The COVID-19 pandemic has also provided momentum. This March, the science academies of the G7 group called for a mechanism to oblige public and private organizations to share relevant data during health emergencies (see go.nature.com/2sjqj2v).

To get the ball rolling, scientists whose work relies on large proprietary data sets should speak out — on social media and at conferences such as NeurIPS — about the perils of corporate data gatekeeping and share their lived experiences with these difficult ethical choices. They should pressure universities to call for changes in current data-ownership regimes and ally with community groups already campaigning for redress from harm enabled by surveillance.

Representatives from academic associations and government bodies such as census offices and national libraries should form an interdisciplinary working group to develop policy for the creation of public data trusts. Computational social scientists must play their part as public stewards of an important collective resource for knowing ourselves and our societies.

The authors

Jathan Sadowski is a research fellow in the Emerging Technologies Research Lab in the Faculty of Information Technology and in the Centre of Excellence for Automated Decision-Making and Society in the Faculty of Arts, both at Monash University, Melbourne, Australia. **Salomé Viljoen** is an academic fellow at Columbia Law School in New York City, USA. **Meredith Whittaker** is the Minderoo Research Professor at New York University and co-founder and faculty director of the AI Now Institute in New York, USA.
e-mail: jathan.sadowski@monash.edu

1. Lazer, D. *et al. Nature* **595**, 189–196 (2021).
2. Dowling, R., McGuirk, P., Maalsen, S. & Sadowski, J. *Urban Stud.* https://doi.org/10.1177/0042098020986292 (2021).
3. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. *Science* **366**, 447–453 (2019).
4. Sadowski, J. *Big Data Soc.* https://doi.org/10.1177/2053951718820549 (2019).
5. Sloan, L. & Quan-Haase, A. (eds) *The Sage Handbook of Social Media Research Methods* (Sage Publications, 2017).
6. Hutchinson, B. *et al.* in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* 560–575 (Association for Computing Machinery, 2021).
7. Raji, I. D. *et al.* in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 145–151 (Association for Computing Machinery, 2020).
8. Viljoen, S. Preprint at SSRN https://doi.org/10.2139/ssrn.3727562 (2021).
9. Lewis, R. A. *J. Contemp. Health Law Policy* **8**, 361–378 (1992).

The authors declare no competing interests.