Meaningful measures of human society in the twenty-first century

https://doi.org/10.1038/s41586-021-03660-7

Received: 3 March 2021

Accepted: 20 May 2021

Published online: 30 June 2021

Check for updates

David Lazer[1,2]✉, Eszter Hargittai[3], Deen Freelon[4], Sandra Gonzalez-Bailon[5], Kevin Munger[6], Katherine Ognyanova[7] & Jason Radford[1]

Science rarely proceeds beyond what scientists can observe and measure, and sometimes what can be observed proceeds far ahead of scientific understanding. The twenty-first century offers such a moment in the study of human societies. A vastly larger share of behaviours is observed today than would have been imaginable at the close of the twentieth century. Our interpersonal communication, our movements and many of our everyday actions, are all potentially accessible for scientific research; sometimes through purposive instrumentation for scientific objectives (for example, satellite imagery), but far more often these objectives are, literally, an afterthought (for example, Twitter data streams). Here we evaluate the potential of this massive instrumentation—the creation of techniques for the structured representation and quantification—of human behaviour through the lens of scientific measurement and its principles. In particular, we focus on the question of how we extract scientific meaning from data that often were not created for such purposes. These data present conceptual, computational and ethical challenges that require a rejuvenation of our scientific theories to keep up with the rapidly changing social realities and our capacities to capture them. We require, in other words, new approaches to manage, use and analyse data.

Sensor technologies have multiplied across many realms of human activity, from tracking devices in cars to online browsing. Satellites scan and digitize the planet at regular intervals. The development of techniques for processing unstructured data such as text, images, audio and video by computer scientists animates the conversion of—for example—books[1], radio broadcasts[2] and television shows[3] into data. In the twenty-first century, human behaviour—from mobility to information consumption to various types of interpersonal communication—is increasingly recorded somewhere and potentially computationally tractable. Past communication technologies, from mail to print to fax, typically left far fewer durable and accessible artefacts; those that did have become computationally accessible only in the past decade or so, as the relevant physical artefacts were digitized. The digitization of books is an example, which enables the computational analysis of a massive corpus of human expression that stretches back centuries[4].

The emergence of these new data streams has often been compared to the development of the telescope. As Robert Merton famously wrote, "Perhaps sociology is not yet ready for its Einstein because it has not yet found its Kepler….”[5]. Merton's provocation was that sociology did not yet have the empirical foundations on which to build great theory. Duncan Watts, in response, writes 62 years later, "…by rendering the unmeasurable measurable, the technological revolution in mobile, Web, and Internet communications has the potential to revolutionize our understanding of ourselves and how we interact. Merton was right: social science still has not found its Kepler. But three hundred years after Alexander Pope argued that the proper study of mankind should lie not in the heavens but in ourselves, we have finally found our telescope.”[6].

We believe in the potential of digital data sources to transform the social sciences. However, the metaphor of the data streams from the instrumented society as a 'telescope' is misleading in important ways. First, the study of societies is different from the study of the stars, because the patterns that characterize human behaviour will generally differ across time and place. Second, the measures built from these streams are potentially suspect in ways that must be actively interrogated, because these sources were not built with scientific goals in mind. We now turn to the first point; the remainder of the paper is devoted to the second.

The unstable logics of society and measurement

Empirical social science is largely focused on finding generalizable but not universal patterns in human behaviour. The part of the social sciences that has the intent of finding such universal patterns in human behaviour (for example, evolutionary psychology) is tiny relative to the whole field. The issue of the instability of the rules that govern human society is exacerbated by the very sociotechnical systems that are gathering the data about people, which are actively (and in some cases intentionally) changing the social world that social science would study. Through what social scientists call reflexivity and self-fulfilling prophecies, humans actively change the world that they are observing

[1]Network Science Institute, Northeastern University, Boston, MA, USA. [2]Institute for Quantitative Social Science, Harvard University, Cambridge, MA, USA. [3]Department of Communication and Media Research, University of Zurich, Zurich, Switzerland. [4]Hussman School of Journalism and Media, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [5]Annenberg School for Communication, University of Pennsylvania, Philadelphia, PA, USA. [6]Department of Political Science, Pennsylvania State University, State College, PA, USA. [7]School of Communication and Information, Rutgers University, New Brunswick, NJ, USA. ✉e-mail: d.lazer@northeastern.edu

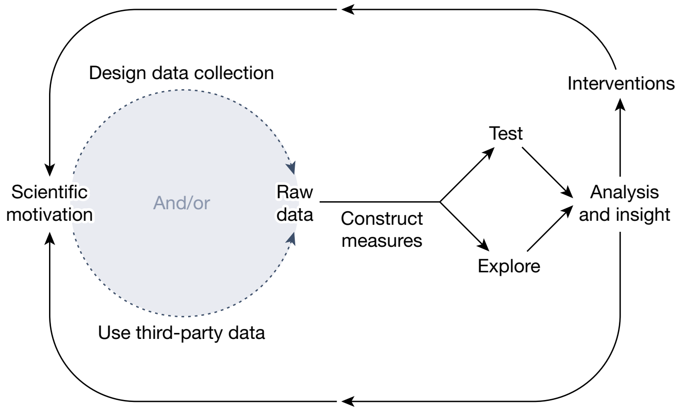

Fig. 1 | Measurement in social science. Measurement is the bridge connecting scientific motivations and data with insight and applications.

by acting on the knowledge gained (in part by measurement instruments).[78]

Reflexivity refers to the loop that links social reality with the theories and the metrics that we devise to explain it. The 'bandwagon' and 'underdog' effects, for instance, have long been identified in the analysis of electoral politics to explain the impact that polls and forecasts have on voting behaviour. If candidates are projected as likely winners, more people may decide to vote for them (bandwagon effect) or, conversely, more people could mobilize to increase support for the candidate expected to lose (underdog effect)[7]. These effects reflect the impact that measurements have on attitudes and behaviour[8,9], and how our measures can distort the phenomena that they were designed to monitor. These distortions, in turn, can be amplified by algorithmic decision-making in public health, law enforcement, sentencing, education and hiring[10,11].

Reflexivity also takes the form of the observer effect, which happens when people modify their behaviour if they know they are being watched[12,13]. Digital technologies have created a new version of the reflexivity problem, amplifying the performative aspect that is intrinsic in social indicators. When Google launched the Flu Trends project in 2008, the goal was to use search queries to estimate the prevalence of flu symptoms in the population. In 2013, however, Flu Trends substantially overestimated peak flu levels. One of the reasons was the flawed assumption that search behaviour was driven by external events, such as having flu symptoms. In fact, Google's algorithms were driving those patterns as well: by trying to anticipate the intent of the users through recommended search terms, Google was distorting the information users would have otherwise revealed[14]. The reaction to the observed phenomenon, in other words, changed the phenomenon itself.

Obfuscation tactics represent another version of the observer effect: we can now disrupt measurements by deliberately adding ambiguous or misleading information to interfere with data collection. Examples of obfuscation include editing profile photographs to prevent facial recognition; using virtual private networking (VPN) to hide one's location when browsing the web; or using group identity (for example, many people under one user account) to obscure specifics about the actions of one user[15]. The reflexivity loop here is created by the awareness that behavioural traces feed into metrics and surveillance, so the meaning of that behaviour is intentionally altered. This is similar to when respondents lie to survey researchers, but on a much larger scale. And because the skills needed to know that surveillance is happening and how to implement obfuscation to address this are not randomly distributed across the population, the individuals whose data will be altered in such ways will not be random either.

The unobtrusive nature of many digital measures suggests that, overall, observer effects may be less of an issue with these new data sources compared to the past when—for example—the gender, age and race of the person conducting an interview could vastly change the answers that respondents provided[16]. However, the loop that connects social reality with the metrics that we devise to analyse it has been strengthened—reflexivity is now embedded in the instruments used to monitor and predict human behaviour. It is as if the Hubble telescope were organizing the placement and behaviour of the stars at the same time as it is observing them. Social media, for example, not only capture human behaviour, but also have the potential to alter important patterns of human society, such as the speed of information flows, the scope of media production and the actors responsible for defining public opinion.

As a result of the fluidity of the principles organizing human society, the meaning of a given measure will also evolve. Part of why the social sciences must accommodate these new types of data is that emerging sociotechnical systems are reducing the relevance of some old scientific instruments used to measure human behaviour. Existing measures of key concepts such as gross domestic product and geographical mobility are shaped by the strengths and weaknesses of twentieth century data. If we only evaluate new measures against the old, we simply replicate their shortcomings, mistaking the gold standard of the twentieth century for objective truth. For example, consider the standard question (originally from 1978) from the American National Election Studies[17] about radio consumption regarding an election: "Would you say you listened to a good many, several, or just one or two speeches or discussions on the radio about 'the campaign'?"

This construction of 'media consumption' as consisting of a countable number of discrete units is an artefact of the technology of the broadcast era. This question bears little relation to how people access digital media today. It would be futile to attempt to capture behaviour regarding social media by asking questions such as 'How many tweets did you see today?' or 'What Twitter accounts showed up in your feed?'. Many of the ways to measure behaviour developed in the early days of quantitative social science were: (1) necessary given constraints on measurement at the time; and (2) grounded in a social reality that was markedly different.

Figure 1 summarizes how measurement fits into the general scientific process. We discuss below the central challenges of turning data from these sociotechnical systems into scientific measurements. We include in this discussion two motivating examples of data streams that have been the basis of much social science research: location data from mobile phones and social media posts on Twitter. The key questions we turn to now are what and whom we measure with massively instrumented human behaviour, focusing on the key principles of measurement summarized in Box 1.

What trace data measure

The goal of measurement using behavioural trace data is to extract meaning from the raw data generated from instrumentation. All scientific data instrumentation confronts this issue, but the leap from raw data to meaningful measures is often particularly large when we use data recycled from systems designed for other purposes[18]. For example, mobility data from mobile phones reporting specific latitudes and longitudes are largely uninteresting without further processing, which enables us to measure proximity, mobility and other socially relevant concepts.

The key challenge is whether our measurement accurately captures the construct that we want to examine. Does it closely match other measures of the same thing? What is the potential slippage between construct and concept (for example, if measuring physical activity from mobile phones, how consequential are the missed stationary activities, such as a treadmill?). When we examine supposedly unrelated constructs, do our measures reflect the expected lack of association? By and large, twenty-first century observational data are not designed

Box 1

Central principles of measurement

Measurement should follow definitions of what matters

Efforts to measure observed phenomena are premised on the identification of relevant questions. What matters is driven by research questions, which may be motivated by normative goals, theoretical debates or empirical puzzles.

Measures must be actively constructed out of data

Instrumentation designed for research purposes often generates scientific data. But data collected for purposes other than scientific research are also frequently repurposed by scholars. Data do not have meaning in themselves—to become measures of some theoretical construct, they must be transformed by methods that make them systematically relatable to one another, and to scientific theory.

Scientific measurement follows from the above principles in a constantly evolving loop

Scientific motivation directs researchers to design a data-collection protocol, use third-party data or develop some fusion of the two. In their raw format, data offer the observations that are processed into the measures that will enable testing pre-conceived hypotheses (in a deductive way) or derive new hypotheses from exploratory analyses (in an inductive, data-driven way). These deductive and inductive analyses aim to offer insights that can then feed back into scientific motivations, inform policy interventions, or—more generally—drive the basic and applied arms of research.

for research and need to be linked to known concepts before we can use the data to answer scientific research questions.

The meaning of measures is derived, in part, from theory. Theoretically driven designs that apply existing knowledge to interpret digital signals can overcome many of the problems of using instrumented behavioural data. Conversely, undertheorized ad hoc operationalizations can make research findings difficult to interpret and inconsistent across studies. As noted previously[19], formal theory is useful not only in generating hypotheses, but also in selecting an appropriate way of measuring constructs with big data.

Consider, for example, the use of mobility data to study the spread of COVID-19. Multiple studies used real-time travel data to track the movement of people from Wuhan to other provinces in China[20,21]. The researchers found that population movements from Wuhan were strongly predictive of the introduction of the coronavirus to a region. Local controls then predicted the subsequent spread of the virus. In these studies there is a well-theorized process based on the assumption that the spread of the virus is driven by the proximity of individuals. The chosen theoretical framework, in turn, informs how generalizable those findings could be to other cases. That is, we might expect similar patterns in the USA[22], but not in Australia, given the rigorous testing and isolation procedures that were imposed on visitors in the latter country. The results of any given empirical study are necessarily local, in both time and space; theory is needed for the appropriate movement of any measurement to a new geographical or temporal context[23,24].

As we conduct more research using high-volume, complex data sources and formats, methods that offer insights into the validity of new measures become especially valuable. One promising approach is to examine classic validated self-reported scales in conjunction with new ways of measuring related concepts. For instance, self-reported news attention and exposure can be used in conjunction with eye-tracking to capture visual attention to online content[25]. A similar triangulation of approaches to measurement can also be useful in confirming the validity and robustness of new behavioural constructs[26]. Researchers have used mobile phone data to design proximity-based measures capturing the amount of time that people spend close to each other[27]. These metrics can serve a variety of useful purposes. They can be used as a proxy for relationship strength, or give us a way to track possible pathways of virus contagion. There is, however, the potential for error—two people whose devices appear near each other as measured by their Bluetooth beacons may, for instance, be separated by a wall or may simply be charging their phones from the same outlet. In cases such as this, triangulation can come from the inclusion of self-reported data, such as sending a message to someone's phone to ask them who else is nearby at the time.

For internet-based research, both basic population characteristics and underlying mechanisms that structure user behaviour on digital platforms remain relatively poorly understood. Many basic concepts remain difficult to measure even in online platforms that offer easy data access to researchers. Despite the thousands of papers based on Twitter data in recent years, social media scholars still find that identifying the demographic characteristics of individual users remains a big challenge. Additionally, researchers still cannot reliably distinguish humans from non-humans (for example, bots, collective accounts or organizations), although there have been important strides made in that direction[28,29]. As a result, the large majority of Twitter research is making inferences about accounts or tweets; very little of Twitter research can reasonably claim to be making statements about the behaviours of humans. For research questions that focus on human behaviour on Twitter, methods that link user accounts to administrative data or to survey responses offer promise in identifying humans and their demographic attributes on Twitter[30].

Even when it is clear that humans are the source of a given behaviour, there may be a challenge in attribution of specific behaviours to specific humans. In its early days, audience research for broadcast television, for instance, encountered challenges with multi-member households[31]. The data in those cases would suggest the existence of someone with a taste for children's cartoons and cable news, when, in fact, there were two different individuals involved. Technological sensors can thus be actively misleading when behaviour is divided across humans (two people using the same Netflix account) or across sensors (the same person viewing Twitter on a smartphone and a desktop). Further exacerbating the issue is that the sensor–human mismatch could rapidly evolve over time. Thus, for example, a finding based on desktop browsing data that news consumption has systematically changed could simply be an artefact of the progressive shift from desktop browsers to mobile apps[32]. The lack of stability of human use of these different systems (and sensors) may make such a comparison over time essentially impossible.

The use of models based on other data can facilitate the measurement of focal behaviour. For example, who uses which device can be modelled from other data, and the outputs of this model will be less sensitive than discrete assumptions about the identity of a device user. The cable news viewer may be the grandparent and the Xbox user the grandchild. However, the data that are included in these models must always come from the past, and the relationship between measures is itself unstable. This is the fundamental problem of induction, and while it cannot be surmounted without a metaphysical revolution, we propose that constantly updated measurements and models represent our best amelioration of the problem. That is, we should plan for the slippage of our measurements and conduct an ongoing assessment of how particular measures capture the current social reality. For example, measures of inflation need to assess how the set of goods that people consume changes over time. This is a useful recalibration, although it also illustrates the limits of this approach, because the emergence of completely new items (no one was buying smartphones in 2000) makes consumption across time inherently incomparable.

The proliferation of communication technologies, driven by the internet, also yields a fragmentation of behaviours into different data silos. Consider a research question that explores whether non-proximate synchronous voice-mediated communication is important to reducing feelings of social isolation. The past half century has seen a steady fracturing of this behaviour into different systems—from government-mandated monopolies (for example, Ma Bell in the USA) to oligopolies to a countless number of internet providers. Furthermore, there are plausibly systematic biases in the data captured in any one of these systems—whom you talk to on your mobile phone might be systematically different from whom you talk to via Zoom, Skype or WhatsApp[33]. Even the tortured linguistic construction used above reflects the sociotechnical complexity: not too long ago 'non-proximate synchronous voice-mediated communication' would have been described simply as a 'phone call'. One important consequence of this technological fragmentation is that measurements relying on a single digital device or service should be interpreted with considerable caution. The answers that we find could plausibly differ from those we could get by measuring the behaviour in a similar but different technology. Ironically, because of that complexity, an accurate picture of whom someone generally talks to may be better captured through a simple survey question than through records from a single platform.

Conversely, behaviours observed in different silos that seem similar might actually be capturing very different phenomena. Just as various name generators that are used in surveys to generate lists of contacts result in the identification of different social ties[34], a friend on Facebook does not denote the same relationship as a Twitter follower or a LinkedIn contact. Moreover, none of these relations denote a 'friend' as used either colloquially or scientifically, although there are very likely some strong statistical connections among these concepts. These systems, furthermore, change over time and their affordances—what they allow users to do—also evolve. This in turn means that the causal processes that underlie our online social actions, relationships and structures are constantly changing. As such, we must now be aware of system-varying properties of measures such as temporal and inter-system validity. The challenge then becomes developing measures that provide some degree of generalizability over time or across systems for a given research question.

Another deep problem is the algorithmic confounding of measurement[35]. Confounding here refers to our inability to distinguish signals that represent typical human behaviour from ones that result from the rules that govern a digital platform. Without knowing how a system is designed, we could easily attribute social motives to behaviour driven by algorithmic decisions. If Twitter's feed suddenly starts to prioritize sports, a user may find out who won an Olympics competition without any changes in their underlying interest in sports. Such changes are often difficult to detect, both because they are sometimes introduced without notice and because they may roll out unevenly, affecting certain user populations before others. This mechanism also functions in more subtle ways, such as how natural human proclivities are enhanced by algorithmic prompts. For instance, if Twitter systematically suggests that you follow back people who already follow you, that can boost our natural tendency to reciprocate social ties[36]. More generally, internet companies aim to manipulate human behaviour so as to increase engagement on their platforms (for example, Facebook, Twitter and Instagram) or money spent on their products (such as Amazon and Ebay). Those machine-learning-based manipulations are pervasive, and any efforts to develop measures from platform data need to evaluate the extent to which algorithms will distort both the measures and any downstream analyses. Because of their importance, those algorithms are worthy of closer study[11,37].

Although an in-depth discussion of causal inference is outside the scope of this paper, we should note that a number of measurement issues identified here present a particular problem for research that aims to establish cause and effect. Lack of stability in measurement over time, for instance, may induce researchers to attribute the changes in a focal outcome to an unrelated external event. The discussion above regarding Google Flu Trends is also relevant here. In that case, there was an implicit assumption that the flu was causally related to flu-related searches on Google. However, if Google around 2013 was proposing flu-related searches during flu season because it had, deep in its complex algorithmic machinery, inferred it was flu season, the measure of exactly the same behaviour in 2013 would mean something very different than it did in 2008.

The malleability of human expression and language also poses general challenges around inferences of attitudes and opinions from language and image data[38]. Expressions of sentiment on Twitter are notoriously difficult for computers to decode, as they typically stumble over sarcasm, irony and hyperbole[39]. How problematic that is depends on the structure of the noise and, again, on what matters—that is, the research question.

Whom trace data measure

Human behaviour is a multi-level concept that often requires measurements at the individual level to make inferences about the distribution of behaviours, attitudes and attributes at the collective level. The research question should make clear what population is of interest to a particular study. That population could include people everywhere of all types, or it could be specific to a certain geographical region (a city or country), a particular community (a hobby group or company) or a myriad of other subpopulations (youths, immigrants, or politicians). Especially when entire populations are concerned, it is not feasible, logistically or financially, to collect data about everybody. In such cases, researchers should ideally collect data about a random sample of the population, which means that each member of the population has an equal probability of being in the sample. This ideal was never quite achievable, and is even less relevant in a world in which response rates to survey requests are below 10%, with uneven rates of accessibility of people across modalities of recruitment[40].

With system-level data, one may be tempted to think that everybody is represented since the actions of all of the users are in the dataset. However, the sampling in this case happens at the level of who is a user of the system from which the data are collected as well as who is most active on said system[41]. It is, at best, a 'convenience census' of the platform under investigation rather than the whole population[42]. If the scientific objective is to make a statement about the people on the platform, that census might be compelling. However, any leap to generalize beyond that platform must be viewed more critically. This is a particular problem for research on Twitter, the most commonly cited source of emerging data, as it is used by only about 20% of the US population and is even less popular in most other countries[43–45]. Importantly, users of social media platforms do not mirror the general population of internet users either demographically[41] or regarding other attributes such as their interests[41,44,46]. In light of recent progress in promoting the representativeness of research populations in other domains[42,47], it is imperative to think carefully about these issues in the realm of social media[47]. We also note that methods that recalibrate data to make reasonable population-level inferences can be particularly powerful when applied to large-scale data[48].

The issue of generalizability is amplified when only a subset of the platform population is studied. The key question is whether and how the nature of the sample affects the inferences being drawn. Thus, for example, a study of Twitter users who include their names and locations in their profiles[49] raises the question: do these findings generalize to Twitter users who do not divulge such details? Similarly, another study[50] examines the consumption patterns of political information, based on the small minority of Facebook users who provided partisan labels in their profiles—but do the resulting findings generalize to individuals who do not divulge their political affiliations? The relatively large sample size in these studies—by social science standards—does not alleviate the concern that the sample is not representative of the population using the platform[51]. This issue is exacerbated by the sometimes large changes in who uses a platform over time (Facebook was once the exclusive domain of Harvard undergraduates), in which case these demographic shifts by themselves affect what happens on the platform.

Other critical problems in generalizability include the fact that different platforms elicit systematically different behaviours. For example, the same person will often behave differently on Facebook and on Twitter[52]. More generally, some human behaviour is highly dependent on the setting—if we could only observe the same people at work, at home or in a religious setting, we might make radically different conclusions about humanity. Generalizability is a function not only of the population, but also of the particular observational contexts. Depending on the research question, this may or may not be a problem. A clearly defined question and population will help to establish how well the measurement lines up with the research intent.

Finally, we note the key measurement question of what are the systematic biases with respect to sampling. Generally, our data collection systems are biased away from minority and, especially, marginal populations; furthermore, our theoretical questions regarding populations typically focus on the middle of the distribution. Representativeness is an issue of transcendent importance in understanding humanity, now and in the past. Consider studies that analyse the text of Google Books (the largest digitized collection of human knowledge), which want to draw conclusions on how linguistic shifts in the texts over the centuries correspond with shifts in, say, national sentiment[4]. This corpus suffers both as a representation of language use, because its composition systematically changed over time (for example, with a much higher representation of scientific texts in the twentieth century)[1] and because even a well-curated set of books will reflect the reality of unrepresentative elites. Not even the largest library ever compiled can cast light on those who—although unrepresented in published texts—still had the ability to act and change the course of history.

These representativeness issues were a major concern in the social science methods of the twentieth century. Reaching respondents through postal mail systematically excludes homeless populations, telephone surveys exclude those without a phone, and surveys conducted in person are subject to people's comfort with and trust in that type of interaction with a stranger.

Observational behavioural streams are potentially subject to similar biases. First, often the instrumentation that collects the data are a consumer good owned by an individual (for example, a mobile phone or a computer), for which costs present a barrier. Second, the instrumentation is often driven by corporate business models aimed at people with money to spend. Third, people more concerned or knowledgeable about privacy matters may be less represented in systems that track behaviours as they opt out of using such services.

However, these data streams have some critical compensating features. Sensor technologies may fill in important data gaps, giving visibility to those who would otherwise be erased from the map. Satellite imagery, for instance, has been used to build indicators of wealth and poverty in the Global South when surveys of household income and consumption do not exist[53]. The banal pervasiveness of modern technology means representation will in many cases be superior to traditional data-collection mechanisms—it is cheaper to own a mobile phone than a home. There are parallels here to the administrative data that W. E. B. Du Bois used to study African-American individuals in the late nineteenth and early twentieth century[54]. The data of an administrative state that enforced racial hierarchy were surely not neutral, yet still had critical value in providing visibility of those most precariously positioned in society.

Furthermore, large sample sizes allow us to look at the behaviour of subsets of the data, for example, minorities (generally construed) and events that are statistically uncommon but consequential (for example, hate speech or misinformation)[49,55,56]. In these cases, sample size and our ability to zoom into smaller populations and infrequent data points matters more than the representativeness of the sample[57]. As Pareto observed long ago, many human behaviours are concentrated in tiny slices of the population[58]; however, twentieth century methods were generally poorly suited to studying that social reality. Perhaps the social theories of the twenty-first century will be able to use micro-level behavioural data to understand how structures of interdependence yield certain macro-level patterns[59].

Access and ethics in measurement

Emerging data streams from sociotechnical systems present two additional challenges, compared with—for example—the data from the Hubble telescope. First, the Hubble telescope is controlled by the scientific establishment, whose goal, presumably, is to answer scientific questions. The institutional goal of a platform, such as Twitter, is clearly not to answer scientific inquiries. The first question is therefore, what can be measured? Second, humans as research participants pose ethical issues that far-away galaxies clearly do not. The question that follows then is what should be measured? We deal with these two questions in turn.

What can be measured varies markedly depending on the system that is generating the data. It is possible to design a small-scale data collection system that relies on consenting participants;[60] however, access to data from millions of people generally requires partnership with a platform. There is a wide spectrum of availability for internet-based communication data with access rules that differ greatly across data holders and time. At the least restrictive end, platforms such as Reddit and Wikipedia allow access to nearly everything the end user can view in machine-readable formats. By contrast, companies such as Facebook and Twitter offer far more restrictive access regimes that are limited by time, data volume and the fact that not all publicly visible data are programmatically accessible. Notably, none of the current major platforms offers individual-level data on what people pay attention to, a remarkably large gap in current internet-based measurements[61]. Furthermore, none of the platforms provide access to information on the extensive randomized control trials (in the form of AB testing) that they do, which could—in principle—enable inferences of the influences of their algorithms on individuals[62]. Generally, any private authority that controls data of interest to researchers can, in the absence of regulation to the contrary, dictate the terms of data access as it chooses. The fact that the actions of platforms such as Twitter and Facebook are a compelling focus for scientific questions of public interest (consider: does a platform amplify the spread of misinformation? What steps does a platform take in response to hate speech?) makes this control deeply problematic[63]. A duty of scholarship in these spaces is to inform public discourse on these important questions. A corollary to the question of what can be measured must be: is it possible to speak truth to power if the power in question controls access to the data used to construct that 'truth'? And, if not, is it (ever) possible to trust any measures that are allowed to be extracted from a given system?

Emerging data sources also pose new ethical challenges. We focus on those that intersect with measurement, and, in particular, on what can and should be measured. More extensive discussions of trace data ethics, as well as alternative models for data access, are available elsewhere[18,64–66]; here we briefly present five particularly pressing concerns. First, although informed consent is a mainstay of research on human participants, anonymous data acquired by third parties are often not considered 'human participant data' and are therefore exempt from review by institutional review boards. What are the ethical obligations of the researcher to consider the circumstances under which the data

Box 2

Data access and ethical issues

Implications of platform control of data access

- Research tools may be rendered obsolete without notice by changes to data access by platforms.
- Private data holders may require external researchers to collaborate closely with them as a condition of data access. Furthermore, products of such collaborations can be subject to review by private data holders before publication. Research under such direct control of platforms cannot be a source of critical insights regarding a given platform.
- Researchers whose work falls outside the scope of interest of the data holders or who are uncomfortable collaborating directly with data holders may be relegated to methods that violate the terms of service of the platforms.
- Examples of incipient models to facilitate access to platform data while maintaining independence of researchers include:

Social Science One: this effort involved external approval and funding of research on aggregate Facebook data that had had differential privacy applied to them[72].

The Facebook 2020 Election Research Project: this project involved the collaboration of external researchers with Facebook, in which the data analysis was performed by Facebook researchers, using pre-registered analysis plans and measures defined by outside experts, who also oversaw the execution of the analyses and had full control over the interpretation of results[77].

Ethical questions

- What are the ethical obligations of researchers to consider the circumstances under which data were collected (for example, through leaks or hacking)?
- How can the research community resolve trade-offs introduced by data anonymization techniques that reduce data utility (for example, by adding noise)?
- What expectations of privacy are reasonable for publicly visible behaviours, such as social media posts?
- How can we manage informational spillovers, in which data collected from consenting individuals reveal insights about others without the knowledge or consent of those people?
- How can we ensure that marginalized populations are adequately and accurately represented in research?

were collected? In a recent example, over 70 gigabytes of data from the far-right social network Parler were publicly released in early January 2021, including GPS-derived location data[67]. Whether researchers can ethically analyse this dataset is a topic of ongoing debate, particularly in light of the use of the website as a planning space for the US Capitol insurrection of 6 January 2021. Far more generally, people are probably unaware of how different systems track them, whether it is through mobility data from phones[68], or browsing data. What then are the ethics of using tracking data from third parties when the targets of that tracking are, at most, nominally aware of that fact?

Second, the level of detail in behavioural datasets means that anonymization that is robust to re-identification efforts is often practically difficult or impossible[69]. It is important to note that de-identified anonymous data can be both the type that cannot be re-identified and the type that can. There have emerged approaches around 'differential privacy' that allow the addition of noise to a dataset that guarantees a degree of anonymity in the data, making it robust to re-identification efforts[70,71]. There is a trade-off, however, because the privacy-enhancing addition of noise diminishes the utility of the data. This was the approach taken in the Social Science One project that provided analytical access to Facebook data[72] (Box 2). One of the struggles confronting the teams granted access was whether the resulting data retained value for answering their questions. (Note: some of the authors are involved in Social Science One and the Facebook 2020 Election Research Project.)

Third, what expectations of privacy are reasonable for publicly visible behaviours, such as tweets? What obligations are incumbent on the researcher to cloak those behaviours? For example, when should researchers avoid mentioning (in publications or presentations) information such as user screen names and complete social media messages, because of the possibility of negative attention or harassment? Some have argued that automatic anonymization of public data may not be the right approach either, rather, content creators should be consulted about their preferences[73].

Fourth, the reliance on the principle of individual autonomy is intrinsically limited, for two reasons. In a world of networked information and insight there will generally be informational spillover from what one person discloses to other individuals. The function of networked media, by definition, is to facilitate interpersonal visibility[74]. An individual who shares their email data, for example, is necessarily providing information from other individuals. The Cambridge Analytica scandal demonstrates the perils of this kind of networked disclosure of information, in which individuals used a Facebook app, which in turn provided access to the behavioural data of the friends of those individuals. However, the risk of informational spillover is a more general principle that is not new with digital trace data: there are almost always potential spillovers from individual disclosure. Genetic data, for instance, potentially shed light on close relatives of an individual;[75] and almost all data about an individual provide information about others. A response from one individual regarding their political preferences provides insights into the preferences of other household members. Knowledge about the drug use of one individual provides insights into the potential drug use of the friends of that individual.

There is also intra-individual informational inference, where information provided (perhaps with consent) enables inferences that the individual may not have anticipated[76]. The practical ethical upshot cannot be that all research for which there is the possibility of informational spillover or inference from disclosure is forbidden; however, it does mean that often there will need to be important limits to data sharing and data visibility. It also highlights the importance of data security.

Building on our discussion regarding 'whom do we measure', care must be taken when attempting to generalize the results of trace-data-based research to populations beyond the platform(s) examined, as well as to the offline lives of the participants[41]. It is essential to find ways to include participants who are digitally underrepresented, especially when such research is used to inform decisions about wide-ranging social or corporate policies.

Conversely, when digital forms of measurement can offer a better representation of marginal groups compared to a traditional twentieth century approach, our ethical obligation should be to use them, as the example of satellite data above highlights. The choice confronting society is not whether digital technologies will be used to measure human behaviour, but when, how and whether anyone outside of corporate or state surveillance will have access to those data. Ideally, large-scale digital data sources would feed into measures that inform nuanced policies and targeted interventions, going beyond one-size-fits-all initiatives, which tend to work less well for minority groups.

Finally, it is a duty of the field to critique decision-making practices that result from problematic measurement procedures. A previously published study, which demonstrated the racial biases of an algorithm used by many hospitals that was driven by errors of measurement, is an excellent example of both the dangers that result from flawed measurement in automated decision-making and the potential for good science to help to rectify those issues[11].

Box 3

Key questions of measurement

What counts?
There are many concepts that we can measure. We must be clear about the ideas, values, priorities and principles that are guiding our selection of research questions and how we frame a topic as worthy of study.

What is the temporal, spatial, structural and cultural integrity of the measure?
Ostensibly similar constructs can be measured in very different ways and the same measure can change over time as the system measuring it changes, or be inconsistent across geographic, demographic, and cultural groups.

Who is counted?
Who opts into the use of various systems such as social media platforms is not random, and nor is what parts of those systems they use and how actively. Accordingly, the trace data of users of such systems that form the basis of many studies may not generalize to the broader population, or even to other, seemingly similar, platforms.

What is accessible to counting?
Science is limited by who is allowed access to what data. The data are limited by the functions, purposes, and protocols of the organizations and technologies that produce them. These limits will never be fully resolved and thus should be a central concern for the field.

What is ethical to count?
Digital data touch many more people than ever before and information can be gleaned about the people not only in a study but around those in the study. This issue of informational spillover is intrinsic in a networked world for all research, which undermines the foundational principle of individual autonomy in current ethical frameworks for research, and greatly increases the duties of researchers to maintain data security. The standards and practices for consent, privacy and confidentiality must take these realities into account.

Outlook

Box 3 summarizes the essential arguments of this paper. The massive instrumentation of global society has enormous potential to transform our understanding of the social world. However, the revolution in instrumenting human behaviour requires a revolution in the measurement of human behaviour. Any new measurement regime needs to match the possibilities of both old and new theories of society, deal with the essential instability of human measurement within these heavily instrumented sociotechnical systems, and develop a new model of ethical research of human participants that balances individual rights and collective benefits.

1. Pechenick, E. A., Danforth, C. M. & Dodds, P. S. Characterizing the Google Books Corpus: strong limits to inferences of socio-cultural and linguistic evolution. *PLoS ONE* **10**, e0137041 (2015).
2. Dietrich, B. J., Hayes, M. & O'Brien, D. Z. Pitch perfect: vocal pitch and the emotional intensity of congressional speech. *Am. Polit. Sci. Rev*. **113**, 941–962 (2019).
3. Dietrich, B. J. Using motion detection to measure social polarization in the U.S. House of Representatives. *Polit. Anal*. **29**, 250–259 (2021).
4. Michel, J.-B. et al. Quantitative analysis of culture using millions of digitized books. *Science* **331**, 176–182 (2011).
In this study, 4% of all books that have been published were digitized and used to examine changes in phonology, word use and the adoption of new technologies over long periods of time.
5. Merton, R. K. in *Social Theory and Social Structure* 39–72 (Free Press, 1968).
6. Watts, D. J. *Everything Is Obvious: Once You Know the Answer* (Crown Business, 2011).
7. Simon, H. A. Bandwagon and underdog effects and the possibility of election predictions. *Public Opin. Q*. **18**, 245–253 (1954).
8. Mutz, D. C. *Impersonal Influence in American Politics* (Cambridge Univ. Press, 1998).
9. Westwood, S. J., Messing, S. & Lelkes, Y. Projecting confidence: how the probabilistic horse race confuses and demobilizes the public. *J. Polit*. **82**, 1530–1544 (2020).
10. O'Neil, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Crown, 2016).
11. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
12. Landsberger, H. A. *Hawthorne Revisited* (The New York State School of Industrial and Labor Relations, 1958).
13. Mayo, E. *The Human Problems of an Industrial Civilization* (Routledge, 2004).
14. Lazer, D., Kennedy, R., King, G. & Vespignani, A. The parable of Google Flu: traps in big data analysis. *Science* **343**, 1203–1205 (2014).
This paper shows that the increasing over-prediction of flu prevalence of Google Flu Trends was largely the result of changes to Google's search algorithm, which altered the terms that people used to find flu-related information.
15. Brunton, F. & Nissenbaum, H. *Obfuscation: A User's Guide for Privacy and Protest* (MIT Press, 2015).
16. Davis, D. W. The direction of race of interviewer effects among African-Americans: donning the Black mask. *Am. J. Pol. Sci*. **41**, 309–322 (1997).
17. American National Election Studies. *1978 Time Series Study* https://electionstudies.org/wp-content/uploads/2018/03/anes_timeseries_1978_qnaire_post.pdf (1978).
18. Salganik, M. J. *Bit by Bit: Social Research in the Digital Age* (Princeton Univ. Press, 2017).
19. Patty, J. W. & Penn, E. M. Analyzing big data: social choice and measurement. *PS Polit. Sci. Polit*. **48**, 95–101 (2015).
20. Kraemer, M. U. G. et al. The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science* **368**, 493–497 (2020).
21. Jia, J. S. et al. Population flow drives spatio-temporal distribution of COVID-19 in China. *Nature* **582**, 389–394 (2020).
22. Badr, H. S. et al. Association between mobility patterns and COVID-19 transmission in the USA: a mathematical modelling study. *Lancet Infect. Dis*. **20**, 1247–1254 (2020).
23. Munger, K. The limited value of non-replicable field experiments in contexts with low temporal validity. *Soc. Media Soc*. **5**, 1–4 (2019).
24. Deaton, A. & Cartwright, N. Understanding and misunderstanding randomized controlled trials. *Soc. Sci. Med*. **210**, 2–21 (2018).
25. Vraga, E. K., Bode, L., Smithson, A.-B. & Troller-Renfree, S. Accidentally attentive: comparing visual, close-ended, and open-ended measures of attention on social media. *Comput. Human Behav*. **99**, 235–244 (2019).
26. Guess, A., Munger, K., Nagler, J. & Tucker, J. How accurate are survey responses on social media and politics? *Polit. Commun*. **36**, 241–258 (2019).
27. Aleta, A. et al. Modelling the impact of testing, contact tracing and household quarantine on second waves of COVID-19. *Nat. Hum. Behav*. **4**, 964–971 (2020).
28. Echeverría, J. et al. LOBO: evaluation of generalization deficiencies in Twitter bot classifiers. In *Proc. 34th Annual Computer Security Applications Conference* 137–146 (ACM, 2018).
29. Ferrara, E., Varol, O., Davis, C., Menczer, F. & Flammini, A. The rise of social bots. *Commun. ACM* **59**, 96–104 (2016).
30. Hughes, A. G. et al. Using administrative records and survey data to construct samples of Tweeters and Tweets. *Public Opin. Q*. https://doi.org/10.1093/poq/nfab020 (2021).
31. Napoli, P. M. *Audience Evolution: New Technologies and the Transformation of Media Audiences* (Columbia Univ. Press, 2011).
32. Yang, T., Majó-Vázquez, S., Nielsen, R. K. & González-Bailón, S. Exposure to news grows less fragmented with an increase in mobile access. *Proc. Natl Acad. Sci. USA* **117**, 28678–28683 (2020).
This study tracked the news consumption of users across mobile and desktop devices and found that most individuals do not self-sort their news consumption by partisanship but, instead, consume news from a diversity of sources including partisan and nonpartisan ones.
33. Haythornthwaite, C. Exploring multiplexity: social network structures in a computer-supported distance learning class. *Inf. Soc*. **17**, 211–226 (2001).
34. Campbell, K. E. & Lee, B. A. Name generators in surveys of personal networks. *Soc. Netw*. **13**, 203–221 (1991).
35. Wagner, C. Measuring algorithmically infused societies. *Nature* https://doi.org/10.1038/s41586-021-03666-1 (2021).
36. Healy, K. The performativity of networks. *Eur. J. Sociol*. **56**, 175–205 (2015).
37. Rahwan, I. et al. Machine behaviour. *Nature* **568**, 477–486 (2019).
38. Neuendorf, K. A. *The Content Analysis Guidebook* (Sage, 2017).
39. Davidov, D., Tsur, O. & Rappoport, A. Semi-supervised recognition of sarcasm in Twitter and Amazon. In *Proc. 14th Conference on Computational Natural Language Learning* 107–116 (Association for Computational Linguistics, 2010).
40. Groves, R. M. Nonresponse rates and nonresponse bias in household surveys. *Public Opin. Q*. **70**, 646–675 (2006).
41. Hargittai, E. Potential biases in big data: omitted voices on social media. *Soc. Sci. Comput. Rev*. **38**, 10–24 (2020).
Using survey data, this study finds that younger, wealthier and more technically skilled people tend to use social media and that there were substantial gender and education differences in which platforms people used.
42. Lazer, D. & Radford, J. Data ex machina: introduction to big data. *Annu. Rev. Sociol*. **43**, 19–39 (2017).
43. Correa, T. & Valenzuela, S. A trend study in the stratification of social media use among urban youth: Chile 2009–2019. *J. Quant. Descr. Digit. Media* **1**, https://doi.org/10.51685/jqd.2021.009 (2021).
44. Mellon, J. & Prosser, C. Twitter and Facebook are not representative of the general population: political attitudes and demographics of British social media users. *Res. Polit*. **4**, 1–9 (2017).

45. Beisch, N. & Schäfer, C. Internetnutzung mit großer Dynamik: Medien, Kommunikation, Social Media. *AS&S* https://www.ard-werbung.de/media-perspektiven/fachzeitschrift/2020/detailseite-2020/internetnutzung-mit-grosser-dynamik-medien-kommunikation-social-media/ (2020).
46. Hargittai, E. & Litt, E. The Tweet smell of celebrity success: explaining variation in Twitter adoption among a diverse group of young adults. *New Media Soc*. **13**, 824–842 (2011).
47. Henrich, J., Heine, S. J. & Norenzayan, A. Most people are not WEIRD. *Nature* **466**, 29 (2010).
48. Wang, W., Rothschild, D., Goel, S. & Gelman, A. Forecasting elections with non-representative polls. *Int. J. Forecast*. **31**, 980–991 (2015).
49. Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B. & Lazer, D. Fake news on Twitter during the 2016 U.S. presidential election. *Science* **363**, 374–378 (2019).
50. Bakshy, E., Messing, S. & Adamic, L. A. Exposure to ideologically diverse news and opinion on Facebook. *Science* **348**, 1130–1132 (2015).
51. Meng, X.-L. Statistical paradises and paradoxes in big data (I): law of large populations, big data paradox, and the 2016 US presidential election. *Ann. Appl. Stat*. **12**, 685–726 (2018).
52. Hargittai, E., Füchslin, T. & Schäfer, M. S. How do young adults engage with science and research on social media? Some preliminary findings and an agenda for future research. *Soc. Media Soc*. **4**, 1–10 (2018).
53. Blumenstock, J. Don't forget people in the use of big data for development. *Nature* **561**, 170–172 (2018).
54. Battle-Baptiste, W. & Rusert, B. (eds) *W. E. B. Du Bois's Data Portraits: Visualizing Black America* (Princeton Architectural Press, 2018).
55. Siegel, A. A. et al. Trumping hate on Twitter? Online hate speech in the 2016 US election campaign and its aftermath. *Quart. J. Polit. Sci*. **16**, 71–104 (2021).
56. Allen, J., Howland, B., Mobius, M., Rothschild, D. & Watts, D. J. Evaluating the fake news problem at the scale of the information ecosystem. *Sci. Adv*. **6**, eaay3539 (2020).
57. Foucault Welles, B. On minorities and outliers: the case for making big data small. *Big Data Soc*. **1**, 1–2 (2014).
58. Newman, M. E. J. Power laws, Pareto distributions and Zipf's law. *Contemp. Phys*. **46**, 323–351 (2005).
59. González-Bailón, S. *Decoding the Social World: Data Science and the Unintended Consequences of Communication* (MIT Press, 2017).
60. Stopczynski, A. et al. Measuring large-scale social networks with high resolution. *PLoS ONE* **9**, e95978 (2014).
61. Lazer, D. Studying human attention on the Internet. *Proc. Natl Acad. Sci. USA* **117**, 21–22 (2020).
62. Aral, S. & Eckles, D. Protecting elections from social media manipulation. *Science* **365**, 858–861 (2019).
63. Puschmann, C. & Burgess, J. The politics of Twitter data. *HIIG Discussion Paper Series No. 2013-01* http://www.ssrn.com/abstract=2206225 (2013).
64. Chen, W. & Quan-Haase, A. Big data ethics and politics: toward new understandings. *Soc. Sci. Comput. Rev*. **38**, 3–9 (2020).
65. Breuer, J., Bishop, L. & Kinder-Kurlanda, K. The practical and ethical challenges in acquiring and sharing digital trace data: negotiating public–private partnerships. *New Media Soc*. **22**, 2058–2080 (2020).
66. Zook, M. et al. Ten simple rules for responsible big data research. *PLOS Comput. Biol*. **13**, e1005399 (2017).
67. Greenberg, A. An absurdly basic bug let anyone grab all of parler's data. *Wired* (12 January 2021).
68. Valentino-DeVries, J., Singer, N., Keller, M. H. & Krolik, A. your apps know where you were last night, and they're not keeping it secret. *The New York Times* https://www.nytimes.com/interactive/2018/12/10/business/location-data-privacy-apps.html (10 December 2021).
69. Sweeney, L. Simple demographics often identify people uniquely. *Privacy Working Paper 3* https://dataprivacylab.org/projects/identifiability/paper1.pdf (Carnegie Mellon University, 2000).
Using census data, this paper shows that 87% of the US population could be uniquely identified by date of birth, postal code and gender; demonstrating the ease with which study respondents can be re-identified from ostensibly anonymous data.
70. Wood, A. et al. Differential privacy: a primer for a non-technical audience. *Vanderbilt J. Entertain. Technol. Law* **21**, 209–276 (2019).
71. Dwork, C. & Roth, A. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci*. **9**, 211–407 (2013).
72. King, G. & Persily, N. A new model for industry–academic partnerships. *PS Polit. Sci. Polit*. **53**, 703–709 (2020).
73. Bruckman, A., Luther, K. & Fiesler, C. in *Digital Research Confidential: The Secrets of Studying Behavior Online* (eds Hargittai, E. & Sandvig, C.) 243–258 (MIT Press, 2015).
74. Marwick, A. E. & boyd, d. Networked privacy: how teenagers negotiate context in social media. *New Media Soc*. **16**, 1051–1067 (2014).
75. Bieber, F. R., Brenner, C. H. & Lazer, D. Finding criminals through DNA of their relatives. *Science* **312**, 1315–1316 (2006).
76. Zheleva, E. & Getoor, L. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In *Proc. 18th International Conference on World Wide Web* 531–540 (2009).
77. Miller, G. As U.S. election nears, researchers are following the trail of fake news. *Science* (26 October 2020).
78. Merton, R. K. The self-fulfilling prophecy. *Antioch Rev.* **8**,193–210 (1948).

Acknowledgements D.L. acknowledges support from the William & Flora Hewlett Foundation. E.H. acknowledges support from the Alfred P. Sloan Foundation.

Author contributions All authors contributed to the concepts, writing and/or revisions of the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to D.L.

Peer review information *Nature* thanks Paramveer Dhillon and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.