Original Research Article

BIG DATA & SOCIETY

Big Data & Society
July–December 2014: 1–11

DOI: 10.1177/2053951714563194
bds.sagepub.com
SAGE

Big Data and the brave new world of social media research

Ralph Schroeder

Abstract

The recent Facebook study about emotional contagion has generated a high-profile debate about the ethical and social issues in Big Data research. These issues are not unprecedented, but the debate highlighted that, in focusing on research ethics and the legal issues about this type of research, an important larger picture is overlooked about the extent to which free will is compatible with the growth of deterministic scientific knowledge, and how Big Data research has become central to this growth of knowledge. After discussing the 'emotional contagion study' as an illustration, these larger issues about Big Data and scientific knowledge are addressed by providing definitions of data, Big Data and of how scientific knowledge changes the human-made environment. Against this background, it will be possible to examine why the uses of data-driven analyses of human behaviour in particular have recently experienced rapid growth. The essay then goes on to discuss the distinction between basic scientific research as against applied research, a distinction which, it is argued, is necessary to understand the quite different implications in the context of scientific as opposed to applied research. Further, it is important to recognize that Big Data analyses are both enabled and constrained by the nature of data sources available. Big Data research is bound to become more widespread, and this will require more awareness on the part of data scientists, policymakers and a wider public about its contexts and often unintended consequences.

Keywords

Big Data, social science, social networking sites, epistemology, research ethics, privacy

Introduction

A recent study by Kramer et al. (2014) has caused much debate in the media and among researchers, arguably more so than previous studies, about using Big Data (a set of sources and reactions can be found in Grimmelman, 2014). In the study, Facebook changed the newsfeed of almost 700,000 users, dividing them into two randomly selected different groups. Over the course of one week in January 2012, the newsfeeds of the two groups were changed such that the positive emotional content of friends was reduced for one group and the negative content reduced for the other. Then, the experiment collected the positive and negative words produced by the persons who had been exposed to these two types of reduced content (122 million words in 3 million posts), to see whether these words were more negative or positive.

The experiment showed that those groups who were shown more positive words also then posted more positive words, and the same for negative words. In other words, the emotional state of users can be influenced by the words of other users – a form of 'emotional contagion' which takes place unbeknown to the user. While the implications of the study have been much discussed, it can be predicted that ultimately this debate about the uses of Big Data (again, like previous ones) will die down. One reason for this is that it is not clear what is new here: surely companies have been doing this kind of research to manipulate people's behaviour for some time, and academic researchers such as psychologists have been carrying out experiments to change people's behaviour (though, unlike in

Oxford Internet Institute, University of Oxford, Oxford, UK

Corresponding author:
Ralph Schroeder, Oxford Internet Institute, University of Oxford, 1 St Giles, Oxford OX1 3JS, UK.
Email: ralph.schroeder@oii.ox.ac.uk

this study, they have gained prior consent from participants). Is not that what a lot of advertising and marketing research does already – changing people's minds about things? Beniger (1986), for example, documents the rise of scientific methods and psychological techniques going back to the middle of the 20th century, and Porter (2008) traces how academic social science and commercial research have been much closer during various periods in the past, also in terms of data gathering. And do not researchers sometimes deceive subjects in experiments about their behaviour? Once the debate peters out, we will be left asking the next time: So what is new?

This paper will argue that this 'nothing new' way of thinking about Big Data research has a serious defect, because there are in fact several issues raised by this research, only some of them new. The new issues revolve around how more powerful knowledge is used in our everyday lives, and we will come to them shortly. However, debates of this type tend to concentrate, understandably, on two more narrow and immediate issues: the first is whether Facebook (in this case) has done anything illegal in carrying out this research and using findings, and the second is about the ethics of research, in this case also involving academic researchers. I will argue that these two issues are not new, whereas the greater powerfulness of knowledge is. The paper will proceed in three steps: the first is to illustrate through Facebook's 'emotional contagion' study what is new and not new about this kind of research. Next, it will address the broader question of how to define Big Data, and data, and in this way identify what new issues are raised in this kind of research, and in areas of rapid scientific advance generally. These definitions will allow us to locate the newness of Big Data research in new sources of readily computationally manipulable data. In the third and final section, it will therefore be possible to pinpoint the possibilities, dangers and above all the limits of how this knowledge can be used. This will also make it possible to make some suggestions for policy.

Legal, ethical and social issues in the Facebook 'emotional contagion' study

Let us turn first to the legality of the study. As the authors correctly point out, the research falls within Facebook users giving informed consent when they sign up to the service (although the clause allowing Facebook to do 'research' as opposed to 'data analysis', as a number of commentators have pointed out (Yarkoni, 2014), was added after the study was carried out). Laws or regulation may be required here to prevent this kind of informed consent by means of a service agreement, which has also been criticized in many other instances. However, new regulations may also be difficult, since it will be hard to draw a line between this experiment and other forms of manipulating people's responses to media. The main constraint in this case is that Facebook may not want to lose users due to negative reactions on their part. Or, as the first author of the study, Adam Kramer (who works for Facebook), acknowledged in a blog post response to the outcry, for users to be manipulated in an experiment via Facebook may 'cause anxiety' (Kramer, 2014), which led him to apologize and promise more care in future studies. In short, this type of research may be bad for business, and hence Facebook may abandon this kind of research (but we will come back to this point). In any event, companies using techniques that users do not like – so that they are forced to change course – is not new.

The legal issue is a moving target: apart from a media outcry, the Facebook study has resulted in legal and regulatory challenges, and the mounting of two such challenges has been mentioned in the media. One is by the Electronic Privacy Information Center, an organization based in Washington (Guardian, 2014a), and one in the UK by the Information Commissioner's Office (ICO) (BBC, 2014). The study will continue to be much discussed (for example, a special issue of the journal *Research Ethics* (Hunter and Evans, 2014) will be devoted to the Facebook study). However, it should be noted that this paper addresses the larger question of how studies using Big Data manipulate users (and the Facebook study is discussed as a key example of many similar studies: for recent critical perspectives on Big Data research generally, see the special issue of the journal *Surveillance and Society*, edited by Andrejevic and Gates (2014), and the new journal *Big Data and Society*). These studies are ongoing and will continue – whatever the outcome of these legal and regulatory wrangles – for the foreseeable future.

The second issue concerns academic research ethics. In addition to the first author (Kramer) who works at Facebook, this study was carried out by two academic researchers. In retrospect, it is hard to see how this study would have received approval from an institutional review board (IRBs, as they are known in the USA), the boards at which academic institutions check the ethics of studies. Here we can briefly trace the media flurry in *The Guardian* newspaper: a list of reasons why the study did not meet ethical approval for academic research, by the British Psychological Society's ethics committee, was published as a letter in this newspaper (Guardian, 2014b). The IRB guidelines at Cornell are available on the university website (Cornell, 2014). However, these rules do not apply since, as pointed out in a statement by the Cornell University Media Relations Office (2014), the

academics 'did not participate in data collection and did not have access to user data' and 'the research was conducted independently by Facebook', so that 'no review . . . was required' at Cornell. This separation calls attention to the thin line between academic and commercial research (it can be mentioned that one of the academic researchers, Jamie Guillory, had moved to the University of California, San Francisco by the time of publication, while the other author, Jeffrey Hancock, was at Cornell throughout, thus further complicating the matter). In any event, the media flurry reached its climax with Facebook chief executive Sheryl Sandberg apologizing, not for conducting the study, but for how it was 'communicated' (Guardian, 2014c).

Perhaps stricter guidelines are needed in relation to academic research too since, firstly, Big Data research is becoming much more widespread in the social sciences and is often based on data from social media like Facebook, Twitter and mobile phone data. Secondly, much – though not all (consider Wikipedia) – of this research entails close relations between academics and social media companies who provide access to this data, and to being able to experiment with the platforms (as here). Again, the ethics of academic research may need to be tightened up to provide new guidelines for academic collaboration with commercial platforms, especially, as in this case, the line between what part of the research was Facebook's responsibility and what part was to be covered by academic ethics review was not clear.

However, this was not the only response, since a number of prominent academics and ethicists defended the Facebook experiment, arguing that the study 'was not an egregious breach of either ethics or law' (Meyer, 2014) and that its value to social science was such as to override concerns. More importantly, the authors argued that restrictions on this type of research would only 'drive social trials underground', making them more pernicious. But this kind of controversy over research ethics and the weighing of rights versus consequences is not new either.

The third issue, which is the new and important one, is the increasing power that social research using Big Data has over our lives. This is rather more difficult to pin down than the first two points. Where does this power come from? Any attempt to pin this down must begin with a definition of Big Data, which can be defined as having access to data of a scale and scope that is a leap or step change from what was available before, and to be able to perform computational analysis on these data (Schroeder, 2014). This definition will be elaborated in more detail below, but it clearly applies in this case, as in other cases we have documented (Schroeder, 2014; Taylor et al., 2014): as mentioned, almost 700,000 users' Facebook newsfeeds were changed in order to perform this experiment, and more than 3 million posts containing more than 122 million words were analysed.

What is important here are the implications of the power that has been gained by means of this new knowledge. To be sure, as the authors point out, this study was valuable for social science in showing, among other things, that emotions may be transmitted online via words, not just in face-to-face situations. The study was also important because it argued that the results were contrary to the claims of a previous study (Turkle, 2011). Secondly, even if, as has been pointed out, the effect size was small in this sample so that its academic value is in doubt, the study nevertheless provides Facebook with knowledge that it can use to manipulate users' moods; for example, making their moods more positive so that users will come to its – rather than a competitor's – website. We will come back to this point, but it can be noted that in this case, it may not matter to Facebook how significant or otherwise the effect size is, as long as some impact is made. (As an aside, it can be noted that the published study claims an important and significant effect, while the mea culpa by Kramer in response to the outcry downplayed the importance and effect of the study!). Put differently, Facebook (or other companies and organizations engaged in this kind of research) may not be as concerned with effect size as researchers are. Moreover, companies or other organizations can of course increase the scale of the manipulation as much as they wish in making use of this research, and even a small effect in, for example, increasing the number of users or the time spent on the site will be a benefit for the company. As the authors of the study themselves note, 'given the massive scale of social networks such as Facebook, even small effects can have large consequences' (Kramer et al., 2014: 8790). All this is to say that social science knowledge, produced in collaboration with academic social scientists, in this case enables companies to manipulate people's hearts and minds (it should be noted that 'manipulate' is used here in a neutral sense, in line with the definitions that will be provided below, equivalent to saying 'to change what people do' or changing the physical environment, which can of course be negative or positive).

This is not the Orwellian world of surveillance by means of phone tapping, which has been much in the news recently due to the revelations by Edward Snowden, and that may come to mind in this case too. Rather, it will be argued that it is more relevant to invoke Huxley's Brave New World, where companies and governments are able to play with people's minds, and do so in a way such that users, knowingly or unknowingly (and it may not be easy to tell the

difference), come to accept and embrace this: after all, who would not like to have their experience on Facebook improved in a positive way? Or, to use a different example, who could object to being 'nudged' online by government to feel better about paying taxes?

It is important to point out different implications of Orwell and Huxley in the light of what follows, because the distinction between top-down surveillance as against a process in which users are implicated, possibly unwittingly, and they slowly adapt to new practices, is at the core of what will be argued about scientific advance below. In the Facebook case, the difference between Orwell and Huxley is highlighted by one reply to criticisms of the study: the motivation of the research is the effort to improve user experience, as Kramer (2014) says in his blogpost. Similarly, according to *The Guardian* newspaper, 'A Facebook spokeswoman said the research ... was carried out "to improve our services and to make the content people see on Facebook as relevant and engaging as possible"' (Guardian, 2014d). Yet, improving experience and services could also just mean selling more products, or manipulating people's political behaviour.

These potential implications are worrisome, and academic social scientists may want to think twice before producing knowledge that supports this kind of impact. But again, we cannot pinpoint this impact without understanding what is new: Big Data is a leap in how data can be used to manipulate people in more powerful ways. This point has been lost by those who criticize Big Data mainly on the grounds of the epistemological conundrums involved (as with boyd and Crawford's (2012) widely cited paper), which argues that Big Data studies often lack scientificity (this is the thrust of the first four of their six 'provocations', and especially the second, that 'claims to objectivity and accuracy are misleading', where they draw on Bruno Latour). No: as we shall see, it is precisely because knowledge is more scientific, more objective and more powerful that it enables more manipulation. Second, the view taken here also departs from Savage and Burrows' (2007, 2009) – again much cited – papers, which argued that private sector companies have more powerful means to advance social science than academics with Big Data, because of greater access to data and tools. Again, this is somewhat misleading since, for one, academics can collaborate with and often use the data provided by companies (though this is not to deny that there are often issues about data access), and second, not all data, including Big Data, are proprietary (again, Wikipedia is a good example) and academic social science may also thrive with the proliferation of these and other non-proprietary data sources, and third, it may be that social science will make more cumulative advances, since commercial efforts are likely to be more applied, and hence restricted to more narrow and non-replicable findings (as we shall see later). For this reason, it is crucial to identify the point or points at which a stop should be put on the slippery slope of increasing manipulation of our behaviours, in both academic and commercial research. Further, there need to be more efforts to specify when access to Big Data on a new scale enables research that affects many people without their knowledge, and to regulate this type of research – at a minimum making it transparent when such research is being carried out.

This brings us back to an earlier point: it is true that Facebook may stop this kind of research for fear of losing customers, but how would we know? Defenders of the study have argued that this research would merely be 'driven underground' (discussed a moment ago) and have noted that the furore over this study will mean that these types of studies are not published. For example, the psychologist Yarkoni (2014) blogged that 'by far the most likely outcome of the backlash Facebook is currently experiencing is that, in future, its leadership will be less likely to allow its data scientists to publish their findings in the scientific literature'. Whether published or not, in this case, academics participated in research that potentially encourages the use of data for a kind of Huxleyan conditioning, as in *Brave New World*. Whatever the legal implications and the issues connected to academic research ethics, it behooves us to consider where social research using Big Data is leading, including when people may come to like the more positive reinforcement of their online behaviours that this research is enabling. Or, in cases where they are not aware of the manipulation, there should be greater transparency and awareness about how these manipulations take place.

The counterargument against this highlighting of the novelty of the study and its implications is that, as already mentioned, despite the fact that this is one of the largest ever manipulations of human subjects on this scale, the 'effect size' was in fact small. Yet as we have seen, this argument does not apply to commercial or other applications of this research. The second point is at once more difficult and more important, which is that despite my argument that this research is novel, there is no single impact of this study or even of several studies of this type. Instead, the problems and the benefits of this kind of research are part of a constantly moving and multi-faceted research front and of how research is being used, and how the effects of this research are therefore cumulative. Put differently, there is simply 'creep'. As with the impact of video games, where there have been constant debates about their impact on violence and the like, Big Data social

research will have a series of scandals or debates, even as the process of engaging in this type of experimentation and the use of this type of research will continue to expand. There is a difference from video games, however, insofar as playing video games is a choice and the effects are transparent to the user. Facebook's experiment and similar research are carried out without the users' awareness.

It is worth considering for a moment why the surreptitious manipulation of our thoughts and feelings is regarded as offensive. Reactions to the Facebook study ranged from outrage to resignation (see, for example, the comments in response to one of the first newspaper stories, Guardian, 2014e). From a social science perspective, however, gauging the implications of this manipulation entails assessing how many people use new media technology which engages in this type of manipulation, and to what ends they were being manipulated. This type of assessment is very difficult since it would require aggregating the effects of the uses of this type of research across the range of its uses in practice. However, a different approach is to note that social media are mainly used for entertainment and consumption, and the manipulation of users by new media companies aims mostly at increasing audience (or advertising) share and online purchases. What is new here is the way in which audience or customer experiences can be manipulated on an unprecedented scale and with unprecedented accuracy. Politically, the main potential uses of social media could be when authoritarian regimes (like China) make use of these techniques in order to browbeat or mollify their populations into subjection. And again, although the scope of these new practices is difficult to gauge, the obvious point is that more and more of our lives are spent online, and the vast bulk of this activity consists of the use of social media which engage in this type of research and manipulation, including searching for information (Google), connecting via social networking sites (Facebook) and communicating with each other (Gmail, Twitter).

Again, it is worth considering that it is the powerfulness of knowledge, not the number of users, that is important for the implications. One way to highlight this point is by comparing the Google+ network with Facebook. As has been pointed out in a newspaper article, despite far fewer users (especially active users), in fact, Google+ may have more powerful knowledge about its users since it can link various data sources about what information users are seeking via the Google search engine, what they are watching on (Google owned) YouTube, what messages they are sending via Gmail, in addition to how they use Google+ social network. The difference is that Facebook, despite far more users, is a single platform: 'Google Plus may not be much of a competitor to Facebook as a social network, but . . . some analysts . . . say that Google understands more about people's social activity than Facebook does' (New York Times, 2014: A1). Perhaps we cannot know which of these two, or any other services, has more powerful knowledge about its users: but this unknowability also reinforces the point being made here, that part of what worries users of these services is that they do not know how much is known about them (unless they go to extraordinary efforts).

What is Big Data?

Before analysing the newness of this type of research by providing definitions, it is worth briefly looking at the wider context. The Facebook 'emotional contagion' study has been discussed is one example; a number of others could have been used (Bond et al., 2012; see also Golder and Macy, 2014, for an overview of both experimental and other Big Data studies in the social sciences). And most commercial research of this type is not reported in academic publications, as in this case. This is one reason why data generally, and Big Data more specifically, has provoked some major and still ongoing debates. The issues relating to privacy and data protection are too well known to reiterate here (Brown and Marsden, 2013: 47–68; Rule, 2007). Recently, this debate has shifted from privacy in general to debates about data, and more recently still specifically digital data. Briefly, privacy and data protection laws are established to safeguard and ensure individuality and autonomy in society. There is currently much ongoing policymaking about data (for example, the 'right to be forgotten' in Europe, see Guardian, 2014f) and the White House review on Big Data (White House, 2014). Thus, privacy and data protection laws are spreading and being adopted around the world (currently in more than 100 countries, Greenleaf, 2014). Greenleaf (2013: 213) argues that 'the effectiveness of data privacy principles comes as much from their ideological effect and their global nature as from their enforcement (which is often lacking). These are more important in terms of establishing guidelines than in implementation'.

There are no definitive, academic definitions of data and of Big Data. Gartner (2014), the consultancy firm, defines Big Data in terms of three Vs: high volume, high velocity and high variety. But this definition does not work for all Big Data studies: some studies of Wikipedia which clearly use Big Data do not meet the criterion of 'velocity' since the data is not produced quickly in real time but is instead static (compared to, say, Twitter, which produces a lot of data quickly). Further, some Wikipedia studies also do not use a

'variety' of data, but instead use data from one or few dimensions (examples include Yasseri et al., 2012 or Moat et al., 2013). However, since specifying what is new about data-driven research is crucial for understanding its implications, 'Big Data' can be defined here as research that represents a step change in the scale and scope of knowledge about a given phenomenon (Schroeder, 2014). Note that this definition does not rely on 'size' per se, but on size in relation to a given object or phenomenon being investigated – where there are so many data points that previously collecting and analysing these data on a sufficiently large scale was difficult, impractical or impossible – and how Big Data research advances beyond previous research about this type of object.

But what is 'data'? In terms of science or valid and objective knowledge, data has three characteristics: First, data belongs to (in the ontological, not legal sense) the object or phenomenon under investigation; it is material collected about the research object. Second, data exists prior to analysis: as Hacking (1992) puts it, the view that 'all data are of their nature interpreted' is misleading: 'data are made, but as a good first approximation, the making and taking come before interpreting' (p. 48). He adds, 'it is true that we reject or discard putative data because they do not fit an interpretation, but that does not prove that all data are interpreted' (p. 48). He also distinguishes data from other related parts of the scientific process, such as the calibration of instruments for data measurement. And third, data is the most divisible or atomized useful unit of analysis.

Apart from pinpointing how digital Big Data is novel, this definition of data has implications for how advance in social science can be gauged, and presumes a realist and pragmatist epistemology (Hacking, 1983) because the definition requires that there is an object 'out there' (realism) about which more useful or powerful knowledge has been gained (pragmatism). Hacking (1983) defines science as the 'adventure of the interlocking of representing and intervening' (p. 146); again, a pragmatist and realist account of the relation between scientific knowledge and the physical or natural worlds. I (Schroeder, 2007: 9) have developed Hacking's ideas by arguing that technology is 'the adventure of the interlocking of refining and manipulating' the world by means of physical instruments or tools. With these definitions, it can be recognized that more powerful tools (for example, computational power) have become available in relation to large-scale and readily manipulable sources of data, thus linking the advance of tools to the increased availability of data sources that are suited to these tools.

These are philosophical ideas about what scientific knowledge and technologies do, or how they provide knowledge about and change the world. The key here is that they provide insight into the implications of data-driven knowledge: a 'realist' conception regards data as becoming available from a source out in the world on a scale that is different from what was available before about similar objects. Here, we can think, as concrete examples, about the data we have about social interactions on Twitter or Facebook or Wikipedia, in the case where all data or large samples (which may or may not have issues about representativeness, though these biases can also be assessed: see, for example, in relation to Twitter, González-Bailón et al., 2014) about these platforms are available, and how this compares with data that is available about landline telephone records, or data about television watching, or about physical letters and their contents and senders and receivers.

There are several consequences of this view of science and data for the nature and uses to which different types of knowledge are put. More powerful 'representing' entails a greater grasp of the phenomenon, and 'intervening' takes place typically in relation to trying to make changes in the natural – or here, in the social – world. For Big Data research, the 'world' of the phenomenon that is intervened in consists of digital platforms or people's digital traces. Here we can note that researchers typically do not have the possibility of intervening in these digital platforms – unless they control the environment from which digital data is gathered. Yet, this is precisely the case with the Facebook emotional contagion study (or, take another example, the Facebook voting experiment (Bond et al., 2012), which experimented with different Facebook messages urging people to vote in the American elections, where, similarly, a small but important effect was found). Technology, for example for manipulating the physical world, needs to control phenomena if more powerful knowledge is to be applied, and with Big Data research, there are many cases where the phenomenon that has been investigated can subsequently be changed to influence user behaviour. On the other hand, if academic researchers do not control these tools and environments, such manipulation is not possible. The power of Big Data research, at least in an academic context, derives from its scientificity, and the possibilities of making advances in understanding phenomena without necessarily controlling them in practice (changing experimental conditions, as in the Facebook emotional contagion study, straddles this divide, but only for the 'laboratory population' within the limits of the academic study – applying the findings comes later). The manipulation of these phenomena is thus a more practical, applied exercise, a more powerful exercise of control over specific parts of the physical or social world for certain purposes (for example, changing Facebook text to increase audience share).

It will be evident from these considerations that quite different possibilities attach to academic and commercial research. Academic social scientists are engaged in research in order to generate generalizable knowledge about human behaviour, not (for the most part) to change it. Working in the private sector or in other applied settings, however, researchers and those who use this knowledge (like marketers and advertisers) will want to do so. Thus, the uses of Big Data for specific applications, influencing the behaviours of people, are not neutral, even if the knowledge generated for these purposes is neutral. Knowledge using digital data applies to human beings treated as abstract material governed by certain statistical regularities, while knowledge generated for use in technological platforms to influence behaviours is much more bound to the context of particular times, places, populations and purposes. There is thus a divide between the uses of Big Data in academic or scientific analyses as against the uses of Big Data in commercial, government and other applied settings: in academic research and science, Big Data is used to generate abstract knowledge, without prescriptiveness about how to use this knowledge to change behaviour. In applied settings, the reverse is true: knowledge is generated inasmuch as it can be used to change behaviour.

This point can be related directly to the definition of data that has been used here. In settings where data is not obtained from 'raw' sources (the physical world), it is nevertheless treated 'as if' it were raw (in relation to human behaviour). Consider, to take a different example apart from Facebook, Twitter data: when tweets are analysed, this is typically done by counting word frequencies or messages sent between accounts 'as if' these were units without context. That is, Twitter accounts are treated as belonging to one unit or person (though that is not necessarily the case) and interactions between units are treated as equal (which, again, may not be true for different contexts). Or again, frequency of words is treated as indicating a certain sentiment or intent without regard to the fact that words may be used in different ways – for example, ironically (for an overview of using social media for sentiment analysis, see Thelwall et al., 2012, who also point to 'irony' and other problems). As such, Twitter data is treated as if it consists of abstract units, whereas in applied settings, this data would need to be translated into specific populations, targeted in particular times and particular places, and with specific messages.

The uses and limits of Big Data research

Data-driven knowledge is an advancing research front because of the availability of new data sources from digital media: the reason Big Data is new is that social scientists and others now have a number of sources (again, social media, Wikipedia, and the like) for studying human behaviour on an unprecedented scale. Yet it should be remembered that there are also limits to what this knowledge can do: for example, even if there are powerful Big Data techniques for establishing what my Facebook 'likes' might be, that is a far cry from obtaining my compliance in, say, making a purchase because of suggestions that have been made to me on the basis of these 'likes'. Put differently, there tends to be a very narrow aim in the case of applied Big Data knowledge, whereas in academic Big Data research, the aim is to obtain the broadest or most generalizable knowledge.

The process of generating more powerful knowledge invariably produces depersonalization, or a more deterministic approach to the world: inasmuch as the world is explained objectively, this leaves no scope within knowledge for individuality outside of impersonal laws or regularities. As Mayer-Schoenberger and Cukier (2013) point out in relation to law, Big Data can help to undermine the idea of personal responsibility, particularly as one of the cornerstones of the modern worldview is the idea of free will. But the issue they point to is much wider than law, since Big Data research also challenges our notions of individuality and self-determination outside of the legal context: if the aim of a study of Facebook is able to predict my personality or predict what I will do, this may not be legally ground-breaking, but it does undermine my sense of individuality on a personal level. Similarly, the very idea of technological determinism – that my behaviour may be not only predicted but *manipulated* by a particular technology – goes against fundamental (self-)understandings of how society operates according to individual and collective decision making. Moreover, it can be mentioned that although deterministic knowledge of human behaviour may seem threatening, for certain social purposes, more powerful knowledge will inevitably be needed – if we think, for example, about people's energy consumption in the face of the challenges of climate change. Further, it is worth recalling that it is not in the interest of firms to violate the privacy of people's data: firms collect personal data in order to influence our purchasing behaviour and the like, and it is thus a resource to be protected rather than shared unless there is a prospect of gain. Similarly, states want to protect populations from threats and obtain more powerful knowledge for policymaking and in some cases 'nudge' the behaviour of populations – not necessarily to diminish their freedoms.

If identifying new data sources highlights the new opportunities and dangers deriving from these sources, it therefore also points to the limits of Big Data approaches: there are only as many such sources as

people who use the objects which provide them (such as social media platforms or other objects which leave digital traces). Hypothetically, once the usefulness of analysing these sources is exhausted – if, say, all possible social scientifically interesting relationships on Facebook or Twitter have been researched – then there will be diminishing returns for social scientific knowledge – though not for commercial or other non-academic uses of Big Data (though similarly here, the practical uses of knowledge are limited to users of particular social media). It can also be noted that in some cases (again, not all: consider Wikipedia), these new sources may not be accessible to academic researchers except if they are purchased (for access to Twitter data, for example, see Puschmann and Burgess, 2014), and purchasing data will often be beyond the means of all but a few academic researchers.

New sources of Big Data have of course become widely available in the commercial world and, to a lesser extent, in government and in the non-profit sector. In these cases, data-driven research is typically carried out with narrow aims: if certain correlations, say, in purchasing behaviours are found, then these correlations can be used to encourage further purchases; or if certain crime hotspots are identified, law enforcement resources can be reallocated to counteract them (Eagle and Greene, 2014). Here it can be noted that while data can be used to target specific individuals, it may not be possible to change the behaviour of individuals (even if it is possible to 'nudge' them). However, in many cases, it may be sufficient that these correlations work at least in a profitable or useful proportion of instances.

Another issue arising from the difference between academic and commercial research that has recently been highlighted concerns the provenance of data, which is illustrated by studies which have used Google searches to analyse flu trends. As Lazer et al. (2014) point out, there are a number of methodological problems with these studies, but a major problem is simply that the data cannot be replicated since it is unclear how the data was arrived at (see also Borgman, forthcoming). This illustration also points to an interesting bind (which goes beyond the use of data from Google, but we can stick to this example here): use of data from Google by researchers from within Google can be high powered because researchers will know the problems with the data and know how to overcome them, but this research will be less useful for academic purposes because it may not be possible to replicate the studies or make these problems public. Academic research that uses open access to these data, on the other hand, will be less useful because these problems are not known, but more high powered because this research will be premised on the idea that it should be possible to replicate the findings and have transparency about provenance of the data. These advantages and disadvantages may not apply to all types of Big Data research – commercial, academic, or both – but they are often problematic where access to social media and commercial data is involved (as boyd and Crawford (2012) and Savage and Burrows (2007, 2009) have also pointed out).

Social implications of data-driven research

The ethics of Big Data research are typically considered in relation to current issues which require urgent regulatory and policy responses. Again, what is overlooked in these debates is the longer-term 'creep' in terms of the effects of more powerful knowledge, derived from Big Data sources, on society. The ethical implications of data-driven knowledge are hard to observe at an aggregate level where data is impersonal and anonymous. Data about individuals, on the other hand, is by nature personal and often sensitive, and the effects of applied data-driven knowledge on the individual are direct. A growing body of knowledge based on digital data is bound to have important social implications, but it does so qua knowledge, at a level that is mostly imperceptible to individuals. For individuals and policymakers, it seems most important to respond to immediate and recognizable issues in relation to data protection, even as the wider social consequences of the growth of knowledge are rather less tangible and less clearly identifiable.

Big Data raises major questions about a loss of human autonomy which arises from deterministic knowledge being applied to human behaviour. These questions revolve around free will and human agency in the face of knowledge which seems to take these away from individuals ('seems', because it does so in certain cases and from one perspective, but this knowledge also enables people to do things from a different perspective, such as gaining access to information about the things they might like to purchase). Big Data extends knowledge into new domains and achieves greater accuracy in pinpointing individual behaviour (which also entails that only people with a great deal of expertise about the workings of computers can avoid this kind of monitoring of their activity), and the capability of generating this knowledge can be undertaken by new actors and with more powerful tools (not just marketing and credit rating companies or large government agencies, but also those with access to web-based or other digital data and the capability to analyse it). All three of these departures are a product of access to new sources of – Big – data.

This paper has identified some of the major implications of data-driven research. In thinking about regulation beyond research ethics and legal questions about using this more powerful knowledge, it is difficult to see how commercial uses could be counteracted in contemporary market-driven societies. After all, there is no compulsion to use any one new social media service. As Google's executive chairman Eric Schmidt argued during anti-trust hearings in the US Senate, with regard to the Google search engine: 'it's also possible not to use Google search . . . the competition is just one click away' (NBC Bay Area News, 2011). The main possibilities for regulation in this case would be if social media (or search engines and other similar Big Data sources) would be regarded as part of an essential infrastructure for citizens, along similar lines as broadcast, communication via phone, energy or transport. In this case, as with Facebook and other services, it can also be argued that these services provide an essential public good. And if these services are seen as essential infrastructures, it will be possible to argue that the large-scale manipulation of people's thoughts and feelings via online media could be deemed harmful. (An alternative proposal, to treat the generation of data as labour that should be paid for (Fuchs, 2013), is neither feasible, nor, in commodifying the manipulation of behaviour, does this get at the core of the problem.)

Note, however, that such thinking towards regulation requires wider considerations beyond research ethics or laws that attach to particular studies. The 'creeping' advance of scientific knowledge is not based on individual studies, but depends on knowing who the users are, knowing their responses to certain stimuli, and having a complete picture of their online activity (which includes linking this to data about offline demographics and the like) – combined, these provide a more powerful picture, and thus a more manipulable population. In this way, knowledge generated by Big Data affects our everyday lives in a novel way. This type of knowledge inexorably moves towards a new type of omniscience (omniscience in the sense that everything that can be known via digital traces should be known, and will lead to a comprehensive understanding of human behaviour from individual actions to interactions at the global level; see, for example, Eagle and Greene, 2014) about an increasingly important part of our lives – life online. It would be easy, on the basis of these reflections, to draw rather apocalyptic conclusions, which are typical in the media and in some academic responses (again, Grimmelman (2014) contains examples), especially in view of the rather wide ranging influence that has been discussed. However, it is also important to reign in exaggeration: the effects of this type of research are confined to uses for commercial advantage, for political indoctrination (in authoritarian regimes) and nudging in democratic ones, and for advancing academic social science, and the limits to these have just been discussed. Instead of exaggerating dangers, it is more accurate to point to 'creep', which is diffuse, pervasive and also a largely invisible process.

Counteracting 'creep', similarly, requires drawing lines in the sand: where should manipulation of user behaviour be regulated to be transparent or subject to explicit (and meaningful) consent? (See Wilson et al., 2012, for consent and privacy in Facebook research.) When are users dependent on a service such that it is an essential part of the social infrastructure which requires regulation? When does academic social science work in support of commercial applications that are ethically unacceptable? These questions will (hopefully) increasingly come into focus, and they provide a different basis for potential regulation – or for opening up data to benefit the public good – than questions of research ethics or law and privacy in individual cases.

Academic research and applied research share the aim of producing powerful knowledge based on large-scale data. Where they differ is insofar as academic research aims at generalizable knowledge, while applied research aims at implementing knowledge derived from a Big Data source into reaching a particular audience with a view, for example, to influence purchasing behaviour and web traffic. The two overlap, but the ethical and social implications of the two are quite different in terms of privacy and data protection. One reason why it is nevertheless important to note their overlap is that both aim in the longer term at omniscience about human behaviour, even if the respective uses of this knowledge remain analytically separable. This omniscience could reach its limits when data from digital platforms and other digital traces no longer has value, but these limits will be quite different in respect to our scholarly understanding of the social world on one side, and how data can be exploited for commercial purposes and influencing people's political or social behaviour on the other.

Author's note

This paper is based in part, among other sources, on more than a hundred interviews with social science Big Data researchers (detailed in Schroeder (2014) and Taylor et al. (2014), though these were not involved in this Facebook study).

Acknowledgements

This paper has benefitted from discussions with Eric T Meyer, Linnet Taylor and Josh Cowls, as well as from several careful and helpful reviewers from *Big Data & Society*.

Declaration of conflicting interests

The author declares that there is no conflict of interest.

Funding

This paper has received support from the Alfred P. Sloan Foundation for the project 'Accessing and Using Big Data to Advance Social Science Knowledge'.

References

Andrejevic A and Gates K (eds) (2014) *Big Data Surveillance*, special issue of *Surveillance and Society* 12.

BBC (2014) Facebook faces UK probe over emotion study. Available at: http://www.bbc.co.uk/news/technology-28102550 (accessed 2 December 2014).

Beniger J (1986) *The Control Revolution: Technological and Economic Origins of the Information Society*. Cambridge, MA: Harvard University Press.

Bond R, Fariss C, Jones J, et al. (2012) A 61-million-person experiment in social influence and political mobilization. *Nature* 489: 295–298.

Borgman C (forthcoming) *Big Data, Little Data, No Data*. Cambridge, MA: MIT Press.

boyd D and Crawford K (2012) Critical questions for big data: Provocations for a cultural, technological and scholarly phenomenon. *Information, Communication and Society* 15(5): 662–679.

Brown I and Marsden C (2013) *Regulating Code: Good Governance and Better Regulation in the Information Age*. Cambridge, MA: MIT Press.

Cornell University (2014) Human Research Participant Protection Program Institutional Review Board (IRB). Available at: http://www.irb.cornell.edu/documents/IRB%20Policy%201%20%28Oct%202013%29.pdf (accessed 2 December 2014).

Cornell University Media Relations Office (2014) Media statement on Cornell University's role in Facebook 'emotional contagion' research. Available at: http://mediarelations.cornell.edu/2014/06/30/media-statement-on-cornell-universitys-role-in-facebook-emotional-contagion-research/ (accessed 2 December 2014).

Eagle N and Greene K (2014) *Reality Mining: Using Big Data to Engineer a Better World*. Cambridge, MA: MIT Press.

Fuchs C (2013) *Social Media: A Critical Introduction*. London: Sage.

Gartner (2014) IT glossary Big Data. Available at: http://www.gartner.com/it-glossary/big-data/ (accessed 2 December 2014).

Golder S and Macy M (2014) Digital footprints: opportunities and challenges for online social research. *Annual Review of Sociology* 40: 6.1–6.24.

González-Bailón S, Wang N, Rivero A, et al. (2014) Assessing the bias in samples of large online networks. *Social Networks* 38: 16–27.

Greenleaf G (2013) Data protection in a globalised network. In: Brown I (ed.) *Research Handbook on Governance of the Internet*. Cheltenham: Edward Elgar Publishing, pp. 221–259.

Greenleaf G (2014) Sheherezade and the 101 data privacy laws: origins, significance and global trajectories. *Journal of Law, Information & Science* 23(1).

Grimmelman J (2014) Personal website, with sources for the facebook emotional manipulation study. Available at: http://laboratorium.net/archive/2014/06/30/the_facebook_emotional_manipulation_study_source (accessed 2 December 2014).

Guardian (2014a) Privacy watchdog files complaint over Facebook emotion experiment. *Guardian*, 4 July. Available at: http://www.theguardian.com/technology/2014/jul/04/privacy-watchdog-files-complaint-over-facebook-emotion-experiment (accessed 2 December 2014).

Guardian (2014b) Facebook's 'experiment' was socially irresponsible. *Guardian*, 1 July. Available at: http://www.theguardian.com/technology/2014/jul/01/facebook-socially-irresponsible (accessed 2 December 2014).

Guardian (2014c) Facebook apologises for psychological experiments on users. *Guardian*, 2 July. Available at: http://www.theguardian.com/technology/2014/jul/02/facebook-apologises-psychological-experiments-on-users (accessed 2 December 2014).

Guardian (2014d) Facebook reveals news feed experiment to control emotions. *Guardian*, 29 June. Available at: http://www.theguardian.com/technology/2014/jun/29/facebook-users-emotions-news-feeds (accessed 2 December 2014).

Guardian (2104e) Facebook faces criticism amid claims it breached ethical guidelines with study. *Guardian*. Available at: http://www.theguardian.com/technology/2014/jun/30/facebook-internet (accessed 2 December 2014).

Guardian (2014f) Google allows Europeans to ask for links to be removed. *Guardian*, 30 May. Available at: http://www.theguardian.com/technology/2014/may/30/privacy-activists-welcoming-google-allowing-links-to-be-removed (accessed 2 December 2014).

Hacking I (1983) *Representing and Intervening*. Cambridge: Cambridge University Press.

Hacking I (1992) The self-vindication of the laboratory sciences. In: Pickering A (ed.) *Science as Practice and Culture*. Chicago: University of Chicago Press, pp. 29–64.

Hunter D and Evans N (2014) Call for Papers: The Facebook Emotional Manipulation Study and the Ethics of Big Data Research – A special issue of *Research Ethics*. Available at: https://app.simplenote.com/publish/WrHZm7 (accessed 2 December 2014).

Kramer A (2014) Blog post. Available at: https://www.facebook.com/akramer/posts/10152987150867796 (accessed 2 December 2014).

Kramer A, Guillory J and Hancock J (2014) Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 111(24): 8788–8790.

Lazer D, Kennedy R, King G, et al. (2014) The parable of Google Flu: traps in Big Data analysis. *Science* 343(6176): 1203–1205.

Mayer-Schoenberger V and Cukier K (2013) *Big Data: A Revolution that Will Transform How We Live, Work and Think*. London: John Murray.

Meyer M (2014) Misjudgements will drive social trials underground. *Nature* 511: 265.

Moat HS, Curme C, Avakian A, et al. (2013) Quantifying Wikipedia usage patterns before stock market moves. *Scientific Reports*, 3, Article number 1801.

NBC Bay Area News (2011) Schmidt on antitrust: competition is one click away. *NBC Bay Area News*, 21 September.

Available at: http://www.nbcbayarea.com/blogs/press-here/Schmidt-on-Antitrust-Competition-is-One-Click-Away-130300333.html (accessed 2 December 2014).

New York Times (2014) The plus in Google plus? It's mostly for Google. *New York Times*, 15 February, page A1.

Porter T (2008) Statistics and statistical methods. In: Porter T and Ross D (eds) *The Modern Social Sciences*. Cambridge: Cambridge University Press, pp. 238–250.

Puschmann C and Burgess J (2014) The politics of Twitter data. In: Weller K, Bruns A, Burgess J, Mahrt M and Puschmann C (eds) *Twitter and Society*. New York: Peter Lang, pp. 43–54.

Rule J (2007) *Privacy in Peril: How We are Sacrificing a Fundamental Right in Exchange for Security and Convenience*. New York: Oxford University Press.

Savage M and Burrows R (2007) The coming crisis of empirical sociology. *Sociology* 41(5): 885–899.

Savage M and Burrows R (2009) Some further reflections on the coming crisis of empirical sociology. *Sociology* 43(4): 762–772.

Schroeder R (2007) *Rethinking Science, Technology and Social Change*. Stanford: Stanford University Press.

Schroeder R (2014) Big Data: towards a more scientific social science and humanities? In: Graham M and Dutton WH (eds) *Society and the Internet*. Oxford: Oxford University Press, pp. 164–176.

Taylor L, Schroeder R and Meyer ET (2014) Emerging practices and perspectives on big data analysis in economics: Bigger and better, or more of the same? *Big Data and Society* July-December: 1–10.

Thelwall M, Buckley K and Paltoglou G (2012) Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology* 63(1): 163–173.

Turkle S (2011) *Alone Together: Why We Expect More from Technology and Less from Each Other*. New York: Basic Books.

White House (2014) The big data and privacy review. Available at: http://www.whitehouse.gov/issues/technology/big-data-review (accessed 2 December 2014).

Wilson R, Gosling S and Graham L (2012) A review of Facebook research in the social sciences. *Perspectives on Psychological Science* 7(3): 203–220.

Yarkoni T (2014) Personal website, in defense of Facebook. Available at: http://www.talyarkoni.org/blog/2014/06/28/in-defense-of-facebook/ (accessed 2 December 2014).

Yasseri T, Sumi R, Rung A, et al. (2012) Dynamics of conflicts in Wikipedia. *PloS ONE* 7(6): e38869.