Big Data, Ethics, and the Social Implications of Knowledge Production

Ralph Schroeder
Oxford Internet Institute
1 St Giles
Oxford, OX1 3JS
+44 (0)1865 287224
ralph.schroeder@oii.ox.ac.uk

Josh Cowls
Oxford Internet Institute
1 St Giles
Oxford, OX1 3JS
+44 (0)1865 287210
josh.cowls@oii.ox.ac.uk

ABSTRACT
This position paper addresses current debates about data in general, and big data specifically, by examining the ethical issues arising from advances in knowledge production. Typically ethical issues such as privacy and data protection are discussed in the context of regulatory and policy debates. Here we argue that this overlooks a larger picture whereby human autonomy is undermined by the growth of scientific knowledge. To make this argument, we first offer definitions of data and big data, and then examine why the uses of data-driven analyses of human behaviour in particular have recently experienced rapid growth. Next, we distinguish between the contexts in which big data research is used, and argue that this research has quite different implications in the context of scientific as opposed to applied research. We conclude by pointing to the fact that big data analyses are both enabled and constrained by the nature of data sources available. Big data research will nevertheless inevitably become more pervasive, and this will require more awareness on the part of data scientists, policymakers and a wider public about its contexts and often unintended consequences.

Categories and Subject Descriptors
K.4.1 [**Computers and society**]: Public Policy Issues – *Ethics*

General Terms
Human Factors, Legal Aspects.

Keywords
Data, big data, ethics, data protection, knowledge, sociology of science**.**

1. INTRODUCTION
That data generally and big data more specifically are the subject of major contemporary debate in society no longer needs belabouring. However, despite many discussions of their implications, neither data nor big data have been defined in the academic literature. Here we will provide such definitions in order to explain why knowledge based on data-driven research is new and why the sources of data in this research are distinctive. Once we have laid the groundwork by establishing what is new and distinctive about data-driven knowledge, we will be in a position to discuss its ethical implications. We argue that apart from current policy debates about privacy and data protection, data-driven research raises larger issues about the role of knowledge in society and about how knowledge can be used in relation to human behavior, with implications for how to create greater awareness among data scientists, policymakers, and a wider public. (This position paper is based on a larger project on 'Accessing and Using Big Data to Advance Social Science Knowledge'[1] which has examined the role of big data in advancing knowledge in the social sciences, and is based, among other sources, on more than a hundred interviews with big data researchers. The lessons that are drawn out here have emerged mainly from this project.)

2. BACKGROUND
The issues relating to privacy and data protection are too well-known to reiterate here [1, pp.47-68; 2]. Recently, there has been a shift from privacy in general to debates about data, and specifically digital data. Briefly, privacy and data protection laws are established to safeguard and ensure individuality and autonomy in society. There are currently many debates about data (for example, the 'right to be forgotten' in Europe[2], and the White House review on Big Data[3]) and privacy and data protection laws are spreading being adopted around the world (currently in more than 100 countries [3]). Greenleaf 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 establishing guidelines than in implementation [4, p.213].

3. DEFINING BIG DATA AND DATA
There are no definitive, academic definitions of data and of big data, but since specifying what is new about data-driven research is crucial for understanding its implications, we do so here. 'Big data' can be defined as research that represents a step change in the scale and scope of knowledge about a given phenomenon [5]. Note that this definition does not rely on 'size' per se, but on size in relation to the object being investigated, and how research advances beyond previous research about this type of object. But what is 'data'? In the definition offered here, data has three

[1] http://www.oii.ox.ac.uk/research/projects/?id=98

[2] http://www.theguardian.com/technology/2014/may/30/privacy-activists-welcoming-google-allowing-links-to-be-removed

[3] http://www.whitehouse.gov/issues/technology/big-data-review

characteristics: First, data belongs to the object or phenomenon under investigation; it is material collected about the research object. Second, data exist prior to analysis: as Hacking 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' [6, 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' [6, 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 are 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 [7] because the definition requires that there is an object 'out there' (realism) about how more useful or powerful knowledge has been gained (pragmatism). Hacking defines science as the 'adventure of the interlocking of representing and intervening' [7, p.146]; again, a pragmatist and realist account of the relation between scientific knowledge and the physical or natural worlds. Schroeder [8, p.9] has developed Hacking's ideas by arguing that technology is 'the adventure of the interlocking of refining and manipulating' 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.

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 are important in relation to 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 in scale 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 about these platforms is 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 the view of science and data that has been presented here 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 is digital platforms or peoples' digital traces; not in the manner of physical objects that can be intervened in - unless the environment from which digital data is gathered is also controlled. This is different for other sorts of technology, where more powerful tools can be used to manipulate the phenomenon under investigation, though again, in the case of digital data, unless researchers control these tools, 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 in them in practice; whereas the manipulation of these phenomena is a more practical, applied exercise, more powerful in exercising control over specific parts of the physical or social world.

4. USING DATA-DRIVEN KNOWLEDGE IN APPLIED AND ACADEMIC CONTEXTS

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 generate generalizable knowledge about human behavior, not to change it. Working in the private sector or in other applied settings, however, researchers and those who use knowledge (like marketers and advertisers) may want to do so. Thus the uses of big data for specific applications, influencing the behaviors 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 behaviors 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 behavior. In applied settings, the reverse is true: knowledge is generated inasmuch as it can be used to change behavior.

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 behavior). Consider, for example, Twitter data: when tweets are analyzed, this is typically done by counting word frequencies or message sent between accounts 'as if' these were units without context. That is, Twitter accounts are treated as belonging to one unit (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). As such, Twitter data is treated as if it consists of abstract units, whereas in applied settings, these data would need to be translated into specific populations, targeted in particular times and particular places, and with specific messages.

5. THE USES AND LIMITS OF BIG DATA

Data-driven knowledge is an advancing research front because of the availability of new data sources of digital data. 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 '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, a more deterministic approach to the world. As Mayer-Schoenberger and Cukier [9] 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 individuality on a personal level. Similarly, the very idea of technological determinism – that my behavior 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 behavior may seem threatening, for certain social purposes, more powerful such knowledge will inevitably be needed - if we think, for example, about peoples' energy consumption. Further, it is worth recalling that it is not in the interest of firms to violate the privacy of peoples' data: firms collect personal data in order to influence our purchasing behavior and the like, and it is thus a resource to be protected rather than distributed. Similarly, states want to protect populations from threats and gain more powerful knowledge for policymaking and in some cases 'nudge' the behavior of populations - not necessarily to diminish their freedoms.

If identifying new data sources highlights the new opportunities deriving from these sources, it 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 analyzing them 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.

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 with narrow goals: if certain correlations, say, in purchasing behaviors are found, then these correlations can be used to encourage further purchases; or if certain hotspots for crime are identified, resources can be reallocated to combat them. Note too that these data can be used to target specific individuals, and even if it is not possible to change the behavior of these individuals, it is sufficient that these correlations work at least in a profitable or useful proportion of cases.

6. THE ETHICS OF DATA-DRIVEN RESEARCH

The ethics of big data are typically considered in relation to current issues which require urgent regulatory and policy responses. 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 are impersonal and anonymous. Data about individuals, on the other hand, are 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 imperceptible to individuals. For individuals and policymakers, it seems most important to respond to immediate and recognizable issues relation to data protection, even as the wider social consequences of the growth of knowledge are rather less imperceptible.

Big data raises major questions about a loss of human autonomy which arises from deterministic knowledge being applied to human behavior. These questions revolve around free will and human agency in the face of knowledge which seems to take these away from individuals. Big data extends knowledge into new domains (such as predicting behavior from online activities), it achieves greater accuracy in pinpointing individual behavior (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 analyses it).

Academic research and applied research share the aim of producing powerful knowledge based on large-scale data; where they differ is in that academic research aims at generalizable knowledge, whereas applied research aims at implementing knowledge derived from a big data source into reaching a particular audience with a view, for example, to influencing purchasing behavior. The two overlap, but the ethical 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 behavior, 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 have value, but these limits will be quite different in respect to our understanding of the social world on one side, and how data can be exploited for commercial purposes and influencing peoples' political or social behavior on the other.

7. ACKNOWLEDGMENTS

This paper has benefitted from support from the Sloan Foundation for the project 'Accessing and Using Big Data to Advance Social Science Knowledge', and from discussions with Linnet Taylor and Eric T. Meyer, our colleagues on the project.

8. REFERENCES

[1] Brown, I. and Marsden, C.T. 2013. Regulating Code: Good Governance and Better Regulation in the Information Age. Cambridge MA: MIT Press.

[2] Rule, J. 2007. Privacy in Peril: How We are Sacrificing a Fundamental Right in Exchange for Security and Convenience. New York: Oxford University Press.

[3] Greenleaf, G. 2014, forthcoming. 'Sheherezade and the 101 data privacy laws: Origins, significance and global trajectories', Journal of Law, Information & Science.

[4] Greenleaf, G. 2013. 'Data protection in a globalised network' Chapter in Brown, I (ed.) Research Handbook on Governance of the Internet, Edward Elgar Publishing, Cheltenham, 221-59.

[5] Schroeder, R. 2014. 'Big Data: Towards a More Scientific Social Science and Humanities?', forthcoming in Graham, M., and Dutton, W. H. (eds.), Society and the Internet. Oxford: Oxford University Press.

[6] Hacking, I. 1992. 'The Self-Vindication of the Laboratory Sciences', in Andrew Pickering (ed.), Science as Practice and Culture, Chicago: University of Chicago Press, 29-64.

[7] Hacking, I. 1983. Representing and Intervening. Cambridge: Cambridge University Press.

[8] Schroeder, R. 2007. Rethinking Science, Technology, and Social Change. Stanford: Stanford University Press.

[9] Mayer-Schoenberger, Viktor and Cukier, Kenneth. 2013. Big Data: A Revolution that will transform how we live, work and think. London: John Murray.