Big Data and the Facebook Scandal: Issues and Responses

Abstract
The recent scandal over the appropriation of users’ data from the Facebook platform serves to surface wider concerns about ‘big data’, relating \textit{inter alia} to the ways in which personal data are obtained, stored and used for commercial purposes. This paper outlines some of the issues involved, and sketches some of the ways in which theologically-inspired reflection and action might begin to address those issues.

Keywords
Big data, consent, cybercrime, data brokerage, ethics, hermeneutics, privacy

Introduction
In March 2018 much media attention was devoted to the way in which data culled from the Facebook accounts of millions of users had been harvested and used by third parties. It has been suggested that the information this data yielded was used to affect people’s voting in events such as the UK’s Brexit referendum, and the election of Donald Trump to the Presidency of the USA.\footnote{1} This story surfaces a number of issues raised by the contemporary ‘big data’ phenomenon, which are only now coming to the attention of the public. The aim of this paper is to highlight some of these issues, and to suggest that theologians and religious communities have access to resources which may be of considerable importance in addressing them.

‘Big data’ is a notoriously vague term:\footnote{2} essentially, it is used to signify the capacity of today’s computers to capture and store enormous quantities of data. A number of commentators have noted that there is a ‘step change’ in moving from the kinds of
circumscribed datasets used by statisticians in the past and those which have become available through today’s computing practices, making the latter qualitatively as well as quantitatively different to the former. The term ‘data science’ has been coined to describe the novel techniques required for the analysis of such extremely large datasets.

Big datasets are sometimes generated by scientific investigations, for example through the collection of data in fields such as astronomy and particle physics, or through the accumulation of information in medical records. This paper focuses on the ways in which data are generated by any individual who makes use of the internet – by their use of social media, by the online purchases which they make, or simply by the storing of their internet browsing histories. Data are also routinely surrendered when people use retail storecards, or their mobile phones (which can record their geographical location in addition to any connectivity entered into by those phones); and facial recognition software has the capacity to identify individuals through their being recorded by security cameras, or through their appearance in photos posted online by others, without their knowledge or consent. As one commentator has observed, ‘[c]ollectively, data footprints and shadows [information about individuals generated by others] provide a highly detailed record of an individual’s daily life: their patterns of consumption, work, travel, communication, play, interactions with organisations, and their thoughts and interests.’ All these data can be stored, and subsequently used, by those organisations which have enabled its generation; and, in a marked departure from past practice, the purposes behind such storage and use increasingly relate to the exploitation of data for economic purposes, rather than for the pursuit of ‘pure’ research – data have come to be seen ‘primarily as an economic asset, not as a research one’.

This paper considers some of the issues raised by the collection and retention of personal data under three broad and overlapping headings: consent, the security and privacy
of personal data, and the commercial use of such data. These three issues will be considered separately, before some theological responses to them are offered.

Consent

It is routinely the case that whenever a person enters voluntarily into an arrangement whereby they put themselves at any kind of risk then the individual or organisation in charge of that arrangement is required to ensure that the person has given their informed consent before they enter into it. If I am to take part in a medical trial (for example), the team running that trial will need to explain to me exactly what it involves (in terms of the procedures to be undertaken, the timescale, and so on), and what the risks might be (in terms of the possible side-effects of the treatment). Well-established procedures monitor the conduct and the efficacy of such consenting processes. However, some issues remain. Suppose, for example, that such a procedure requires me to surrender genetic information about myself, which is then retained. In future, some new use for this information might be found, for which I have not given consent. Should I then be re-contacted so that fresh consent may be obtained? Should I give some kind of ‘blanket consent’ at the outset, covering any use to which the information I have surrendered might be put – in which case, how might my consent reasonably be described as ‘informed’?

When people use an online service, whether Facebook, or Google, or a retail website, they are required to acknowledge that they have read the ‘terms and conditions’ applying to that service. These cover issues such as privacy, and the extent to which data surrendered by users may be re-used by the service provider; and acknowledging the terms and conditions is notionally the way in which the users’ informed consent is gathered by the service provider. However, it has been observed that ‘[t]he notice and consent paradigm assumes that citizens are able to assess the potential benefits and costs of data acquisition sufficiently accurately to
make informed choices. This assumption … [is] a legal fantasy today, for a variety of reasons including the increasing use of complex and opaque predictive data-mining techniques, the interrelatedness of personal data, and the unpredictability of potential harms from its nearly ubiquitous collection’. The situation is further complicated by the fact that those gathering the data ‘typically attempt to minimise the ability of the person about whom data is being gathered to comprehend the scope of data, and its usage, through a mixture of sharp design and obscure legal jargon’. Put plainly, this leads to situations in which ‘people consent to the collection, use, and disclosure of their personal data when it is not in their self-interest to do so’.

There are clearly questions raised here which do not admit of easy answers. Can consent processes for online services be rewritten in such a way as to make them understandable and transparent for all users? Is there any way of ensuring that people are fully aware of the consequences when they agree to terms and conditions for internet services? And this raises the bigger question: to what extent can the full implications of all that people are ‘signing up’ to readily be apparent at the point where consent is sought and given – to those who are accruing the data, as well as to those whose data is being accrued?

Security and privacy

Once data have been generated, there are obvious issues around their safe keeping, and around maintaining the privacy of those who have surrendered them. We are all aware of cybercrime, whether it is directed against organisations (for example, through the use of ‘ransomware’ to extract financial gain) or against individuals (through the various types of malware that can infect PCs). All personal data which has been stored may potentially be attacked by hackers. For example, an attack on the Sony organisation in 2011 resulted in the theft of the ‘names, addresses, email address, birthdates, usernames, passwords, logins,
security questions and more’ from peoples’ accounts with the Sony PlayStation Network. No personal data, it appears, are immune to such attacks. No matter how carefully such data may be protected, they will always be at risk from those who are determined to access them, so that security software must constantly be updated in order to remain one step ahead of the hackers.

Not only this. In some cases, the data held about an individual in one particular location may be anonymised (or ‘de-identified’), which might appear to be a reliable way of protecting that individual’s privacy. However, it has been shown that when multiple de-identified datasets are combined, it is possible to break the anonymity of individuals by triangulating between those datasets. For example, in the American context it has been pointed out that ‘knowing an individual’s ZIP code localises that person to one in 30,000 (the average population of a ZIP code). Linking a ZIP code with a birthdate reduces the pool to approximately one in 80, while further connecting gender and year-of-birth are sufficient, on average, to uniquely specify an individual’. In other words, if different anonymized databases which contain my ZIP code, birthday, gender and year of birth are linked, it is highly likely that I can be identified. It has been maintained that ‘[p]rivacy and big data are simply incompatible’, and this does not appear to be an overstatement.

It would appear, then, that the data which users surrender to internet service providers cannot be guaranteed to be protected from theft by third parties; and that any such data which might in principle be thought to be protected through processes of anonymisation in practice are not. These are serious issues, and we might ask whether the risks associated with the storage and handling of data are ever explained to individuals when their consent is sought for the retention of their data. If not, the extent to which their consent is ‘informed’ is surely questionable.
Data brokerage

The main driver behind the rise of big data is, of course, financial. The economic benefits to be gleaned from harvesting and analysing the data which people surrender to online platforms is enormous (for example, it is reckoned that one third of Amazon’s sales are generated by recommendations based on the analysis of data which users submit to the site). This has led to the phenomenon of data brokerage, ‘an entire industry [which] has developed around the gathering and sale of consumer information, the analysing and enhancing of customer databases, and the sharing of customer lists.’ This brings with it a fresh set of problems. Attention has been drawn to the capacity for such analyses to increase inequality, for example when it comes to making decisions about who should and who should not have access to bank loans. It has been noted, moreover, that ‘many American companies that specialize in the collection and disclosure of personal information are largely unregulated.’ Kitchin notes that this means that data brokers ‘are not required by law to provide individuals access to the data held about them, nor are they obliged to correct errors relating to those individuals.’ Such errors might arise through an individual’s own actions (perhaps through the accidental input of incorrect data, or perhaps through the retention of views or activities posted years ago and since repudiated), or they might arise through the erroneous handling of their data by others. In either case, such erroneous information might come to have a huge effect on the services and opportunities which are available to people.

Mention of legal requirements raises the important issue of legislation around the gathering and trading of data. This is a complex matter, since data may readily be transferred across national boundaries where different legislatures obtain. At the time of writing (May 2018), the UK’s Data Protection Act of 1998 is about to be superseded by the EU’s General Data Protection Regulation (GDPR) of 2018. This legislation ‘expands the rights of individuals to control how their personal data is collected and processed, and places a range
of new obligations on organisations to be more accountable for data protection.’ It covers issues of consent and privacy, it provides strict frameworks within which data may be transferred outside the EU, and it requires those storing and processing data to demonstrate their compliance with the requirements of the Act. This is to be welcomed as a response to many of the issues raised in this paper; but given the ease with which data may be communicated across national boundaries it will be interesting to see just how globally enforceable legislation of this kind is, should it ever lead to prosecutions. (A current test case in the UK which has been set in train by the Facebook scandal, requiring information acquired under UK jurisdiction to be made available to a citizen in the US, may set important precedents in this respect. 18)

As things currently stand (and as the recent Facebook scandal has demonstrated), the commercial exploitation of potentially sensitive personal data is proceeding apace, and there is worryingly little that any individual can do about it.

**Responses**

How might theological responses be made to the various challenges raised by big data? To begin with, we might wish to critique the assumptions about human nature which are being made by data scientists, specifically that human beings are reducible to the kinds of thinking and behaviours which can be captured through the retention of data concerning them. Once information about humans becomes reduced to numbers and commodified, there is a danger that human beings themselves may come to be thought of in similarly reduced and commodified terms: that our value will be seen to lie in our potential for commercial exploitation, rather than in anything intrinsic to our shared human nature. In contrast to this, the theologian might wish to defend a radically different understanding of human nature, and to urge that (as Marc Cortez puts it), ‘Human beings are mysterious beings. There are depths
to humanity that will always escape our attempts to understand and define the human essence’ (italics in original).\textsuperscript{19}

From the changed perspectives regarding human nature which are afforded by theological reflection, there follow changed perspectives regarding the ethical issues surrounding big data. It has been suggested that data science is an entirely new way of doing science, with a consequent need for developing its own skills, practices, codes of ethics, and so on\textsuperscript{20} (for example, it has been suggested that data scientists should have their own version of the Hippocratic Oath\textsuperscript{21}). Such codes are indeed essential; but whence are they to be derived? In addressing ethical issues, purely pragmatic or functional approaches might be challenged or augmented through an appeal to theological virtues, such as justice, forgiveness and humility. With regard to justice, theologians might wish to ask (for example): should the use of data in such a way as to reinforce social stratification be condemned as an unjust practice? With regard to forgiveness, we might wish to ask: should people remain permanently accountable for opinions or behaviours which they post online as teenagers, which they may later repudiate? With regard to humility, some claims made on behalf of big data may be criticized as hubris-laden, for example the claim that big data represents ‘the end of theory’ – that it has made science, as traditionally understood, obsolete.\textsuperscript{22} Might it rather be urged that a more humble approach to what the analysis of big data can achieve can lead to a more balanced assessment of its true potential?

On the platform of such theological critiques, it might be urged that there is an important role to be played by churches (and, indeed, other faith communities) in raising public awareness of the very significant ethical issues which big data has brought to the fore.\textsuperscript{23} There are many precedents for authoritative bodies within the Church making statements and facilitating discussions on ethical issues: for example, the bishops of the Church of Sweden have issued a statement on climate change,\textsuperscript{24} and in the UK projects such
as ‘Scientists in Congregations’ have resourced congregational groups discussing topics such as ecology and medicine.\textsuperscript{25} Indeed, any local church community is in a position to provide a public forum where important issues may be discussed, and views on them shared. Insofar as such authoritative bodies and fora are in a position to inform the public, and directly and indirectly to influence legislation, they have a crucial role to play in shaping practically how society at large responds to the challenges of the big data age.

A further intriguing possibility is that the advent of big data might provide a new forum for engagement in the decades-old dialogue between theologians and scientists. I have suggested elsewhere\textsuperscript{26} that an important skill which has been developed by theologians in the pursuit of their discipline is that of hermeneutics: the search for meaning(s) within texts. A similar skill is clearly required of data scientists as they search for meaning(s) in very large datasets – not least because, as has been pointed out, ‘When the amount of data is sufficiently large, you can find almost anything you seek lurking somewhere within.’\textsuperscript{27} Here, then, is a way in which skills developed by theologians might usefully be shared with the data science community.

**Conclusion**

The recent Facebook scandal reveals the tip of an iceberg in terms of the potential abuses of personal information that have become possible in the age of big data. Knowingly or unknowingly (and the ineffective use of consent mechanisms suggests that much of the time it is the latter), individuals are offering up significant amounts of personal data to online service providers, and they are insufficiently protected against the possible abuse of that data in the future. Moreover, the creation of a multi-billion dollar industry trading in this data\textsuperscript{28} means that strong vested interests have been created which might well resist the rectification of this situation through legislation.
Theologians, together with Churches and other religious institutions, are in a unique position to serve the community over these matters, through raising awareness of them, through offering fora in which they may be discussed, and through offering practical, ethical and theological critiques addressing the use and abuse of big data. Such critiques might question both the ways in which data are interpreted and the models of human beings which they presume. All these are potentially crucial discussions in the ongoing debate around big data which the recent Facebook scandal has surfaced – and the need for them is becoming daily more urgent.


4 Kitchin, The Data Revolution, p. 167.


12 Mayer-Schönberger and Cukier, Big Data, p. 52.


15 CIPPIC, On the Data Trail, p. 4.

16 Kitchin, The Data Revolution, p. 44.


See Mayer-Schönberger and Cukier, Big Data, pp. 70-72.


Kitchin, The Data Revolution, p. 42.

Michael Fuller is a Senior Teaching Fellow at New College, University of Edinburgh. He is Chair of the Science and Religion Forum, and Vice-President for Publications of the European Society for the Study of Science and Theology.

This article contains 3540 words, including the biography and endnotes.