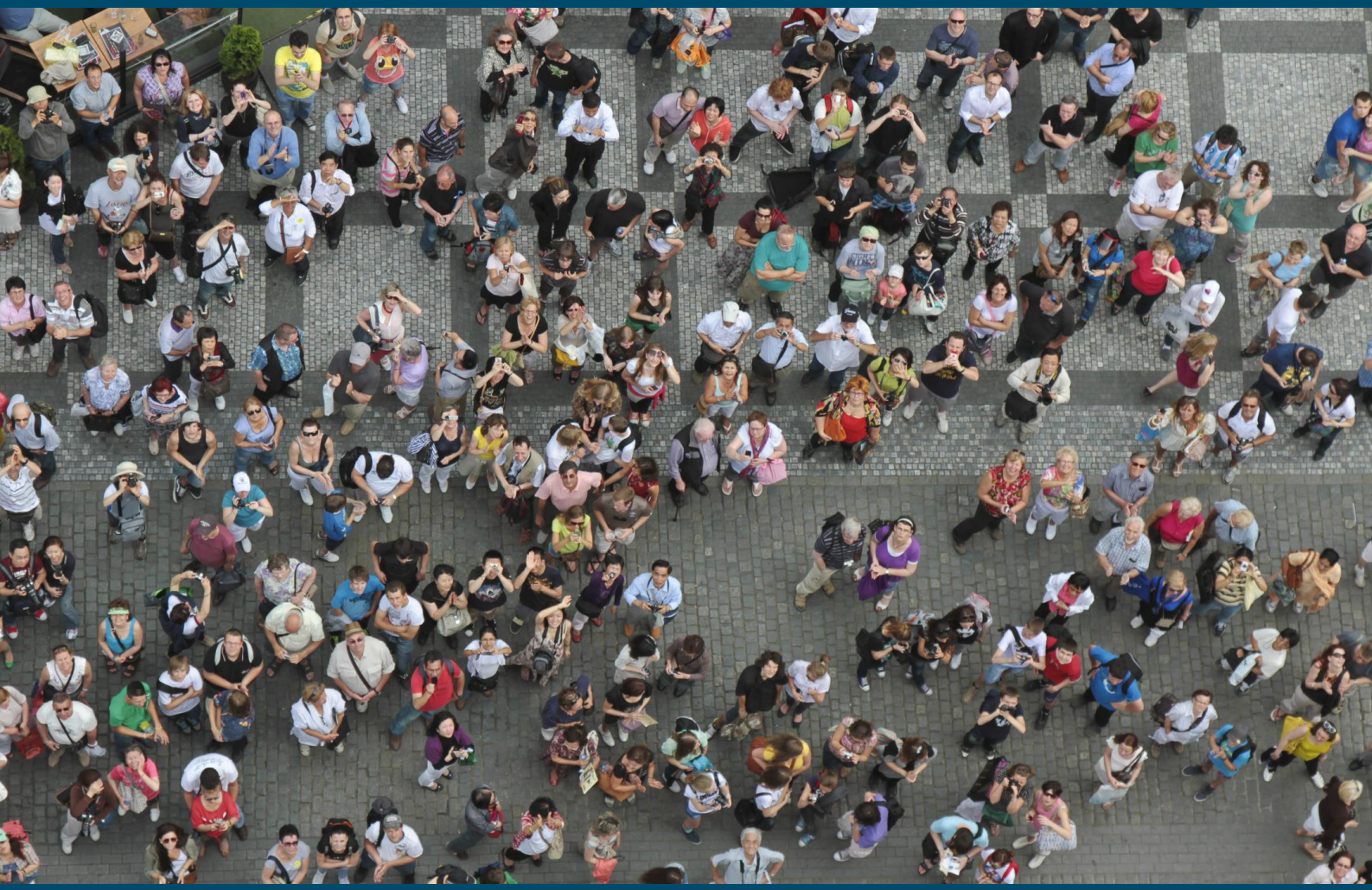
ROYAL STATISTICAL SOCIETY

DATA | EVIDENCE | DECISIONS

The Opportunities and Ethics of Big Data



Workshop report

Report of a consultation run by St George's House in November 2015 in association with the Royal Statistical Society, supported by the British Academy and SAGE

Introduction and acknowledgements

This Royal Statistical Society (RSS) report summarises the programme and discussion that took place in a two-day consultation convened in partnership with St George's House on 27th-28th November 2015, on the opportunities and ethics of Big Data.

The Consultation aimed to start thinking about the ethics of big data, bringing together participants from a wide range of sectors (listed on p. 10), to explore the issues. We note and value the further and wider follow-up from participants who are building up understanding and expertise on big data, which could in many cases contribute towards our recommended Council for Big Data Ethics. We are hugely grateful to the Chair, Professor Denise Lievesley, and to participants in the event. We also recognise and call for a far wider programme of public engagement to fully address scientific and ethical concerns.

The RSS would like to give special thanks to St George's House for co-organising and supporting the event, and to SAGE and to the British Academy for their support. The event took place at St George's House under the House Protocol, which encourages full and open discussion with no comments attributed to individuals.

Contents

1. Big data, opportunities and ethics

1.1. The big data landscape

The term ‘big data’ is frequently used. There is no agreed definition, but in public discourse it tends to refer to the increasing ubiquity of data, the vastness of datasets, the growth of digital data and other new or alternative data sources. More technically, people have talked about the increasing variety, velocity and variability of data which indicates we are in a ‘big data age’.

The new landscape of big data has featured in public and private debate in a widely acknowledged ‘big data hype’. This is framed and influenced in varying ways by those with a stake in the discussion and the inquiries and decisions that they seek to make. Within this, there are important ways in which work with data and statistics has been framed:

- Big data has helped people to look differently at data. Leading this trend has been the growth in availability of data as a by-product of routine processes in almost every area of life, such as satellite data and its transformation into visualised or mapped data. This makes the use of data far more applicable and appealing. Statistical methods for interrogating big data, such as machine learning, have also become more sophisticated to deal with these new types of data. The new insights derived from these methods boost enthusiasm for data analytics as a whole, as big data presents a powerful and innovative context for making inquiries.

- Big data has the potential to change the type of evidence that is available for policy makers to consider. Policy makers could refer more to computer models and predictive analytics as a basis for their decisions.

- Big data challenges pre-existing governance and policy. It implies a need for a comprehensive data infrastructure, so that data sources are appropriately organised, and can be accessed for appropriate use. It opens the case for legislative changes, for example to join up data from disparate sources for appropriate goals. It also opens up questions of who is able to access data, and whether ethical frameworks are fit for purpose to guide access and use.

- Big data challenges public understanding and public trust. Ethical concerns include questions about personal privacy, individual consent, data ownership and transparency. Research by the Royal Statistical Society has indicated that there is a ‘data trust deficit’ whereby the public have lower levels of trust in institutions to use their data appropriately, when compared to their general levels of trust in that organisation.[1] There are also worries about algorithms making important societal decisions – for example the predictive policing

[1] ‘RSS research finds ‘data trust deficit’, with lessons for policymakers’, *StatsLife*, 22 July 2014. Available at: www.statslife.org.uk/news/1672-new-rss-research-finds-data-trust-deficit-with-lessons-for-policymakers

model used in Los Angeles, which some argue could lead to unwarranted discrimination against particular sections of the population.[2]

In order to generate the gains that can come from analysing and linking big data sets, we need to consider the ethical frameworks and principles that will maintain public trust. In the United States, the National Science Foundation helped to set up a Council for Big Data, Ethics and Society. There is no equivalent in the UK. Do existing ethical, regulatory and legal frameworks need to change, or can they accommodate big data? Do professional bodies need to change their professional codes in light of the changing nature of data? How can we use the increasing amounts of data in society for public good and with public support? These are among the issues that were raised in November, and that continue to be explored.

1.2. Opportunities and ethics: what should we expect for the future?

It was agreed that there are common issues that will be confronted in the use of big data, and that these will apply across sectors and interest groups. Those that were apparent even at the outset of discussion included:

- data ownership
- data quality
- where we access data
- public perceptions and public understanding
- the legislative framework
- data ethics, and professional and organisational culture

Ethical debates about data in 2015 have been dominated by questions of privacy, convenience, surveillance, encryption and anonymisation. Such debates will move on, in a long-term view, to feature questions about autonomous technology and artificial intelligence.

As technological developments continue apace, it was raised that we lack an understanding of harms. What is the equivalent of an old-fashioned industrial injury in the big data world? It was considered that there are not only concerns about harms to privacy, but other concerns as well.

Two primary means of interrogating big data were identified, which are to question **trust** in the findings, and **truth** in the methods.

This was first discussed with regard to the widening breadth of uses of algorithms to provide services. Algorithms developed through fundamental mathematical research are excellent at identifying patterns in data, but we need to ensure that they are picking up what matters. This relies on understanding the context in which the algorithm should function. There is a need to guard against dangerous assumptions that algorithms are near-perfect, or more perfect than

[2] Brayne, S. Rosenblat, A. Boyd, D. (2015) *Predictive Policing - Data and Civil Rights* (PDF), Available from: http://www.datacivilrights.org/pubs/2015-1027/Predictive_Policing.pdf

human judgement. Errors introduced by algorithms are often not random and have systematic effects. It was argued that effects should be looked at across society, to address any large-scale unintended consequences. In particular, there is a risk of systematic discrimination in the provision of services, and for this to happen with a lack of human interest or decisions with regard to this. It was agreed that big data exists alongside politics, and needs to be coupled with space for politics, so that there are mechanisms to hold power to account.

For statistics, big data represents the opportunity to utilise data from a wide range of different sources, as the volume of data that might be collected, often on an automated basis through the use of technology, has increased. Big data is often regarded as 'exhaust data', or 'non-deliberative data', data that is a by-product of other data collection that takes place. Statistics is a use that was unanticipated at the time when the data was collected. Access to data from the use of technology (the internet and smart devices), and access to government administrative data, were each highlighted as important opportunities which could help statistics draw closer to providing real-time information about society. Management of these data sources could address challenges in traditional methods, for example where large scale surveys struggle with reduced response rates. Coupled with this is a need to manage and address substantial concerns about the quality and representativeness of the data. There is a need to cross-reference findings from big data against those from other national or international sources.

It was noted that big data is also qualitatively different from data that has previously been gathered, carrying with it both opportunities and ethical concerns. Data generated by citizens in their use of social media or the internet represents their concerns in a different way, compared to the views reported to government-designed surveys. Technology crosses boundaries, and makes existing boundaries irrelevant. Automated systems, such as bitcoin and blockchain, are capable of carrying data with no oversight or human intervention.

Questions were raised about societal consequences. When an enormous body of data is used to predict outcomes for individuals before a decision has even been made, what is then the response (among individuals and in society) to services that are more determined by prediction? Other countries, including in the developing world, are investing heavily in data-driven development. The outcomes of this investment will be substantially different in different contexts. This implies that across countries and across borders, we may carry with us wrong assumptions about the use of data, or wrong assumptions about politics. Knowledge we get from big data coalesces with the political environment, and political responses to the same data can change over time. Information used to help the public in one country might be used to harm the public in another, and this depends on how decisions are governed.

2. Public understanding and concerns

There are many scientific and technological innovations that prove controversial, suggesting a need for better public engagement on the part of decision-makers, to perceive and remedy concerns. Present debates about big data are reminiscent of dilemmas that have also been faced in relation to other scientific and technical developments. Where big data interlinks with personal or individual-level data is a prime example of this. Many people have concerns about the uses of personal information without consent, and about untested or unintended consequences. However many also expect that the government and other providers already use personal data to address problems in administration or services.

Research findings have shown that the approach and context for new initiatives affects people's concerns about proposed uses of data. Public views change for example depending on how informed they have been, whether by being exposed to various forms of evidence, or by being asked to participate deliberatively in discussions.

It was agreed that the relationship between ethics and public perception is not straightforward. It was emphasised that a broader ethical framework is needed. To the public, it may otherwise seem that there are no mechanisms for accountability outside of public outcry.

Several mechanisms were raised as helpful:

- Professional competence and due diligence on data protection
- Appointment of champions who are competent to address public concerns
- Transparency, across all dimensions.

It was noted for some data driven tools it is not technically possible to make mechanisms transparent or open to independent scrutiny, and that even where possible, transparency can face numerous hurdles. This suggests a need to prioritise what are the key transparency mechanisms that are both applicable and needed in the interests of public trust.

It was also agreed that the public needs to be more widely and proactively engaged in debates about big data. A far wider programme of public engagement would be needed to fully address scientific and ethical concerns.

3. Policy opportunities

The key feature of big data for policy making is that there is a lot more data that could be accessed or made available. There is more computing power with which to handle data, and this computing power is very cheap. With all of this, it is technologically more feasible to draw insights from data.

The opportunities in local services were explored in discussions. These ranged from tracking staff performance against the local demography of public needs, to demonstrating and showing where resources are lacking and investment needed, to identifying what approaches work better for the public at a lower cost. There is energy and investment behind identifying what works in public policy making, and an awareness that local decisions accompanied by evidence are more likely to secure investment from central budget holders.

At the same time, in many local authorities data analysis is not part of 'business as usual', and the only techniques that are used are those that are very long established. A great deal of data on society is non-personal and is not subject to data protection concerns. Much of this is held within organisations, and is not analysed or published for others to use. Under use of data reflects a need to address organisational culture, skills and training. Service providers need specialist knowledge to deal with data: how is data procured, how it is licensed, what is the data repository and how it is maintained? Local authorities struggle to draw leaders together to address this, in part because local issues are addressed by multiple service providers. With care commissioning groups for health, separate provision of adult social care, and academies for education, there are many more players at the local level, and without special intervention they may be worse at sharing data than ever before.

For both the public and the private sector, it was agreed that there should be awareness of the boundaries to rolling out effective use of data and technology. People need the skills and facilities to properly engage with it. Culture change needs to be supported, and organisations need to talk to each other. Big data has arrived across many sectors of industry, it is already part and parcel of many services, yet approaches to handling it could improve. For example to take an approach that establishes what 15 fields of data are crucial, rather than having 500.

It was agreed that we need to be clearer in our use of regulation, as issues raised with regard to data privacy and data security affect and challenge effective policy making. These issues regularly need to be addressed when considering data on households, businesses, or individuals.

It was argued that there is a divide between big government and small data. Everyone has their own internal idea about what's ethically right and wrong. There are ethical tensions that need to be resolved between countries', cities', and individual families' interests. It was agreed that it will be important to set out the ethical questions to weigh different developments against.

It was also agreed that in the wider landscape of government data, mistakes are often made. Standards are needed at the centre of government to raise the quality of data infrastructure for the UK. For example, government departments need to identify the key data that they hold and establish complete and coherent registers of these data, cleaning them up at source. Data standards also need harmonisation so that comparable statistics are produced across devolved governments within the UK. The quality of statistics needs to be assured, and central standards and guidance for public service contracts could help to remedy data gaps. Organisations

contracted to provide public services should have a feature of their contract that ensures they provide key data back, as a steady supply of data from services is needed to understand public needs and to invest appropriately for the future.

4. Big data governance

It was noted that for companies and government, big data poses a challenge for governance and legislation, and the role for each was discussed.

4.1. Governance

Part of the role of data governance is to minimise harm and maximise benefit from the use of data. It should include consideration of risks and risk management. There are multiple strategic challenges to address. Is there a risk in a technical sense? Is there public sensitivity to the data being used, or to how that information will be used?

An example raised of the issues for governance to address was a tool designed to derive ethno-cultural background from a forename and surname and also from publicly available Twitter accounts. The software produced from this an estimation of ethnicity and even the religious make-up of a population. The first question raised was the truthfulness of findings from this, given that rudimentary testing shows false positives, such as religious background being identified incorrectly. The second question was with regard to ethics, and whether social media shared in public should legitimately be used to deduce private information as part of a research process.

It was raised that there are multiple dimensions of research methods that affect public trust, for example:

- Who is accessing the data – are they trusted?
- Why do they want to access it – do they have a good reason?
- Where would they access it – is it a safe and secure setting?
- What would they access – is it personal information? Is it anonymised or otherwise protected?
- What will be the output or outcome – is it worthwhile?

Transparent understanding of research methods may also be missing at the operational level, with decisions drawn from the findings but only partially informed. Enforcement of good practice and of penalties may also be lacking, and the strength of the law to deal with specific areas of malpractice such as illegitimate re-identification was raised as an issue.

It was argued that governance needs to fill the gap between what organisations say they will do with big data, and what they say about how they do it. Legal terms and conditions are at present more geared toward resolving corporate liability than addressing public understanding.

4.2. Regulation

Three levels of regulation were seen as areas for action.

Guiding principles

It was suggested that guiding principles should be based on three sets of values:

- *Transparency, honesty and fairness.*
  It was said that these values are already underpinned by legal frameworks, and should be brought into effect.

- *Robustness, resilience, adaptability and usability.*
  There can be a trade-off between the robustness and resilience of data security, and its adaptability and usability. When asked to consent to usage of personal health data, individuals can make quite sophisticated trade-offs between e.g. the uses of data for health, and data privacy. The utility of sharing data in the short term needs to be reconciled with its robustness and resilience for long-term use.

- *Innovation, enterprise, and efficiency for providing new public goods.*
  As data is made usable, innovation, enterprise and efficiency are some of the key outcomes that are valued by policy makers. It was considered that the benefits of using big data need to be recognised and have weight in the decision-making process.

Ad hoc policies

Ad hoc policies are flexibly introduced to address different ethical and cultural challenges posed by different data sharing scenarios. Among data controllers who need to collaborate with each other, there may be risk aversion toward legitimate data sharing. Ad hoc policy might therefore encourage collaborating services to share data with other services where they can, understanding where there is and is not a data protection risk, and ask data controllers to justify if they cannot share it. In other contexts, for example among data brokers illegitimately sharing and selling data with a substantial privacy risk, policy needs to robustly work in the opposite direction to counter this illegitimate use.

Influencing in practice

It was considered that policies and regulations will only address emerging issues, rather than addressing all issues at source, and that there is a need for influencing in practice. Concepts of security by design and privacy by design are gaining recognition. More focused attention on 'ethics by design' could represent a broader missing piece.

For government practice in particular to improve, it was considered that all relevant civil servants should be schooled in good practice with data, and data leaders proactively appointed who have respect for legal and ethical boundaries but who are also ready to push for data developments that would benefit the public. Guidance exists, such as the statutory code on data sharing from the Information Commissioners Office, but both training and leadership are needed for wider adoption.

5. Bringing it all together

Discussion of the opportunities and ethics of big data highlighted that big data presents complex, multi-dimensional challenges. The use of data and evidence is increasingly recognised as a major opportunity, but is also an opportunity that needs questioning. As technological developments continue apace, two primary means of interrogating big data were identified, which are to question trust in the findings, and truth in the methods.

It was understood that the relationship between ethics and public perception is not straightforward, and that a broader ethical framework is needed. The strength of legal enforcement of data protection was raised as a related issue. To the public, it may otherwise seem that there are no mechanisms for accountability outside of public outcry.

It was highlighted that there are multiple dimensions of research methods that affect public trust, such as: who is accessing the data, why they want to access it, where they would access it, what they would access, and what the intended output or outcome would be. Some important mechanisms for accountability to the public were identified, and include:

- professional competence and due diligence on data protection,
- appointment of champions who are competent to address public concerns
- transparency, across all dimensions.

It was agreed that an independent, neutral national entity, such as a Council for Big Data Ethics, is needed to formulate and uphold an authoritative ethical framework. Such a Council should draw upon a wide range of expertise, knowledge and interests across public, private, academia and other lay persons.

For government data in particular, it was also agreed that policymakers should strengthen central data functions in government to address strategic objectives, for example to:

- Build incentives for data sharing.
- Harmonise data standards for ease of use.
- Support trusted, federated data infrastructures with common rules for data access.
- Get data back for use that's lost through fragmentation of service providers and contracting.
- Gain the public mandate, and inform the terms of the social contract.
- Establish a network of 'data leaders' widening support for good practice.

For the long term and across the UK's technology and research base, it was agreed that investment in analytical capabilities, training, and public engagement will help the improvement and uptake of data infrastructure. It was also noted however that this consultation discussion was on a small scale with a select group. Conclusions could not be drawn on public views, and ethical solutions to big data problems should be informed substantively by wider public consultation in big data debates.

Workshop Participants

Abdool Kara | *Chief Executive, Swale Borough Council*
Andrew Collinge | *Assistant Director – Intelligence & Analysis, Greater London Authority*
Anjana Ahuja | *Science journalist, including science correspondent for the Financial Times*
Anthony Walker | *Deputy CEO, Tech UK*
Bernard Silverman | *Chief Scientific Advisor, Home Office*
Bobby Duffy | *Managing Director, Ipsos MORI Social Research Institute*
Christopher Graham | *The Information Commissioner*
Claudia Pagliari | *Programme Director Global ehealth, ADRC for Scotland*
Denise Lievesley (chair) | *Principal, Green Templeton College, Oxford*
Diane Coyle | *Professor of Economics, University of Manchester*
Emer Coleman | *CEO of DSRPTN, digital consultancy*
Evelyn Ruppert | *Editor, Big Data & Society*
Francine Bennett | *CEO, Mastodon C*
Helen Margetts | *Director, Oxford Internet Institute*
Hetan Shah | *Executive Director, Royal Statistical Society*
Howard Covington | *Chair, Alan Turing Institute*
Jane Elliott | *Chief Executive, Economic and Social Research Council*
Marion Oswald | *Head of the Centre for Information Rights, University of Winchester*
Mike Hughes | *Chair of National Statistics Advisory Group, Royal Statistical Society*
Natasha McCarthy | *Head of Policy, British Academy*
Olivia Varley-Winter (rapporteur) | *Policy and Research Manager, Royal Statistical Society*
Paul Maltby | *Director for Data, Government Digital Service*
Penny Young | *Librarian, House of Commons*
Rob Kitchin | *Author of 'The Data Revolution' and Geography professor at Maynooth University*
Roeland Beerten | *Director of Policy and Public Affairs, Royal Statistical Society*
William Barker | *Deputy Technology Leader, Department for Communities and Local Government*
Vanessa Lawrence | *Non-Executive Director, Satellite Applications Catapult, Senior Strategic Global Advisor for Geospatial to the World Bank*
Ziyad Marar | *Executive Director, SAGE*

About the RSS

The Royal Statistical Society (RSS) is a learned society and professional body for statisticians and data analysts, with almost 8000 members around the world. As a charity, we advocate the key role of statistics and data in society, and have done so since we were founded in 1834.

One of our six key strategic goals is for statistics to be used effectively in the public interest, so that policy formulation and decision-making are informed by evidence for the good of society.

The Royal Statistical Society
12 Errol Street
London EC1Y 8LX

+44 (0)20 7638 8998
rss@rss.org.uk
www.rss.org.uk
www.statslife.org.uk
@RoyalStatSoc