THE FOUNDATION
FOR SCIENCE AND
TECHNOLOGY

DEBATE SUMMARY

Is a paradigm shift taking place in the ways individuals and organisations access, analyse and protect data?

Held at The Royal Society on 25th May, 2016.

The Foundation is grateful to Haskel Family Foundation, the Royal Commission for the Exhibition of 1851 and Willis Towers Watson for supporting this debate.

The hash tag for this debate is #fstdata .
Audio files of the speeches are on www.foundation.org.uk .

Chair: **The Earl of Selborne GBE FRS**
Chairman, The Foundation for Science and Technology

Speakers: **Professor Sir Nigel Shadbolt FREng**
Chair, Open Data Institute and Principal, Jesus College, Oxford
Dr Mike Lynch OBE FRS FREng DL
Founder, Invoke Capital and Deputy Chair, The Foundation for Science and Technology
Professor David Hand OBE FBA
Chief Scientific Advisor, Winton Capital and Professor, Imperial College, London

Panellist: **Baroness O'Neill of Bengarve CH CBE FBA FRS FMedSci**
House of Lords

Introducing the speakers, THE EARL OF SELBORNE explained that the Foundation had organised a round table on the topic earlier in the afternoon, and now was the opportunity to share the ideas discussed with a wider audience. The United Kingdom had strengths in data science with innovations in data access, search and archiving. The economic potential of data technologies was considerable but so too was the increasing dependence on data from the Internet and thus on its supporting infrastructure, for example the timing signals from GPS satellites on which global financial trading depended. More and more information about each of us was being collected, stored, mined and monetised in data sets that required protection. Concerns about the effectiveness of privacy protection drove regulation which in turn if not well directed could inhibit innovation. The issues for debate were many and pressing.

Introducing the question, SIR NIGEL SHADBOLT suggested that the full case for open data still had to be made. There were exemplary instances where the combination of open data, mobile devices and social media had made a difference, and he illustrated this with the example of crowd-sourced mapping of damage after the Kathmandu earthquake. The association of data points with geo-location was especially powerful. An example of this was the mapping of prescription data in England and Wales revealing areas where patented rather than generic drugs were being routinely prescribed, at significant extra cost (some £200m a year) to the NHS. Another mapping showed where over-prescription of antibiotics was likely. Making basic data available through an API (Application Programme Interface) stimulated innovation, such as had occurred with apps showing waiting times for buses and tubes, or providing directions, built on TfL open data feeds.

Continuing, Sir Nigel Shadbolt referred to the useful distinctions drawn by the Open Data Institute (ODI) between closed, shared and open data sets. Some closed sets might only be available to employers. Some shared sets could be specified in contracts and others available to specified groups for example for the purposes of authentication. Others were best seen as assets available to anyone. The creation and curation of large data sets should be seen as part of national infrastructure, to be invested in just as in traditional transport and other infrastructure in the interests of the economy. Longstanding taxpayer funded (or subsidised) examples, now digitised, could be seen at Ordnance Survey, the Land Registry and Companies House, together providing reference data essential for administrative geography. In the future there could be a social fabric of performance and reference data on every aspect of economic and social life but only if we planned now what registers would be needed. Such data would be a public good, and care would be needed to work out how much of it government needed to seed and how much could be left to the private sector, and if the latter how best to ensure that the data was genuinely open to others (although it did not have to be free).

Concluding, Sir Nigel Shadbolt suggested that the current platform-centered nature of the web may change into a re-decentralised web where individuals had their own social personal data store. Individuals might manage their personal data with flexible architectures that could act as points of contact for those wishing to use the data. In such a technological ecosystem, many issues could be addressed within a system that respected the autonomy of the data subject in providing limited abilities to control self-presentation[1]. The consumer would need to feel empowered in such ways if innovation was not to be stifled by privacy worries, as illustrated by the arguments following European Court of Justice concerns over data retention and over the adequacy of 'safe harbour' arrangements for EU citizen data processed by US Internet companies. The recent market for mobile devices and sensors that monitored personal fitness and wellness and stored data in the cloud was an example of the privacy issues that would arise in exploiting such data. Where would the data be processed, would the wearer retain control of the data thus generated, and who would have the right to use the data, or to share it with others such as insurance companies? Useful research had been carried out and published by US and EU privacy experts in 2015 to identify practical steps to bridge gaps between the existing approaches to data privacy of the EU and the US in a way that produces a high level of protection, furthering the interests of individuals and increasing certainty for commercial organizations[2].

MIKE LYNCH drew on the main lines of discussion in the Foundation round table held earlier in the day (full record circulated with this note). Open data sets were of increasing importance to the economy and given their benefits deserved to be considered strategically as part of the national infrastructure and invested in accordingly. Open did not have to mean free and we should not be shy of companies realising the monetary value of data. Nevertheless, where the taxpayer had funded the creation of the open data set, as in NHS derived information, care would need to be taken in the conditions of access by commercial companies to avoid the taxpayer then subsidising private enterprise, for example by buying services or drugs built on its own data. Many data sources will never be open and how they are handled and the regulation of actions that follow their use also has to be covered as part of the debate.

Mike Lynch suggested that there were sufficient differences with big data because of their volume, ease of access, and the new technologies of analysis and exploitation to justify using the term paradigm shift:

- The amount and nature of data was fundamentally changing with the ability to manage unstructured data, including prose, video and audio, with the potential to generate unexpected results from data that would not previously have been tractable.
- Mobile sensors and devices, as seen in the coming Internet of Things (IoT), will generate unprecedented volumes of real time data creating new opportunities for innovation.
- Data fusion techniques running queries and testing hypotheses across very different data sets had the potential to answer previously unanswerable questions without first having to invest heavily in creating and applying data standards.
- Data analytics had advanced very rapidly, for example in finding insights from dirty and even inaccurate data on a 'good enough' basis provided that data was not systematically biased, overtaking the previous necessity to clean data sets.
- Machine learning could provide powerful algorithms to mine big data and recognise patterns but were dependent on unbiased and representative learning data sets.

Continuing, Mike Lynch argued that privacy and security went together: the former could not be assured without adequate arrangements for securing data. Anonymisation of data relating to individuals was hard, and with advances in data analytics could not be guaranteed as more data sets became openly available that could be correlated. There would at times be unintended consequences from big data use, and further developments were hard to predict. Efforts to obtain informed consent to data sharing will become less meaningful as it becomes harder to specify in advance the uses to which data can be put. Ways of authorising the repurposing of data would be needed, within an ethical framework, focusing on the beneficial uses to which it could be put rather than the status of the data as personal information. The key to public acceptance of data sharing was perception of benefit. Data analytic skills were in short supply, but the UK was relatively well placed internationally.

In order to illuminate further the issues already raised, DAVID HAND explored the contrary proposition, that the changes taking place in big data should be seen as cumulatively building on earlier ideas rather than being seen as a paradigm shift, in the sense in which Thomas Kuhn[2] had introduced that term[3] to describe a fundamental change in the basic concepts and experimental practices of a scientific discipline. He acknowledged that there were powerful new tools for analysis

[1] See O'Hara, K, Shadbolt, N. and Hall, W. *A Pragmatic Approach to the Right to be Forgotten,* Ottawa: CIGI (2016) prepared for the Bildt Global Commission on Internet Governance, available at https://www.cigionline.org/activity/global-commission-internet-governance, accessed 1 June 2016.

[2] Kuhn, D. *The Structure of Scientific Revolutions*, (1962) Chicago: University of Chicago Press.

emerging from the new science of data analytics, and there was more data about more things more readily accessible thanks to the Internet. Trends in data capture were certainly encouraging with digitization generating near continuous data streams from Internet connected sensors. Analysis in real time of data was enabling new applications such as helping counter fraud and detecting engine wear. The fall in the cost of storage was leading to the storage of far more digital data. Repurposing of data from new angles was now common. But example of each of these developments could be found before the so-called age of big data.

Concluding, David Hand agreed that there were many important if incremental developments that provided exciting new possibilities for using data constructively. Big data was capable of illuminating what people actually did rather than surveys that reported what they said they did. It was certainly the case that having the computer as the intermediary between the individual and the data introduced novel features. But most if not all of the ethical and legal issues that these developments were throwing up such as over ownership of derived and aggregated data were not in essence new, and had arisen including in cases already tested in the Courts. The shift in behaviours exhibited on social media perhaps came closest to being properly described in paradigmatic terms.

BARONESS O'NEILL illustrated the longstanding European Union approach of regulating privacy through data protection, not data use. The Data Protection Act 1998 governs the processing of personal data in the UK, translating the European Directive into law, as regulated by the Information Commissioner. In future, the European Data Protection Regulation, as a regulation rather than a directive, will have legal force directly. Under the Act what makes data personal is that it 'relates to a living individual' in terms of the data itself or other information in the possession or likely to come into the possession of the data controller. In the words of the guidance on the current Act:

> 'Personal data means data which relate to a living individual who can be identified –
>
> (a) from those data, or
>
> (b) from those data and other information which is in the possession of, or is likely to come into the possession of, the data controller,
>
> and includes any expression of opinion about the individual and any indication of the intentions of the data controller or any other person in respect of the individual. It is important to note that, where the ability to identify an individual depends partly on the data held and partly on other information (not necessarily data), the data held will still be "personal data"'.

The class of personal data was therefore by no means the same as that of sensitive information over whose use the individual would in practice want control. But what mattered today in judging lawfulness of processing data was whether the content of the data counted as personal data, that is capable of identifying an individual to whom it pertained with reasonable means, rather than the nature of the act of using it. Baroness O'Neill suggested that approach was not very useful. Confidentiality was hard to operationalize. Difficulties arose, for example over informed consent and repurposing of data sets. So-called 'informed consent' cannot be given when it is not possible to anticipate what future transactions might use the data. Data retention and use could also be problematic when it involved information on such matters as juvenile and spent offences.

Concluding, Baroness O'Neill suggested that a policy shift to controlling who gets to use the data, what they can do with it, and sanctioning misuse when the rules are broken would be more effective both from the point of view of the protection of the position of individuals and of maximising the utility to society of big data, but would represent a radical reform for which the necessary political support was not yet evident. Nevertheless, there were promising developments of ethically robust data governance through 'safe haven' structures, an approach used in the UK Biobank research programme and in the Scottish Health Informatics project. With any workable ethical framework, however, it had of course to be accepted that accidents and mistakes would happen.

In open discussion (continued after dinner), several participants questioned where the boundary of shared data should be drawn in cases where it could be argued that the social value of sharing outweighed the invasion of privacy of the individual. The emergency services for example routinely tried to identify the location from which 999 calls were made, accessing potentially life-saving data in urgent cases. Such data use was best regulated through 'social consent' in which democratic representatives, at national or in some cases local levels, agreed the rules in the public interest and in a transparent way. The same issue affected access to some individual medical data where there was a threat to life. In the future, accountable computing techniques might be able to ensure that the appropriate conditions of use travelled with the data. Nevertheless, the data environment was changing rapidly and would continue to change so it would be wrong to expect regulations and rules to remain stable. And it was likely that we were stuck with the public not taking on board the implications of the availability of so much of their data.

In discussion of business models, there was support for the view that repurposing of data at scale could lead to significant new profitable opportunities. As a generality the value of data had risen in recent years as data analytics had developed and as the network effect had its effect. The global nature of the Internet also meant that business models based on greater volume were more profitable. For many Internet based companies, data was their principal asset as well as their raison d'etre. Fixed capital requirements were relatively small, meaning that SMEs could enter the market, and the barriers to entry related not to fixed investment but to access to data of which the ubiquitous sensors of the Internet of Things (IoT) would generate large amounts. The private sector could be relied upon to come up with new innovative ways to capture data through sensors, although recovering costs would require charging regimes which it was noted need not conflict with the open nature of the data.

It was recognized that the technology continued to generate new opportunities. The power of cloud computing had already enabled voice activated query systems such as Apple's Siri, machine transcription such as Dragon, and real-time language translation programmes such as Babelfish. A useful distinction, it was suggested, was between reference data, such as GPS or the Land Registry and event or transactional data such as an Internet purchase.

It was recognized in discussion that the UK would require more data analysts to exploit the opportunities, and this needed to be prioritised by further education. Those most needed would not necessarily be computer scientists (of which it was said the UK appeared to have a sufficiency), but would be those skilled in analyzing and fusing the range of data involved including unstructured prose text, video and audio rather than the structured numerical data with which statisticians were familiar. New jobs in data would be created. It was already evident that there was interest in schools in the subject. It was also pointed out that having better information from data analysis about social issues and problems and their causes, such as homelessness, would not of itself lead to change. The next generation of policy makers needed to be educated in using the modern open data evidence base, recognizing that the mastery of the results could well be disruptive.

Maintain the integrity of data was seen as a future concern, as dependence upon it increased for the proper functioning of the economy and society. Biases could be hard to detect in autonomous decision systems involving machine learning where the algorithms had been developed using training data that was not representative on the population over which decisions were being taken. Badly cleaned data sets might generate worse results than using modern analytics on the original data set. There were key sets of data such as crime statistics and property prices that already demonstrated the pitfalls in interpreting information provided by contributors with vested interests in the results. And, it was added, a public that would derive most of its information about the outside world through web sites and Internet feeds needed education in the biases that were inevitable in such reporting.

Concerns were expressed by some that, looking ahead, the public would come to push back against Internet technology. The business model of the Internet companies amounted to covert manipulation of spending through advanced Internet marketing using consumer data. Resistance could build to the persistency of data capture where youthful indiscretions and spent convictions remained accessible, indexed on the web. Despite apparent relaxation of inhibitions about sharing personal information with friends through social media there could be concern over the same information, and very much more, being held and exploited by the Internet companies themselves. In the future, such companies whose services were used by the general population might well not just be based in the US but in jurisdictions with very different political systems.

There would always be material, for example on mental health issues that individuals would regard as deeply sensitive. The present situation was not necessarily troubling to most people, but an economic down turn could lead companies to feel compelled to exploit data more aggressively to survive. On the other hand, it was argued that the benefits of being part of the digital world were already considerable and provided individuals continued to see advantage in their use of social media and online commerce they would accept the monetization of their data. And, for example, they would increasingly accept offers, such as those from insurance companies to lower premiums in return for installing logging boxes and road cameras in their vehicles. Government access would, however, remain more sensitive and citizens would expect to see regulation of access, such as was set out in the Investigative Powers Bill 2000 currently at Report stage in its passage through Parliament. In both private sector and government the best way of preventing push-back was such transparency over the uses to which data could be put and the authority required.

In concluding discussion, there was general agreement that we should expect transformational changes to follow the greater availability of large data sets and the development of data science to provide powerful tools for analysing them. The resulting issues of standards and regulation were global. Issues of trust around Internet data were complex, perceptions varied between generations, and might well evolve. Above all, it was essential for the economic health of the UK to have the right skills in the workforce to make the most of the huge potential of open data.

Sir David Omand GCB

A report of a round-table discussion on the same question immediately follows this report.

Open this document with Adobe Reader outside the browser and click on the URL to go to the sites on the next page.

Useful reports:

European Commission – Protection of Data
http://ec.europa.eu/justice/data-protection/

Royal Academy of Engineering Report
Connecting data: driving productivity and innovation
www.raeng.org.uk/publications/reports/connecting-data-driving-productivity

Data Science Ethical Framework launch: Matt Hancock, Cabinet Office - speech delivered on 19th May, 2016
www.gov.uk/government/speeches/data-science-ethical-framework-launch-matt-hancock-speech

Competition & Markets Authority Report
Retail banking market investigation: provisional decision on remedies
www.gov.uk/government/uploads/system/uploads/attachment_data/file/523755/retail_banking_market_pdr.pdf

Open Data Institute Report on Data Infrastructure
www.theodi.org/data-infrastructure

Open Data Institute Report on The Open Banking Standard
Unlocking the potential of open banking to improve competition, efficiency and stimulate innovation
www.theodi.org/open-banking-standard

Useful links:

Google UK
www.google.co.uk

Imperial College, London
www.imperial.ac.uk

Information Commissioner's Office
https://ico.org.uk/

Invoke Capital
www.invokecapital.com

Open Data Institute
www.theodi.org

Research Councils UK
www.rcuk.ac.uk

Royal Statistical Society
www.rss.org.uk

Winton Capital
www.wintoncapital.com

The Foundation for Science and Technology
www.foundation.org.uk

A Company Limited by Guarantee,
Registered in England No: 1327814,
Registered Charity No: 274727

SEE THE NEXT PAGE FOR THE ROUND-TABLE DISCUSSION REPORT

THE FOUNDATION
FOR SCIENCE AND
TECHNOLOGY

ROUND-TABLE DISCUSSION SUMMARY

Is a paradigm shift taking place in the ways individuals and organisations access, analyse and protect data?

Held at the British Academy on 25th May, 2016.

The Foundation is grateful to Haskel Family Foundation, the Royal Commission for the Exhibition of 1851 and Willis Towers Watson for supporting this debate.

The hash tag for this debate is #fstdata .

Chair: **Dr Mike Lynch OBE FRS FREng DL**
Deputy Chair, The Foundation for Science and Technology

Speakers: **Gavin Starks**
Chief Executive, Open Data Institute

Mike Warriner
Engineering Director, Google UK

Introducing the speakers, Dr Lynch explained that the Foundation had organised a dinner discussion later in the day on the same topic, at which the issues around big data could be shared with a wider audience. Big data was increasingly being generated and held by both government and the private sector. Social and political scientists, lawyers and ethicists needed to work with computer and data scientists in order to study issues such as whether public attitudes to data privacy and use were changing, the adequacy of the legal framework and to map the impact of further developments of the technology including the exploitation of the data that would be generated by the coming Internet of Things. Past data science had been based on the analysis of structured data, involving considerable effort to classify and clean data before meaningful analysis could be conducted. Today the more interesting areas of data science involved analysis of unstructured data such as video and voice, and where data volumes meant that significant results could be obtained even from dirty data sets. Big data was a transformational technology.

GAVIN STARKS highlighted the speed with which the technology had developed and was continuing to evolve. The world wide web was only some 10,000 days old and yet had attracted already some 1 billion sites and over 3.5 billion users globally. The 5 billion or so devices already connected via the Internet were generating huge quantities of data, often in real time. We were in the process of moving from an Internet defined by text to one defined by data. He explained the approach of the Open Data Institute to the spectrum of data, from closed data, with access defined strictly by the owner of the data, shared data where institutions, companies and groups could be authorised to access it, and open data where anyone had the right to access the information, although not necessarily without paying for it. Obtaining the advantages of open data did not have to mean providing it free.

Continuing, Gavin Starks stressed the transformative nature of big data. He recommended that big data should be seen as a form of national infrastructure. Like roads, there were open free public routes, essential to the smooth operation of the economy, but also shared roads where tolls applied, and closed roads on private property. Government ought to see investment in the data infrastructure as essential to future economic prosperity through reducing friction in the economy in the same way as improving transport and other public infrastructure. Taking an open data approach to government data could potentially add 0.5% to GDP and the same would be true for private sector data. The main economic argument was that of the network effect: the value of a piece of data increased the more people were connectable to it. Access could be time limited for specific purposes permitted by the data holder. But unlike investment in conventional assets data did not get worn by use and thereby lose its value.

MIKE WARRINER agreed that a paradigm shift was taking place, driven by the possibilities of accessing information collected through the Internet and by the number of mobile devices capable of accessing it. The Internet represented a great global leveller. The volume of data was now massive, including accurate locational information from devices and sensors. Developments in data science now made this information hugely valuable in many different ways, for example in the creation of so-called smart cities. Machine learning techniques provided new and powerful means of problem solving, provided of course that the data sets used for training were representative and unbiased. The economic and social potential of big data was evident, but exploitation could be held back by sensitivities over the personal nature of much of the data, creating issues over permissions for use. It was often the case that the value of personal data to the individual was much less than its value to society if shared, for example in transport usage, public health and medical science. Previously intractable problems were being solved. The major Internet companies held big data sets on their users (not least covering location, travel, calendars, and Internet usage) capable of generating important new insights. Users were sensitive over with whom they shared such personal data and companies needed to retain customer confidence that their data would be held securely, used in ways of which they generally approved, and that misuse to their detriment would be sanctioned. So educating the public as to the social value that could be derived from allowing their data to be shared was important. The user generally needed also to perceive a personal benefit that accompanied the sharing of and use of personal data.

Concluding, Mike Warriner drew attention to the global shortages of data scientists with the skills needed for the world of big data. The UK was so far relatively well placed with several innovative institutes and companies, for example in the field of machine learning, but much more attention would be needed in further education to ensure the economy had sufficient people with the right skills and experience of the emerging data technologies to meet the inevitably growing demand.

Baroness O'Neill drew attention to the longstanding European Union approach, most recently seen in the new European Data Protection Regulation, of regulating privacy through data protection not data use. What mattered therefore today in judging lawfulness of processing data was whether the content of the data counted as personal data, that is capable of identifying an individual to whom it pertained with reasonable means, rather than the nature of the act of using it. She suggested that modal approach to data was defective and was giving rise to difficulties, for example over informed consent and repurposing of data sets. What is in practice regarded by an individual as 'personal' depends upon context. Confidentiality was hard to operationalize. So-called 'informed consent' cannot be given when it is not possible to anticipate what future transactions might use the data. Data retention and use could also be problematic when it involved information on such matters as juvenile and spent offences. A promising avenue was the development of ethically robust data governance through 'safe haven' structures, an approach used in the UK Biobank research programme and in the Scottish Health Informatics project. A policy shift to controlling use of data and sanctioning misuse would be more effective both from the point of view of the protection of the position of individuals and of maximising the utility to society of big data, but would represent a radical reform for which the necessary political support was not yet evident.

In discussion, several participants drew attention to the lack of public understanding of the value of big data, and its social utility when aggregated and analysed, for example in the ability to map data sets to uncover previously unsuspected patterns, such as in detecting the early stages of a global epidemic. Part of the answer was for institutions, inside and outside government, to agree to surface the existence of their data sets and to open them for research. New commercial applications might well follow. In some cases opening up the data freely might lead to greater sales; in other cases where the data related to a core service a charging regime would be more appropriate. Contracts for public services such as transport and public service agreements should contain data sharing clauses. It was argued, however, that simply publishing data did not equate to having a digital infrastructure that encouraged agile exploitation and provided sufficient commercial incentive. The skill shortages that had been mentioned could hold back the realization of the value of the data. Adopting an all of government approach was hard and some government datasets were unlikely to be of high enough quality for use by business (although it was suggested that advances in data science were making that less of a problem).

It was argued that it was important not to foreclose future applications of big data through over-regulation today. It was best to start with a light touch, and to develop from case law as problems emerged and were prioritized and decided. Nevertheless, there were issues to be addressed around data ownership: if a local council installed sensors in domestic dustbins to report when they were full enough to require refuse collection was the data stream that of the user of the dustbin, the owner or landlord of the property or the Council that installed the system and processed the data? Who would own the data stream from a sensor in a washing machine built in by the manufacturer in order to measure wear on the main bearing? Who would determine access to the data for other purposes and how to prevent its misuse? How would consistent application of such definitions of data 'ownership' across sectors be ensured and by whom? Locational data held by an Internet company could be mined to identify to the authorities dangerous stretches of road where vehicles were habitually breaking the road speed limit and that might be generally regarded as an acceptable use because of its social value, but if the same commercial data were to be used to identify and enable the sanctioning of offending drivers of such vehicles the public reaction would be liable to be strongly negative. The future value of digitized personal data should therefore not be seen solely in financial terms. The modern individual would have no choice but to engage in the future digital economy if the full benefits of citizenship were to be realized. But a digital divide could open up with an impoverished minority still using non-smart phones and disenfranchised by being cut off from the ready access the majority would enjoy to new digitized government and commercial services.

A number of legal issues were raised, and a case made for a new legal framework that bridged the gap between current law on the one hand and accepting an absence of law on the other. The current approach to legal liability for software and applications could, for example, prove inadequate when autonomous systems and driverless vehicles became commonplace. On the other hand, the transparency provided by open data sets could improve the regulation of insurance and financial services, and technical developments in distributed ledger technology could make it harder to hide improper transactions. Even within today's legal framework there were issues over when and how to act to regulate digital data use. Big data could reveal that policies were inadvertently resulting in discrimination against certain groups, for example in insurance premiums. We needed to be prepared too for big data to reveal regional disparities and patterns of association in human behaviour that would be politically hard to discuss, let alone accept as genuine.

The potential for misuse of data science existed, and was illustrated by the growing Chinese example of scoring of individuals' Internet use according to government-determined criteria of socially acceptable behavior (and disapproved of behavior such as interest in political dissidence) with such benefits as cheaper insurance and the opportunity to apply for government jobs following good behavior.

Issues around the exploitation of intellectual property were also identified in discussion. A machine learning algorithm would be able to analyse the content of human artifacts such as books, music and works of visual art, distill their uniqueness and 'create' entirely new works. Would such efforts be regarded as the result of a creative process on the part of the machine or an exploitation of the intellectual property of the human creators behind the training data set? And data scientists would be to be very attentive to the risk of inadvertent bias in machine algorithms introduced through the use of training data that was not fully representative. And system design would need to provide for rapid investigation of claims of discrimination or bias in future autonomous machine decisions, such as the allocation of benefits or the premium charged for insurance.

In further discussion, no clear consensus emerged over whether social attitudes to the sensitivity of personal data were evolving in the direction of people being less concerned over privacy. There was evidence of young people's increasing disinhibition over sharing revealing photographs amongst their peer group suggesting that what would once have been considered embarrassing to have recorded had changed. Other changes in attitudes to privacy were to be expected as people grew up as long term users of social media. But the permanence of digital records had yet to sink in to public consciousness. And surveys suggested that privacy was seen as an important human right with around one third of young people concerned that 'big brother' government agencies could be legally authorized to access their communications even for purposes regarded as legitimate. On the other hand, 'little brother' already existed in the form of citizens equipped with mobile devices with high resolution cameras and microphones able to record everyday events, and crowd source identification or use the now powerful facial recognition software available through the cloud.

It was pointed out that the protocols and standards of both the Internet and the web had been designed without serious consideration of cyber security. There had as yet been no paradigm shift in security. The pioneers had not foreseen the rise of criminal exploitation nor of irresponsible State surveillance and offensive cyber activity. Yet privacy was impossible without adequate security of data. The large expansion expected in the Internet of Things (IoT) would greatly increase the attack surface for those of malign intent. It was very important that adequate security was built from the outset into IoT sensors and applications; retrofitting security, as demonstrated by the Internet itself, was expensive and far less satisfactory. The public was not technically aware of the data security issues that would inevitably arise with the commercial use of big data. A useful step would be to require publicly listed companies to include in their Annual Reports details of data breaches and to provide statements of assurance, as for financial control systems, that security arrangements had been independently audited.

Given advances in data science, guarantees could not be given that de-anonymisation of data back to the individual would not be possible, an issue pointing back to the value of controlling data by sanctioning misuse, rather than regulating by control of 'personal' data, if progress in using big data was not to be slowed down. It was suggested that a form of 'social compact' was needed between the public, legislators and companies as to the balance of benefits and risks, and it was noted that in the narrow area of giving government agencies a new licence to access Internet data for law enforcement and intelligence surveillance that 'social compact' approach was being taken through the Investigative Powers Bill 2016 (currently at Report stage in Parliament).

In concluding discussion, there was general agreement that the pace of technological change would continue to be rapid. Information and computing were ubiquitous and in the hands of everyone. In historical terms we were just at the beginnings of the digital revolution, and it would be a mistake to imagine that there could at this stage be a stable legal and social framework of regulation. Attitudes to sharing data would change as the advantages became apparent, although it would be wise to expect there to be missteps on the way. It would be important to keep checking what people really wanted from the digital economy as business shited from a product to a service model. The best approach was indeed to see big data sets as part of the national infrastructure upon which normal life would increasingly come to depend and that merited investment. All could agree that more research was needed to illuminate the issues to come in a world of big data.

Sir David Omand GCB

Open this document with Adobe Reader outside the browser and click on the URL to go to the sites on the next page.
Useful reports:

European Commission – Protection of Data
http://ec.europa.eu/justice/data-protection/

Royal Academy of Engineering Report
Connecting data: driving productivity and innovation
www.raeng.org.uk/publications/reports/connecting-data-driving-productivity

Data Science Ethical Framework launch: Matt Hancock, Cabinet Office - speech delivered on 19th May, 2016
www.gov.uk/government/speeches/data-science-ethical-framework-launch-matt-hancock-speech

Competition & Markets Authority Report
Retail banking market investigation: provisional decision on remedies
www.gov.uk/government/uploads/system/uploads/attachment_data/file/523755/retail_banking_market_pdr.pdf

Open Data Institute Report on Data Infrastructure
www.theodi.org/data-infrastructure

Open Data Institute Report on The Open Banking Standard
Unlocking the potential of open banking to improve competition, efficiency and stimulate innovation
www.theodi.org/open-banking-standard

Useful links:

Google UK
www.google.co.uk

Imperial College, London
www.imperial.ac.uk

Information Commissioner's Office
https://ico.org.uk/

Invoke Capital
www.invokecapital.com

Open Data Institute
www.theodi.org

Research Councils UK
www.rcuk.ac.uk

Royal Statistical Society
www.rss.org.uk

Winton Capital
www.wintoncapital.com

The Foundation for Science and Technology
www.foundation.org.uk

A Company Limited by Guarantee,
Registered in England No: 1327814,
Registered Charity No: 274727