HOUSE OF COMMONS

Science and Technology Committee

Oral evidence: Social media data and real time analytics, HC 245

Monday 23 June 2014

Ordered by the House of Commons to be published on 23 June 2014.

Written evidence from witnesses:

- Horizon Digital Economy Research Institute
- Economic and Social Research Council (on behalf of RCUK)
- Web Science Trust and the EPSRC project SOCIAM
- Professor Liesbet van Zoonen, Loughborough University
- UK Computing Research Committee
- The Open University
- Big Brother Watch

Watch the meeting

Members present: Mr Andrew Miller (Chair); Jim Dowd; Stephen Metcalfe; Stephen Mosley; Graham Stringer

Questions 80-155

Witnesses: **Professor Derek McAuley,** Professor of Digital Economy in the School of Computer Science and Director of Horizon, Horizon Digital Economy Research Institute, **Professor David De Roure**, Director, Oxford e-Research Institute and ESRC Strategic Adviser for Data Resources, Economic and Social Research Council, and **Sir Nigel Shadbolt**, Professor of Artificial Intelligence, Web Science Trust, gave evidence.

Q80 Chair: Gentlemen, thank you for coming in this afternoon. First, let me apologise on behalf of some of the Committee members who are unable to be here. We have squeezed this in out of our normal cycle and we are clashing with N other things in this building, but we are grateful that you have been able to come. For the record, I would be grateful if you would start by introducing yourselves.

> ***Sir Nigel Shadbolt***: I am Sir Nigel Shadbolt. I am professor of artificial intelligence and head of the web and internet science group at the University of Southampton.

Professor McAuley: I am Professor Derek McAuley, professor of digital economy, and I run the Horizon Institute, a research institute, at the University of Nottingham.

Professor De Roure: I am David De Roure, director of the Oxford e-Research Centre, which is a multi-disciplinary research centre at Oxford. I am a representative of ESRC in my role as strategic adviser for social media.

Q81 Chair: First, I want to clarify the use of English. A lot of people use the word "infrastructure" in this discussion about social media and real time analysis. What do they mean by infrastructure? Are we talking about people, physical technology or software? What is the correct definition to use?

Sir Nigel Shadbolt: It is all of the above. You have to believe that in this area the hardware process the content in anything like real time at sufficient scale and the software analytics are clearly an essential component, but without analysts capable of drawing pertinent and relevant conclusions to understand the context of the setting you have hardware and software that has little value to it.

Q82 Chair: Is that an accepted definition?

Professor McAuley: I would accept that definition. It is often easy to get confused about the infrastructure being purely the hardware, with perhaps too much emphasis on that, since a lot of this is available commercially, in terms of at scale, and we really want to think about the software tools, the technologies and the people.

Professor De Roure: I would accept that broad definition. It is very good to hear software emphasised because very often it is something that is forgotten or taken for granted. Digital infrastructure is perhaps different from roads. One of the things about digital is that a lot of the infrastructure can go to the person rather than the person to the infrastructure.

Sir Nigel Shadbolt: Data themselves are an infrastructure. Perhaps this sounds a little odd to some ears, but the structure of the stuff we are analysing has an intrinsic infrastructure, so working out what those standards and interoperability points could be will turn out to be quite important.

Q83 Chair: Within the UK are there gaps in that infrastructure?

Professor De Roure: The experience of consulting to prepare for this emphasises repeatedly the skills gap. I know that is the theme of today, so we will be discussing that. I also highlight the software issue I mentioned just now. It tends to be taken for granted and not seen as infrastructure, but it absolutely is.

Professor McAuley: In interacting around social media especially, we might look to the sort of move the Library of Congress in the US has made to grab hold of this as a national resource. One of the things we are doing at the moment, which perhaps is not most effective, is that everyone is negotiating for access to these datasets independently, hence not having any negotiating power to get access.

Sir Nigel Shadbolt: It is a very important point. I am sure we will talk about the gaps around capability. The other thing to note is that we are in a curious situation in which the private sector, which is driven by the interest and intent to understand everything from sentiment to buy and intent to purchase, has substantial holdings of data that hold really interesting social insights. Some of the companies themselves are the very social media companies. There is an interesting question about the balance between the public and private holding and collecting of those data, and whether we need to take a much more balanced view about how we can get the private sector to work in common cause with us, because it has some of the most valuable content.

Q84 Chair: Presumably, from that you would be arguing that the Government should work with the private sector to develop that infrastructure, sticking to that one theme—

Sir Nigel Shadbolt: Yes.

Q85 Chair: —and also to help address some of the skills issues.

Sir Nigel Shadbolt: Yes. If we look even generally at indicators, there was a survey just this year by Eurojobs.com that looked at the 6,000-odd jobs that mentioned the words “data” or “data science” in their title. Half of those—over 50%—came from Britain. The closest, nearest EU competitor was Germany with about 9%. There are lots of jobs out there, and the sense is that currently the skills in that area could provide about a third of the necessary jobs, so a very broad gap is opening up. The area itself is set to grow, by some accounts, by around 100% in three to four years’ time.

Professor De Roure: The kinds of analytics we do with social media data have different infrastructural requirements from other forms of data. It is probably worth highlighting that as well. I would like to introduce the words “real time” at this point. We do not talk so much about real time analytics with other data sources. That has very significant infrastructural implications.

Q86 Chair: In terms of Government, we are in a rapidly moving field. I can name the first Minister I saw not all that long ago—I will not do so today—with a computer on their desk. Things are moving incredibly fast. Do Government really understand its requirements and what can be developed out of the technologies around?

Sir Nigel Shadbolt: It is moving so fast that researchers are challenged too. The rates of change are so enormous. To give just one statistic, every 10 years the amount of information being held is roughly a thousand-fold greater than we currently have. These are extraordinary rates of change. Do we have the capacity? I think we begin to see the requirement. One of the promising things is the way in which the Government and previous ones have begun to talk about the power of information and understand that this is an asset they have to invest in, and the associated skills to bring it to the fore. As Professor De Roure says, the real challenge here is understanding this blend of skills. It is not just about having the ability to ingest the data and crawl over it. It is not just about big data in fact; it is about data of various stripes, and being able to apply a range of skills, from social science, behavioural sciences through to software science, to understand how to make sense of this torrent of data.

Professor McAuley: For example, one might look at the internet of things as a new concept. I was at a workshop at Stanford 20 years ago on the internet of things. We talked about light bulbs on the internet. At one level, the technologists have been thinking about many of these things. What we are really seeing now is a convergence of the technology with the social and the human: technologists who understand the challenges in dealing with human data; likewise, on the social science side, social scientists who understand what is possible with computational techniques. It is in this space that we are really challenged. It is not that everyone has to be an expert on everything, but it is the dreaded t-shaped person who has experience and understanding across a number of disciplines, and there is an awareness of that, yet depth in one area. That is where we are seeing it is not purely technology-driven any more. I am a computer scientist. I will be the first to admit that it has been a learning experience for me to get to the point of understanding all of these social nuances about how we have to deal with human data. One of the key challenges is not just the purely technical skills but broader awareness of the challenges in dealing with these sorts of data.

Professor De Roure: Reflecting on those comments, I think the agility of that capability requires us to be able to remove barriers, not repeatedly investing in existing silos of activity in research areas. Something like social media really is interdisciplinary, as we have heard, so we need to be more strategic in those investments. That is certainly the mood of the research councils to work across.

Q87 Jim Dowd: Sir Nigel, if I may pursue a point you made a few moments ago, if I understood it, the amount of information and data around grows by a factor of 1,000 every 10 years. How does that correspond to the half-life of facts?

Sir Nigel Shadbolt: One of the great challenges is the dictum: a rumour is halfway round the world before truth has got its boots on. The web and social media applications behind it are often seen as propagations of rumour mills. There is no end to the amount of opinions, rumours and gossip that might be exchanged in this milieu, but, in truth, the increase in our understanding of basic facts in this space is also exploding. So it is interesting to see whether real facts and opinion are keeping track of that, but because of things like the internet of things we are simply instrumenting so much of the environment that there are new kinds of facts being stored and laid out all the time. Some of these will be material to making judgments about social mobility, social attitudes and social interactions.

Q88 Jim Dowd: Where does the dictum that the more you know, the more you know you don’t know fit into all this?

Sir Nigel Shadbolt: That is a really interesting point. One of the things we are all aware of is that there is no one discipline that has the essential insights into how to make sense of these data. Our written submission was from the Web Science Trust. A set of laboratories networked around the world tried to look at the web as a systems level object. It is not a piece of technology; it is not a set of software programs; it is humanity-connected. As such, you have to pay attention to understanding the methods you need to make sense of this, because we are literally seeing new forms of interaction the mathematics for which barely exist to do the analysis. The sociology behind it has been in the background for some time. It is now possible for sociologists even to test some of their theories and to put them to empirical test. We are seeing a new set of disciplines emerge in the face of this new data opportunity.

Q89 Jim Dowd: I have a few questions on the amount of what is called social media data necessary for social media organisations to hold. Should they be allowed to require rather than just request information? I recognise that organisations of all kinds, whether in the private or public sector, can never have too much information about you. I was putting a flashlight app on my phone the other day, because the one I had before was not very good, and the phone told me that I had to switch on my locator before I would be allowed to download it. Other than simple acquisitiveness on the part of the company wanting to know who is out there, why on earth do they need to know where I am before they will let me have light?

Professor McAuley: Of course they don’t. We are now going through a period when many of the things being deployed are being, let us say, a bit duplicitous in their behaviour and they are asking for things. They are presenting one experience, yet asking for a lot more information. I know of a research project, which would not have gone through ethical approval at my university but did go through somewhere else, involving a simple psychometric fun program. What sort of person are you? Its purpose was to determine your social graph. It would have sent the heebie-jeebies up my ethics committees if they had ever seen something like that. A lot of this is going on.

At the moment I am leading a working group looking at the principles of managing personal data for the Information Economy Council. We are reporting in two weeks’ time on ethical guidelines for how companies should process this sort of information. In one sense, industries are crying out for this. Many small companies, even large ones, want to be seen to be behaving ethically and are getting somewhat annoyed at some of the unethical behaviours of others. I would put that flashlight program in that category. It does not need that information to do its job; it is obviously after it for some other reason.

Q90 Jim Dowd: You never know whether that is for its own purposes or to sell it on, because this information has value.

Professor McAuley: Indeed.

Q91 Jim Dowd: Would it be practical to introduce a restriction on the amount of data these organisations are allowed to request up front? I understand the argument, “You have to accept our terms and conditions or you don’t do business with us,” or, “If you want to do business with us and don’t want to accept our terms and conditions, you don’t do business with us.” Is there a practical role for imposing restrictions and limitations on the amount of data and information these organisations can either request or certainly hold permanently?

Professor McAuley: My comment would be that we already have legislation in this area, and possibly forthcoming legislation under data protection regulation. That gets us to a position where, if the user is considered to have given consent, you can take what you want. There is a question, again, as to whether that is ethical. We might want to be thinking about some frameworks, even if they are voluntary, such as those related to organic food supply. There is a regulation about what food products can be on the shelves, but there is another branding associated with things that are organic. There may well be a branding associated with highly ethical personal information products in the future, rather than it being perhaps a requirement, because the underlying legal basis is one of consent. This would be some sort of critical kitemark to say this meets a higher standard, and perhaps then the population might be able to think about it and reflect on what they do and do not use.

Professor De Roure: I think that contention is very important. There is definitely a case for increased awareness. How few people read the terms and conditions they are signing up to is documented.

Sir Nigel Shadbolt: They are usually totally impenetrable of course. There is some very interesting research on this at the University of Nottingham, looking at the reading age and how complex these terms and conditions are. Some of it is more complex than Shakespeare and needs a reading age of 19.2 years to get through it. It is an area where a degree of practical simplification could help.

Q92 Chair: How do you do that, because most of this is written for American corporates and not our systems?

Sir Nigel Shadbolt: Yes, although I suggest that the interest those companies have in a UK or European audience is such that they would take it seriously if there was a real sense that without that they would not get access to those markets. I think that is beginning to shape this.

Professor McAuley: I highlight the recent ruling of the European Court that has been publicised mostly as the right to forget, but the precursor judgment to that was that Google had to comply with Spanish data protection law. That was the more important judgment, because that led to the second part. The multinationals would be based not even on new incoming regulations but on the existing human rights legislation, and everything else would be required to operate under national laws. I think it is a landmark ruling. One of the things we might aim to do is make the UK a better place to do business by having a set of, say, template contracts—a bit like food labelling—where it is quite clear what you are getting and what is happening to your data, not necessarily as the final straw but as a voluntary code for people to sign up to. “We make our tools and services easy to understand” could be part and parcel of the development of that, and setting up the UK as, “We are an easy place to do business. Consumers can understand what is going on.”

Q93 Jim Dowd: As highly experienced and trained practitioners in this field, are you as amazed as I am about the volume of personal information that people would give away for nothing?

Sir Nigel Shadbolt: There are good studies on this, and apparently for a free cup of coffee you give away a surprising amount, except that when the consequences of that are played back to the individual it is clear that the notion we have given up on privacy, or that we will trade almost anything, has not been the experience in areas where we have gone the wrong side of the argument. You might give it away, but the presumption as to how you give it away and how it will be used turns out to be very important. Even in the case of the commercial situations, when people are presented with the consequences of that, they think about whether there should be a better balance. The problem with deciding that these data shall be collected and these data shall not is that there are always proxies for the information you decide not to give that can still reveal a huge amount. It is unlikely that a transaction between you and your supermarket would be something you would want to alter or change, but it is clear from the entire history of that, as we know in the famous example of the Target supermarket in the US, that 20 products can predict whether or not you are pregnant and pretty well your due date. That is the reality of understanding shopping basket compositions. While that is happening, the regrettable other side is that we cannot get into information-sharing situations with these corporations where we can understand that over-the-counter non-prescription cough mixtures and medicines might help us predict and see outbreaks of flu and the common cold in a population, because they will not share the data with the Government. There seems to be a degree of asymmetry here about how analytics are used and who gets to use it and who gets the insights. That would be one of the major correctors we could look at.

Q94 Stephen Metcalfe: You said the Government cannot get this information shared. The Government undertake quite a lot of horizon scanning activities, looking at emerging technologies. In this particular field have the Government done enough to be ahead of this and understand the implications of what is happening, or is it another case of them napping while the world around them turns and develops?

Professor De Roure: We really welcome the emphasis on data and data sharing and all the issues that go with that. I would suggest that looking at data sharing alone is not sufficient. It is essential but not sufficient. It is what you do with the data that counts. Sharing the tools, the results and practice, and building best practice, has to be part of sharing. Investment in data is worth while if it results in data that can be discovered and reused, and preferably accrues value through reuse by the community or the beneficiaries of that analysis.

Q95 Stephen Metcalfe: But the question was: are Government ahead of the curve on this?

Professor De Roure: No. I think social media data are special, which is a point I will continue to make, partly because it is social, which goes back to the previous point. It has to be treated as separate. It is easy to say it is a new form of data, like business data, supermarket loyalty cards and admin data. It is that, but it is something else as well. If you look at something like the summer riots, that was partly mediated by the use of social media. New social processes are being created—for example, on Twitter—which deserve analysis in their own right. That is a new and important phenomenon to the country. I would strongly argue that we need to treat social media at least with the attention we are giving to other new forms of data; and it has, as we said before, its own infrastructural requirements, especially in terms of analytics and real time analytics.

Professor McAuley: It is seven years since the Research Councils UK launched the digital economy programme. I think that was foresight, given that the iPhone is only seven years old. If we can remember what the state of social media was seven years ago, it was a bit of a mess. Maybe people had heard of Twitter, but it was still viewed as a slightly eccentric thing to be engaging in. The rate of change in terms of uptake, especially promoted by the fact that the technology is suddenly available in everyone's hands, and the velocity of this is so high that we have all been slightly bamboozled by how much people have engaged with it.

I put a caveat on that. Often, folks will say, "We'll take the current trends we have observed over the last few years and project them, like some hockey stick, into the future." One of them is that many of the initial social media mechanisms were very open by default. You were publishing to the world. With something like Twitter, where that is its modus operandi, it is broadcast in full, so that is understandable. Many people got caught out by social media that was by default broadcast and did not really internalise that. That is why in the last few years we have seen the emergence of new social media which regard privacy and small group interaction as much more important. My comment would be that the speed at which this has moved has defied many of us and we are working hard to keep up with it. I certainly would not blame Government, or anyone else, for not having had the foresight seven years ago to see this one coming.

Q96 Stephen Metcalfe: Is it now catching up? Has it caught up? Is it doing what it can to support this technology and the potential? We are told it is worth hundreds of billions over the next four years. Is it now getting it right?

Professor McAuley: I am sure David can speak to this from the point of view of ESRC, but I think there is still a lot more to be done. In particular, we need to be careful that we do not focus on the social media of today, because in two years it could be gone.

Sir Nigel Shadbolt: That is exactly right. The remarkable thing about this is that the research to tell us what we can infer from this social media is only now landing in front of us. Your sexual orientation, disposition to vote a particular way, your religion and ethnicity are predictable with a very high degree of certainty from your Facebook likes. This is new research. It will take a while to internalise just how we deal with that. What are the processes to put in place? We can imagine any number of people who would be interested in knowing and understanding everything from the last Obama campaign to what one can expect in the election campaigns here. This will become material information that is argued over, and the research will simply go into this area.

Is the infrastructure there? Is the capability there? Are the Government doing enough? It is easy to confuse big data with insights in social media. We are making significant investments in this area. The Square Kilometre Array and the announcements about the eight great technologies that the Science Minister has made are good investments, but in this particular area of social analytics it is a different kind of content, skillset and blend, and it is moving so very fast that we have not had enough time to put in some of the infrastructural elements we need to support it.

Professor De Roure: I agree with that and I should underline that the intent is there. It may not have been realised yet, but the degree of interest within the research community is there as well. The researchers are very enthusiastic to work with social media, both to study it but also to use it as a tool for their studies in social science. I would hope for investment in the coming year or so in social media which perhaps would mirror some of the investments in admin, retail and business data.

Q97 Stephen Metcalfe: Bearing in mind the speed at which this is all developing and changing, if we want to maximise the financial benefits of this potentially, who should be leading the development of the products that will realise its potential? Is it the private sector, or, because it is moving so fast, does it need to be academia?

Sir Nigel Shadbolt: The UK has led the way in a couple of areas, not in social data but in open data. It led the way by releasing it and thinking about the kinds of institutions and infrastructure that need to be set up to support its analysis, and how to stimulate the demand side as well as the supply side. We might think the market will take care of all of this because there is such a strong set of retail and commercial interest in social media, but their analytics will tend to be very much closed. The methods they use will generate insight for them and their clients. The reports are available typically for very large amounts of money. The challenge is to provide a public set of insights so that the public sector—the public good—can be as well served as the private good.

I am not saying the commercial sector does not have an interest in collaborating on this—it absolutely does; but in some areas it will regard this as particularly valuable to its own positioning and customer insight, so we may well here have a situation where the provisioning of tools available for wider research for the kind of data analytics that Government have to do will need to be made available on different sorts of licensing conditions such as open licences.

Professor McAuley: I talk to quite a number of companies that have had access to a lot of personal information for many years. They already are challenged ethically around what they do. They choose not to do certain analyses—for example, to determine who you might vote for, what your sexuality is or any of these personally invasive things. They choose not to do it. There may be a valid research purpose in asking some of those questions; it may be something that as a society we want to understand, but, quite sensibly, certainly established businesses with a respectable brand would not wish to ruin that by going down that path for a commercial reason. They would not want to be seen to be doing that analysis for a commercial reason, even though that sort of analysis done anonymously and correctly, as we might do with the census, statistically and at aggregate level can still provide an enormous social good. But they would not wish to do it because of the dangers of being seen that they were pursing this only for commercial gain.

Professor De Roure: Social media are a great opportunity for businesses and academia to come together with common interests and methods. An additional role for academia is in the production of people with the skills and business needs. To come back to the “skills” point, we are told time and time again that we are not producing enough data scientists who can do this kind of work.

Q98 Stephen Metcalfe: What do you think should be the role of small business in helping to develop some of these products that might analyse this?

> ***Professor McAuley***: I also work with the connected digital economy catapult. One of the things we have been doing there is reaching out to business to ask them what their challenges are. I would bluntly categorise small companies into two types: those who are extremely sensitive to the ethical issues around the data, and those who have not even thought about it. It is not that the latter group is in any way evil; it is simply that it is driven by technology and a can-do attitude. On the one hand, the former group are appealing to us, "Please could you come in and certify that we handle these data in a certain way, such that you can evidence this and build it? We have no ground to rely on, so how can people trust us?" versus others who say, "We have this technology; it's great; we can do this stuff." It leads to tension between innovation and regulation. Some of the most innovative things have come about by slightly stepping over the line, realising that and then stepping back. I see huge amounts of creativity in the SMEs, and one thing we could usefully do is make them more aware of some of the ethical challenges that this sort of processing causes.

Q99 Stephen Metcalfe: That would be improved by better collaboration.

> ***Professor McAuley***: For example, yes.

Q100 Stephen Metcalfe: How would you go about it? What steps are needed to improve the interaction between small businesses and academia?

> ***Professor McAuley***: I find that for small businesses, especially at the start-up end, keeping food on the table is the important thing. They need very practical advice that is directly related to the sort of thing they are doing. They would probably find most of the papers we publish in the academic journals impenetrable, given that they are written for a very particular audience. In a sense, we need a translation from what is often very generic, abstract work into very practical sector-specific, even context-specific, areas. In many ways, it is not necessarily a property of the data. Data by themselves have in a sense no value or ethics associated with them; it is what you do with them, and each of these companies needs to have a case study or something they can understand and relate to. That translational work between the very generic and abstract academic research into the very practical lessons for these SMEs is vital.

Chair: Graham—sorry, Jim.

Q101 Jim Dowd: Okay; thank you. You caught me out going out of order like that. It shows an unexpected degree of originality from you, Chair. *[Laughter]*

If I may make a variation on the earlier line I was pursuing as to the tension between having good, sound analysis of data and the need for informed consent, to save you tiring yourself with the cliché that it is a question of balance, which option would be the primary concern, and why?

Sir Nigel Shadbolt: A prior problem is that the presumption is that it is a one-way deal at the moment for most people who give up their data either to Governments or commercial organisations. One of the things that needs to be rectified—we are starting to see this happen, partly post-Snowden—is the rectification of this information asymmetry. People want access. They do not want to control it necessarily, but they want to know they can get their information back and out, and that they have some rights and entitlements over it. We will start to see a swing very much in the judgment, opinion or assumption that it is our data, whether it is health records or our social interaction record. This is something we are going to see come increasingly to the fore.

Then the question is: what are the analytics that can and cannot be performed on that? We will then find people being asked much more to consent for all sorts of reasons, whether it is medical research or social analytics of a broader sort. There may be some mandated areas. The census is a good example where we can imagine using many of these as proxies for certain sorts of census data.

A lot of the challenges here are trying to get a mindset where we can imagine that in this world the device formerly known as the mobile phone—this device—is your personal information device. You take your photographs on it; you record your music on it, and why should you not be carrying increasing amounts? It does not all have to disappear into the cloud. There is a very significant sense in which you can act as the custodian of this information. The question then is: do you get personal analytics on this? A very good example is well-being. There are any number of fitness monitors that you carry around to monitor how and where you jog. That data could become hugely important at the stage before primary care. The assumption will be that those are my data. Unfortunately, at the moment I upload my data through an application. That application goes to somewhere in Finland, and the small company that started this fabulous app called Moves gets acquired by Facebook. Now all my data, exquisitely detailed, have gone somewhere else. We have not really understood the food chain enough to realise that the analytics you might do could end up in a place you did not anticipate or did not expect, and did not think you had permissioned.

Professor McAuley: It is increasingly the case that we can find a lot of these interesting applications by the fact we are sharing data. A lot of the analytics could be performed on my personal data in isolation. That is what I refer to as the 1990s Microsoft model of business, which is, "We'll ship you some software that you can use in a private context to edit whatever documents you want," to be contrasted with the modern CloudWorld where I have no idea who is looking at my documents if I am online editing it on Google Docs. I have no auditability of that or visibility of what is going on. As for a lot of the existing models of, "Ship us your data and we'll process it and give you this value," there is absolutely no reason these data have migrated into the cloud and the analytics on your own personal platform.

Then we might turn it around and say, if someone gave me an application that was really valuable and said, "By the way, would you mind sharing these derived statistics from your information?", I would say, "Yes, that's okay, because how many miles I walked today is not particularly sensitive. Of course you can have that." Many of these companies would say that those are the data they want. They do not in a sense want your personal data; they want some statistics from your data. There is a question technologically about how we might implement that in the future.

Q102 Jim Dowd: Could they not just anonymise it?

Professor McAuley: They can anonymise it; that is one thing.

Q103 Jim Dowd: But they won't.

Professor McAuley: The key thing is that, if you are among the truly paranoid, would you believe they have anonymised it? What measure would you have to know they have done that, rather than remembering your IP address and unique identifier from your phone? They have asked for your contacts, and there are many other things. The truly paranoid would say, "Even if they say that, how do you know?" For the companies that are honest and true, one of the things they might welcome would be, "Come and audit us." In fact that is one of the requests we have had. We did an online survey recently where 60% of the respondents said they would support third-party auditing of their privacy policy. They want someone to come and verify that it does what it says on the tin. They say, "This is our privacy policy, and, yes, by God, we do it." That includes some very big companies.

Q104 Jim Dowd: That would imply a regulator of some kind. Whether it is industry-funded or Government, there would have to be an interventionist body of some kind to do that.

Professor McAuley: Indeed. I simply look at other aspects of human life. Whether it is organic food supply, green building regulations or financial audit, in order to have faith in the system as human beings we traditionally require a third-party expert to look over it and say, "Yes, actually, they are doing this."

Q105 Jim Dowd: Like Arthur Andersen.

Professor McAuley: I would not like to jump to who might do it, but that might be one. Financial audit is a very good model to use, because it has to scale from the small company with three shareholders to the multinational. Arthur Andersen do not deal with a company with three shareholders, so one would need a mechanism to scale from small to large.

Q106 Jim Dowd: The surveillance of social media by GCHQ has become an issue of late. How do you reconcile—if, indeed, you do—obtaining user consent with the demands of national security, given the fact that, as Professor McAuley says, to those who are truly paranoid "national security" is simply a cloak used by Government to do nasty things to their citizens?

Professor De Roure: That is a very important conversation. I am concerned that by focusing on GCHQ and surveillance we are looking at only part of the picture. If we expand that to law enforcement and the use of social media to help the police in enforcing criminal legislation, we can see clear benefits. A word I would like to use here is "asymmetry," which Nigel used in a different context earlier. We would not want through legislation to constrain our law enforcement, or our national security, to have less power through the use of social media than those who are threatening us in some way. I think the principle of symmetry is really important.

The other way in which this discussion enlarges very importantly is that it is not just about the individuals, but the companies, the universities, the security of the data and the use of social media data in all those contexts to protect companies, IPR and ensure there is data integrity in our organisations. So it is a much bigger discussion beyond GCHQ and surveillance space.

Sir Nigel Shadbolt: Also, it is important to understand that there are legitimate issues to be addressed, which I know are very much in the minds of politicians and the agencies themselves, around bulk collection. One of the challenges, because the technology exists, is to collect bulk and then decide how you pay attention to it. People imagine that that is essentially the world we are in, whereas in many aspects of our digital age the data are paid attention to only if there is some sense that it exceeds a threshold; otherwise, it simply boils away. When you go past speed cameras, typically they activate only if you are breaking the law and then your image is taken. That feels a much more proportionate response.

In some sense, we are going to move to a situation where we will be living in this extraordinary digital panopticon where huge amounts of information are coming in all the time. To imagine that we have to find some safe place to store it where we can pore over it at leisure seems in a sense disproportionate, and possibly a poor use of financial resources and technical capability. It is much better to have a sense of what you think you are interested in so that you can pay attention to the important signals. That is one area where social media research is very important. It is not as if this stuff does not have a background theory. We have to have a sense of why we are looking at the information in the first place. Historically, completely open, large-scale fishing trips have not been extraordinarily successful. You need insight to direct your search in these contexts.

Professor McAuley: Many of the protocols that have existed for years would have made the mass surveillance we have had not possible, but for commercial reasons, because it costs a few more CPU cycles, or whatever, they have not been deployed. We are now seeing the reaction to that, which is the deployment of, "We will encrypt all e-mail connections," because this behaviour was not expected.

I go back to asymmetry. It would be one thing to say that in the future we will find that many of these channels will be encrypted, so interception and flight will not be possible, but for it to be possible for some nations to access everything no matter where it is on the planet, by the issue of a court order in those countries, whereas it is not applicable the other way, does not seem equitable, and one may want to ask what is reasonable internationally with partners in this regard.

Professor De Roure: Reflecting on the comments made about this and the previous question, we seem to be giving lots of examples that would cause citizens to have a reluctance to release the information. I am a little concerned about that, because we should also be explaining the benefits of the analysis of this information, which we have discussed less. I would like to balance that a little bit. The positive stories resulting from social media analysis need to be propagated as well.

Q107 Chair: Professor McAuley, you mentioned in response to a question just now that 30% of users wanted a third-party audit.

Professor McAuley: Sixty per cent.

Q108 Chair: You referred to some research. Is that published research?

Professor McAuley: It is a survey, and that was the data as of last Wednesday night. We are preparing to present this to the Information Economy Council on 1 July. We floated a document for consultation on practices around personal information which had, in my opinion, two substantive elements over and above what has been said so many times about personal data, access and control and consent. One point was simplification, which includes templated contracts instead of bespoke legal contracts for every website, and an iconic representation of them: "This is what we do with your data." The final one was: what was the appetite for third-party compliance checking against that, including ethical process for approval of new products and services, and privacy by design; was the system designed to ensure your privacy?

We had an open meeting last Thursday and had an online consultation the week before that. The statistic was that 70% agreed, or strongly agreed, that third-party regulation would be a good thing; 20% sat on the fence; and 10% said they disagreed. To me, that shows there is an appetite in the UK among companies for this. It means there would be resistance to legislation, but if 70% of them agree, and strongly agree, I think it would be good to respond to that request.

Q109 Chair: This is all coming into the public domain formally on 1 July.

Professor McAuley: Yes.

Q110 Chair: Perhaps you would be good enough to let us have a copy of it.

Professor McAuley: I will do.

Q111 Graham Stringer: My apologies if you covered part of this earlier. I missed the first 10 minutes of this session. Sir Nigel, you are responsible for open data strategy across the UK Government. Can you explain what that is, how close we are to achieving standards for open data, and just what that means? I find it difficult to know what a standard for open data would be.

Sir Nigel Shadbolt: Yes, indeed. We began this work in 2009, and it has been very successful. It was all about releasing primarily non-personal Government data that Governments collect, usually at taxpayers' expense, which is everything from infection rates in hospitals to where bus stops are located and transportation and education spend data. These have typically not been social media data; they have not been at the personal level. When we refer to open data, we are very keen that the debate respects and understands the varieties of data now in play on the web, from personal data, to anonymised data, to non-personal data, to big data and relatively small data, but nevertheless incredibly valuable data, that sit on spreadsheets in various parts of Government. We have released much of this data.

When we talk about standards, we mean the ability not to stay locked within a particular piece of software and not be used by another piece of software. Famously, we do an awful lot of processing of spreadsheets and are familiar with that. There are ways to store the data in a way that another spreadsheet program can ingest the information, not a particular product issued by a particular software provider. There are interoperability standards. There are standards around how you refer to and describe the data, and how they relate to social media data. Often, you can argue about whether social media data are open or closed. Things like your Twitter stream are typically open; it is going out there into the open.

A very important aspect of the work going forward in this area will be to agree standards as to how we can represent and allow that data to be analysed, because at the moment, frankly, people are dreaming up a particular format or way of representing this information and going ahead with it. One of the interesting research disciplines is to work out how we can, with a little engineering, allow for a much easier flow of information between our analytics.

Q112 Graham Stringer: Are you telling this Committee that you have achieved the impossible and got different Government Departments to be able to access each other's data?

Sir Nigel Shadbolt: We have made pretty good progress, actually. Technically, it will be no surprise to this Committee that the real challenge is often organisational or cultural. We know how to fix the technology, but it is a matter of understanding how we can allow that information to flow. This will be an interesting challenge. If Government are serious about trying to use the product and insights of social media for themselves and their own Government Office for Science and departmental analytics, this sharing will be essential.

Q113 Graham Stringer: How important is access to Government data to organisations that are interested in analysing social media?

Sir Nigel Shadbolt: Very much.

Q114 Graham Stringer: Can you put figures on it at all?

Sir Nigel Shadbolt: Certainly, particular outputs, like the index of multiple deprivation and a variety of ONS datasets, are hugely valuable, because they talk about social mobility, and I think it is intrinsically billions of pounds-worth of value.

The first instinct is to say, "We must be charging these people for this." In that case, the whole premise of open data is that by making them openly available you allow a whole range of companies to participate in and use those data to drive insight around business and the provisioning of services, but the Government-generated data that relate to everything from education to health and social welfare will be of highly material significance. The kind of work on which the ESRC—the Economic and Social Research Council—focuses much of its attention is precisely this sort of so-called administrative data. David, is there a value for this?

Professor De Roure: We could go away and calculate that.

Sir Nigel Shadbolt: It would be very large.

Professor De Roure: Yes.

Q115 Graham Stringer: You gave a definition of standards partly in answer to this question. I do not know if it is a meaningful question. What should confer standards on our open data? Are standards the doorway to making sure you are as open as it is possible to be?

Sir Nigel Shadbolt: I think it is one of the underpinning foundational principles. It seems very obvious. The Government have been doing some quite good work in this area. The Government Digital Service has been looking to mandate open standards, so these are not standards that are proprietary or owned by a particular company or sector. These, along with licensing and a whole range of now quite well understood components, remove friction from the system. Standards have bedevilled everything from health care to welfare to credit payments over the years. The web is a great example—perhaps the best example, along with the internet—of a global fabric which has worked because its standards are open, understandable and freely available.

Professor McAuley: I would add a comment on the importance of standards in releasing data. We want to compute with this data. It is no good releasing it in a text document or PDF. It was data in a spreadsheet, and it gets turned into a table in a printed document. That can classify as open data. The data are not open. This requires human beings to process all this data and probably type it back in. One of the key things about open standards is to get to the point where the data are released so they can be immediately reused computationally rather than with human intervention. A key tangent on this is that we have been working with cultural institutions to release their content as data, not as a website for people to experience, or for a human being to experience, but so that it can be repurposed and reused. The key thing is to get the data so that they can be reused, repurposed and used by many tools if they are in a standard form and accessible to a computer, whereas, "We have published this; it happens to be in this document" does not make it reusable and repurposable.

Q116 Chair: Sir Nigel, you said in response to Graham Stringer that within Government, sharing is essential. You may be aware that I did some work with Government at the beginning of the 1990s when the issue was not about the technology but the people issues and creating a management structure that has fuzzy boundaries between Departments. Is it still the same in this era of dealing with social media data as well?

Sir Nigel Shadbolt: We are starting to see an understanding that the value of information and data exchange is so high that silos are understood to be unfortunate and get in the way.

The real issue is to incentivise officials at various levels to understand why this is beneficial, there are cost efficiencies and better information flow as a result.

To come back to the whole issue about capabilities, unfortunately this is an area where effective and agile use of data is something the Government have to look at. Have they really got the mix of skills in the civil servants? It has them in a few specialist areas, but generally is there an awareness of what is required here? Even if you can incentivise and get the cultural and organisational pieces straight, do you have people who would actually know what it meant to say what Professor McAuley has just said about the whole idea of a PDF versus an open exchange format for a spreadsheet?

Chair: Of course, all of that is much harder than the technology on many occasions.

Q117 Stephen Mosley: That leads very nicely to what I was going to ask, which was about skills and the UK skill base. Professor De Roure, the ESRC have said that the UK data capability strategy was not sufficient in the area of training. Could you explain what you meant by that and say what you think should be included?

Professor De Roure: It is very good as far as it goes, but what we are looking for is to drill down further in the implementation, very much along the lines we have been discussing here. This is one of the reasons we welcome this inquiry because it will emphasise the areas that need greater attention.

"Data capability" is a good phrase. because it is not just about the data but the capability, which requires all the things we have been discussing from the human skills to the analytics skills, the core science and situating that in real examples of UK life, working with the software and the exchange of methods and practice that we have been discussing here. I like the PDF example, because even in our scholarly communication process we are still using techniques that are 350 years old—we should celebrate it—to exchange information. In the digital world and the world of social media there are many new ways of doing these things. There is an emerging social media methodology. New and best practice is beginning to form, and that really needs attention.

Q118 Stephen Mosley: Expanding out a bit, do you see a lack of individuals with the right level of skills? If so, what effect is it having on the social media analysis sector?

Professor McAuley: At the digital economy catapult we have been hiring a number of senior technical architects. We had to pay 20% more for the people who had the data science skills, so it is clear that there is a premium in the market for them right now, which would indicate to me that there is a shortage, never mind that they took a lot of finding. When it comes to social media, it is even worse. Those are pure data science skills. People who have the appreciation, when dealing with what I refer to as human data, that the consequences of the analysis could do harm to individuals and have a sensitivity to that is something we need to be training up. For example, if we look at something like the Turing Institute announced in the last Budget, I would like to make sure that they have an ethical strand to all the statistics and analytics they do, because that is where we have seen a lot of the impact happening. There is a shortage.

The data capability strategy of the research council mentioned that they were bringing forward a proposal for a series of data analytic centres. Much of the data they were talking about was human data—it is not particle physics or the Square Kilometre Array—and that requires a sensitivity to the social sciences that we have a severe shortage of.

Sir Nigel Shadbolt: We have not fixed this if we just throw tin and fat pipes at it. The big computing infrastructure is required, but it is also about human analysts, and it is not just retooled computer scientists. I speak as a computer scientist. We need to bring in people with the social science skills and skills in statistics and experimental design so that they will know how to take the data and turn it into information and put it into a wider context. These are not assemblies we have routinely been putting together in our universities. We are starting to see an awareness of that, but I would urge that, when we think about supplying capability from the higher education sector, we understand that we need these mixed methods. A broad set of disciplines is required from mathematics, statistics, behavioural and social sciences to computing.

The requirement to get data analytic centres which have a range of perspectives and orientations could be really powerful here. I refer to a bunch of people who look at the data from the point of view of retail analysis, financial engineering or social exclusion. We have not talked much about that today. What about those sub-populations who essentially are not well represented in social media? Can we identify the missing footprints as well as the digital footprints at play here?

Q119 Stephen Mosley: So far there has been very much focus—it is understandable, given your backgrounds—on the public sector: universities, research councils and the Turing Institute. Does the private sector—business and start-ups—have a similar problem, or does it find it easier to recruit?

Professor McAuley: Start-ups present their own special case, but I would like to take a couple of large companies: a retailer and a telecommunications operator. I will not name them because it is under commercial confidence. They are both talking about a charter that they would publish concerning how they process data so that their customers would know what they are and what they will and will not do, and try to build up that relationship of trust about how they will process the data. It would be very easy for a large retailer to find you, given a customer loyalty card, on social media. They currently do not do this and do not wish to do it for fear of what the backlash would be. For those companies it has been a painful learning experience, including having the data analytics people saying, “We can do this,” the marketing people saying, “It’s great to have more information,” and the people who own the brand saying, “Please don’t mess with our brand by doing this bad thing to our customers.” In that sense, they are the ones who are representing the social science side.

Increasingly, even large fast-moving consumer goods companies are concerned precisely about this issue and are engaging with social scientists, as well as technologists, on this. There is a real demand in industry. The problem with SMEs is getting together enough people who have that skillset when you are a five-person operation. You need a sociologist. You cannot afford a full-time sociologist; you are too busy. You have three developers, or whatever, and someone running the company. What you need is guidance and information about what to do. What is safe and secure and what would customers expect, rather than unnecessarily being able to bring together these multi-disciplinary teams?

Sir Nigel Shadbolt: We have talked about the UK context here, but the Web Science Trust has fostered a global lab-based network of 14 where people are doing analytics in this space and are trying to put together the interoperability to allow us to share insights. There is an issue here about how we learn from and can inform the global social media analytics in this space.

On your point, I do think these companies are well aware that at the moment there is an absolute race to secure the best of a very scarce talent—the so-called data scientist.

Chair: Gentlemen, thank you very much. It has been an extremely interesting session.

Examination of Witnesses

Witnesses: **Professor Liesbet van Zoonen**, Professor of Communication and Media Studies, Loughborough University, **Professor David Robertson**, Head of School of Informatics, University of Edinburgh, representing the UK Computing Research Committee, **Dr Mathieu d'Aquin**, Research Fellow, the Open University, and **Emma Carr**, Acting Director, Big Brother Watch, gave evidence.

Q120 Chair: Good afternoon. Thank you for coming in this afternoon. All of you have been sitting here listening, so you understand the format. It would be helpful if you could introduce yourselves for the record.

Professor van Zoonen: I am Professor Liesbet van Zoonen. I am professor of media and communication at Loughborough University. I lead a big research project about public taboos and desires around identity management and the sharing of personal data. It is very much focused on what people want and do not want.

Dr d'Aquin: I am Dr Mathieu d'Aquin. I am a researcher in data technologies at the knowledge media institute of the Open University. I am currently focusing a lot on the data infrastructure for the Milton Keynes future cities project.

Professor Robertson: I am Dave Robertson. I am head of the school of informatics at the University of Edinburgh, but I am really here representing the UK Computing Research Committee, which is an expert panel of the British Computer Society.

Emma Carr: I am Emma Carr, the acting director of Big Brother Watch, which is a civil liberties and privacy campaign group based in the UK.

Q121 Chair: Thank you very much. We touched on this directly and indirectly with the previous panel, but who in your opinion is most likely to misuse data—the Government or the private sector? Do the Government need to approach social media data usage differently from how a private company might?

Professor van Zoonen: From our research, we can say that the UK citizens' approach to this is that they tend to believe that the Government are a bigger risk than private corporations, although that is changing very rapidly due to practices by Tesco, for instance, that are becoming more well known. If you look at what people think about this, they still think the Government are the biggest risk.

Q122 Chair: That is Government per se. Does it break down into different Departments having a greater degree of trust?

Professor van Zoonen: Yes. It also breaks down into local government having more trust than national Government. The further away people are from national Government, the bigger the distrust of national Government and the trust in local government is. Of course, those different levels of Government are not included very often in the discussions we have about big data. We tend to presume that it is a global, national discussion, rather than also a local one.

Q123 Chair: A while ago I saw some data that suggested that one of the Government Departments fairly high up the list in trust—relatively—was the Treasury, which seemed to me to be an absurdity, but there are some odd human reactions to Government Departments.

Professor van Zoonen: Yes. It also depends a bit on which people you ask and what their political views are. There are definitely groups of people who have a high degree of trust in the legal institutions, the police and the secret service, and groups of people who do not have that. In that sense, there is not an overall picture.

Dr d'Aquin: Unlike Professor van Zoonen, I cannot talk about people's perception, but our research shows—besides what we might not know about Government capacity—that the organisations that will have the best ability to misuse personal data are certainly private companies, especially large-scale private companies not located in the UK, including the big social media platforms in the US, as well as companies that are not necessarily directly associated with social media but that collect personal data as a side effect of their activities. Of course, Google comes to mind. Some of our research has shown that, looking at the web activities—the web use—of average users, the biggest amount of disclosure of personal data will go to systems such as Google Analytics. That also includes mobile phone providers, internet providers and all the companies that have the capacity to collect very large amounts of personal data, which could, if they wanted, be largely misused.

Professor Robertson: It is really difficult to identify where the threats will come, but there is a special responsibility for a Government in all of this. The one thing a Government can do that is really hard for others to do is set an example. Open data is a good example of that. Government can set an example by backing initiatives of that kind. That is the limit of it—you do that and see what follows and so on and so forth. To some extent, you can have regulation as well, but that kind of light-touch example setting has already been pretty effective in many ways. Maybe you could go further, but it has done a lot.

The other way round is to apportion blame. I will leave that to people who are more closely involved, but I expect it is really hard. It is really brittle as well. It is very difficult to know at exactly what point you lose trust in some of these systems because they are so pervasive. It is difficult to know where the rubber hits the road with the system and where it will have an influence on people—or not, as the case may be.

Emma Carr: Depending on what kind of data you are talking about and what the use of those data is going to be, the polling seems to indicate that the public have a very similar outlook on the public and the private sector in terms of how safe their data are with them. However, generally, people hold the Government to a much higher standard when it comes to data protection. That may be because people know that private companies are usually using their data to make money—for advertising purposes. When we see any indication that the Government are making money out of our personal information, especially when it is being sold to, say, insurance companies—Care.data comes to mind—people seem to get extremely upset about data being used within the public sector generally.

Q124 Chair: I will come on to Care.data. Professor Robertson, the UKCRC said that the Care.data initiative was "out of step with public opinion." Could you explain that?

Professor Robertson: The reason for choosing that example was that it is a good example of what I was mentioning before—how brittle these systems can be. We had no particular reason to believe that there was any notion of malice or impropriety in that particular initiative. As far as I am aware, the people running it were doing the best job they could possibly do, but it got slightly out of step. By out of step, I mean that the people whose data it was were not fully aware of what was going on. Because they were not fully aware in the right sort of way—in their culture or their understanding of what is going on in the scientific community; you can frame it how you like—it was badly rolled out. Because of that, there is a scientific consequence, which is that research that could have gone on quite legitimately and successfully, had that been done just a little bit differently, will probably be held up.

Q125 Chair: Are there particular examples you can draw our attention to in terms of that initiative?

Professor Robertson: Emma Carr is nodding and may be able to give better feedback than I can. The simplest thing is that people simply did not know that this was going on until it leaked out in the press. In Scotland we have not had this kind of impact because we did not take that step, but when we talk to people about this sort of thing quite often their view is, "Oh, aren't researchers using my data for something useful that will help to make people better? Wouldn't that be good?" That is very different from suddenly finding out that people are going to do something you were not told about that may be partly commercial and partly about harvesting data and is now going to become institutionalised. It just plays badly, regardless of what the motives are. That is the problem—it is the public perception of what you do. It makes the problem very difficult. It is the reason you have Caldicott guardians for medical data. You have to have people whose job it is to understand how to reconcile those dangers.

Q126 Chair: Isn’t that a rather difficult area? There are circumstances where exactly the same practice could be used for a commercial benefit. Let’s take the use of cameras as an example. There has just been an understandable hue and cry about terrible cases of abuse in a care setting. I heard one person advocating for the families of the victims say that there ought to be cameras in every care setting. That was for a perfectly good moral reason, but, equally, cameras can be used to understand customer behaviour and so on. You cannot generalise, can you?

Professor Robertson: That is the problem—you cannot generalise. This is moving very fast. Some form of data analysis is embedding itself pretty deeply in domains; health care is one, but you can choose pretty much any other domain you like. If you look at what industry predicts, a big proportion of the data available to it will be social data—data derived from the living population. If that is true, it will embed itself very deeply indeed. It is really hard to reconcile some of these issues without understanding domains very closely. Care.data could not have been predicted by a computer scientist working alone. I suspect it probably could not have been predicted by a sociologist or a social anthropologist—or even a medic—working alone. Somehow the mix has to work. You can see signs of these kinds of interdisciplinary groupings building up, but it is an avalanche. It is probably impossible for us to move so fast that we could make it risk free.

Q127 Chair: Do you want to add to that?

Emma Carr: Big Brother Watch did some polling at the beginning of the year. We found that 69% of people had not been informed of the right to opt out of Care.data. This is entirely the point. Whether we are talking about Government use of data or private sector use of data, it is all about consent. You cannot just presume informed consent because, as in the case of Care.data, a leaflet has been sent out. Especially when it comes to something like medical data, where if people do not have trust that their data will be kept secure it may have hugely detrimental effects both on their own health and on the public sector as a whole, you have to ensure that there is a higher standard than for just tweeting or Facebooking about your day. If you stop doing that as a result of not trusting Facebook or Twitter, it does not really have the same consequences.

Q128 Stephen Metcalfe: Talking about trust, where has this lack of trusting the Government with data of any sort come from? Where has it grown up from? It strikes me that inherently the Government are trusted less than almost anyone else.

Professor van Zoonen: There are a couple of reasons you could think of. One is the particular context of the UK and the reason why data are collected here, which is very much set in a security framework. That has historical reasons and has to do with the alliance with the United States. If you compare the purposes of data collection for the UK Government with the way in which that discourse is framed in continental Europe, for instance, you find that in continental Europe the whole discourse about data collection from citizens is about providing service to citizens—making sure that a Government can offer better services to its citizens. The discourse in the UK is a safety discourse. That is a helpful discourse for citizens in the immediate aftermath of a crisis, when people buy into that and think it is really necessary, but after half a year people think, “Right, that is not that good an idea.” The transaction is a little bit abstract. You give your data and get security, but the longer that takes, the more abstract it becomes. In the Indian situation, for instance, and in Europe, you give your data but you get a better service. That is a really different way of presenting the policy.

Q129 Stephen Metcalfe: So in the UK the citizen does not see the link between sharing data and better services.

Professor van Zoonen: No, not yet, because the whole discussion here is always about security. Look at the surveillance cameras, the Patriot Act and its aftermath and data protection—it is all in the context of protecting your citizens, not of offering better services. That is there, but not as the first thing.

Dr d'Aquin: I agree with that view. There are quite a number of different elements. I am not exactly sure we can say that citizens do not trust the Government, but what private companies—especially the big private companies—give back in exchange is more direct, sensible and visible. There is the notion that if you want to collect information from users in a commercial system, or even in an academic system, you have to provide incentives—you have to have something in return. From a commercial perspective it is much easier to make that something in return visible to the user.

There is another aspect. Our research experience suggests that a lot of the lack of trust comes from fear of interpretation. If data are collected by Government, users need to have an understanding of what is going to be done with them. While it is still very obscure in most cases why personal data are being collected all round and what your social media platform will do with them, the obscurity is much bigger when it comes to Government. As has been mentioned, there is no clear indication of what the exact use and benefit globally of this particular data will be.

That is where the fear of interpretation comes from. In some of our research, we have seen that, if you put users in front of the personal data their employers collect, reactions are often very pragmatic—"Of course," "That is fine," "That is normal," "That is part of operational services"—but on certain very specific aspects the user will want to have a right to respond to their personal data. For example, a user spending a lot of time on the vacancy pages of their own organisation will feel that they need to provide context to that particular information so that it can be interpreted properly. The lack of trust also comes from this—the inability of the user or the citizen to provide the context for interpreting the data they make available. It is a lack of control, really.

Q130 Chair: In these two answers there are some interesting contrasts, aren't there? On the one hand, I suspect that just about everyone in this room has a piece of plastic provided by some banking organisation or another where we have given a lot of data. The trust vehicle there is that, if they go wrong, they commit to making good their mistake financially, broadly speaking. On the other hand, I suspect we all have a Government piece of paper—an Oyster card—in our pocket. The payback there is the convenience of not having those wretched queues at the ticket office, isn't it? So there is a point where the public give in to some of their worries about the role of Government in exchange for a service—the point you made, Professor van Zoonen.

Professor van Zoonen: Going back to the Care.data discussion, if you look at the way that is brought to the public—at least, in the website information—it is all about helping the NHS improve its services. That is interesting, because there is no direct link to how that helps the individual patient; it is all about helping the institution. The other side of that—what you get for the transaction—is underplayed, so to speak. Again, that is an example of an institutional interest—a Government interest—taking precedence over the individual citizen or patient. It is a matter of framing your data-sharing policy, in a way.

Q131 Stephen Metcalfe: Do you want to comment on the original question of where this lack of trust comes from?

Professor Robertson: This is a personal view, but I do not believe that Government is necessarily trusted less in all respects than some companies. I do not think there is a tremendous amount of trust in some of the large companies dealing with data—it is just the way that it is articulated. You have to choose where your pedigree examples are from. In Edinburgh, there is probably not a lot of trust in the local authority building tram systems because it did not go tremendously well for us, even though we now have a tram system—which is good, sort of. It is that kind of thing.

There is still a monolithic feel to Government IT, which does not square well with the picture that was painted for you earlier of very lightweight, streamlined services, relevant to the user and so on; it just does not feel that way. I feel for people in Government, because that is quite a difficult one to shift. You do not have the luxury of being able to say, “Oh well. We’ll start with a clean sheet of paper.” There is a whole lot of baggage to do with public service and Government that you would not have if you were a start-up company, so it is very difficult. On the other hand, that is the way the world appears to have gone. From apps and all the way through, people are much more accustomed to being able to chew things up in small bits and understand vicariously the notion of a service that is delivered over the internet. That was not true five years ago. I should think it is tough for Government; that is part of the reason.

Emma Carr: There has been far more in the media about data loss in the public sector than about data loss in the private sector. Part of that is Big Brother Watch’s fault, I am afraid, for putting that into the public domain. Quite often you hear about USB sticks with child protection information being left on buses or mobile phones going missing and the relevant security not being there to protect all of that information.

Q132 Stephen Metcalfe: Because that never happens in the private sector.

Emma Carr: Apart from one or two high-profile cyber-security attacks involving people’s credit card details or whatever, it is on a much smaller scale. Obviously it happens, but—

Q133 Chair: How do you know it is on a lower scale? You can’t possibly know that.

Emma Carr: How do I know what is on a smaller scale?

Q134 Chair: That loss of data in the private sector is on a lower scale.

Emma Carr: I think it is probably on less of a large scale than in the public sector.

Q135 Chair: How do you know that?

Emma Carr: That is entirely the point. We find out because it is usually to do with credit card details. People generally have to find out about that because there is a financial aspect. When it is in the public sector, we generally find out because we have put in a freedom of information request, which we cannot do with the private sector, unfortunately. Again, that is more to do with having a much higher standard for information relating to people's security or health, which people care about much more. Coming back to Care.data, when you have the public sector asking for larger datasets and wanting to do more with them, it raises the question whether the safeguards are there to ensure that those data will be kept to a higher standard than people would necessarily expect from the private sector.

Q136 Stephen Metcalfe: So it is not necessarily the individual coming to the conclusion that they should not trust the Government when the data are passed over, but someone telling them they should be outraged that the Government want to use those data to provide a service. Do you think that is why the Government are now pulling away from sharing more data? Are they fearful of media and public reaction because people are being told through the television, the radio and the newspapers, "You need to be outraged"?

Emma Carr: I disagree that the Government are pulling away. I am on several advisory boards that are helping the Cabinet Office and NHS England to try to go forward with data-sharing practices. We can all agree that it would be beneficial to the public sector to have that extra level of data and to be able to analyse it to help public services—it is just about going about that in the right way. Care.data highlighted exactly the wrong way to go about it—by not telling people what you are going to do with the data, who will have access to them and how they will be shared.

Q137 Stephen Metcalfe: Do you not think that the damage has now been done—that now no one will ever want to share those data?

Emma Carr: Exactly. That is why I am on the Care.data advisory board. We are working extremely hard to ensure that those mistakes are put right and the public can have confidence, when it does go ahead, that they have been taken away. We need to ensure that at all levels of data sharing, whether it is in the NHS or at local authority level, people can see from A to B exactly what is going to happen with their data. Unfortunately, there have been too many occurrences where that just has not been demonstrated.

Q138 Stephen Metcalfe: On that one area of Care.data, were you proposing that people should positively sign up to allow their data to be used, rather than negatively pull out?

Emma Carr: We would always advocate an opt-in rather than an opt-out as best practice, but we understand that that is not always practicable. However, we suggest that you do not tell people by means of a leaflet that you intend significantly to introduce different means of sharing medical data.

Q139 Stephen Metcalfe: So somewhere in the middle there is a way to be found. Do you think that the Government Departments that are potentially in possession of large quantities of data and might have opportunities to share them are not skilled enough to know how to use them and how to prepare the public for sharing them?

Emma Carr: We have been involved in workshops in the Cabinet Office about using big data to solve fraud, error and debt. Interestingly, when we ask what data they want to have and currently cannot have access to and what they want to do with it, often they do not have the answer to that question and take quite a long time to respond. Often in our experience at Big Brother Watch we find that they know big data is available—it is talked about and is seen as something of a new, modern silver bullet to aid public services—but there is something of a question mark against what they then want to do with it, perhaps because they do not have the necessary knowledge inside that Government Department.

Q140 Stephen Metcalfe: Going back to the public's view on how all of these data are stored or used, is the concern about their collection and storage the fact that they might end up on a memory stick on a bus or a stolen laptop, or is it not about where they are stored but how they are then used? Where does the main concern lie? Has any research been done into that?

Professor van Zoonen: It is also about how and where the data are stored. In general, if there is a central storage system, there is much greater suspicion about what is going on than if storage is decentralised. In Estonia, for instance, they have a national identity card that offers a lot of different commercial and Government services. The trick there is that all of the data that you hand in to those services are kept on different databases. They do not go into a central database. That makes quite a big difference to how much people trust such a system. The other thing that makes a difference is whether people know that their data are being collected or whether they are being remotely collected, either by remote biometrics or by commercial monitoring on the internet. Having a certain level of awareness and a decentralised storage system really makes a difference.

Q141 Stephen Metcalfe: So identity assurance providers who collect data and assure you that it will then be protected are not the answer, because that would be one central database.

Professor van Zoonen: There is a level of choice already. Lessons were learned, so people can choose to whom they want to give their data. That is an important difference already. They are aware that they are giving data and that presumably there is a purpose for giving those data as well that is not breached.

Dr d'Aquin: It is interesting that you connect the two. The identity assurance framework explicitly states distribution of information as a core principle, for the reason that it helps to establish trust by ensuring that no one private organisation will have total control over the entire system. Care.data does exactly the contrary. As I understand it, the current proposal gives the entire provision of software services to support it to one private company. That is one private company that people may or may not trust, but the understanding of how that private company might gain benefit from having such total control over their data is certainly not helping with the trust issue.

Going back to your original question, it is true that for a lot of the researchers and journalists who are looking at this aspect the core computing infrastructure—where the data are stored, how they are processed and how secure they are—is a very important issue. Based not on concrete research but simply on an anecdotal understanding of research in other areas, I would say that these technical aspects are far from the concerns of the majority of users. The one question is, “What are they going to do with it, and how can I understand what is going to be done with it in such a way that I can be reassured that the interpretation, understanding and use of the data will not go against my interest?”

Q142 Stephen Metcalfe: How would you ever arrive at being able to reassure someone that the data would be used in a way that they could even begin to understand? We have heard how complex some of the links are between different pieces of data. How can you persuade someone that what is going to be done will be done ethically, sensibly and safely on their behalf?

Dr d’Aquin: It is a good idea to start by saying what is going to be done.

Q143 Stephen Metcalfe: But there are people out there who will instantly say, “Don’t believe them. Don’t believe them.”

Dr d’Aquin: There are mechanisms that do not necessarily achieve that amount but allow a certain feedback mechanism to be included in the whole system. They include some of the elements that have been talked about in the mydata program, which give access not through a complex procedure of requesting it under data protection but by the simple click of a button. A single authentication into one system gives access to the full set of data somebody else might have access to, showing what kind of analytics can be achieved on those data.

In our own research we have done that, if only on a small scale, for the relationship between employees and employer. Our research has shown that when employees are faced with the data that their employer collects about them, the majority of them react by saying, “That is reassuring. I knew data were collected and that some things can be done. Seeing them makes me confident that I can entrust my employer with this type of data.” That is a very simple mechanism that is reasonably obvious. Giving the data back to the citizen—making them accessible to the citizen and attaching to them information about organisations and individuals who might have access to every part of them—seems an obvious mechanism, in addition to distribution, of course, to try to improve trust in these sorts of systems.

Professor Robertson: The issue is really, “What is the evidence base?” If somebody says, “Why should I trust this?” you should have some evidence. The important thing I want to add is that it is not completely a black art. There are things that you can apply that are reasonably well understood. You can have various mechanisms that are just to do with the data artefact that protect it. You can have various protocols. There are all sorts of ways you can try to make sure that that is secure in the narrow, technical sense. There are things to do with the process by which you move data around. You can have processes that are inspectable, that you could accredit and so on. Those are quite often used in health care when you are interested in doing that sort of thing.

The third point is the authority argument, which is to do with professionalism. Just because you are using social data, it does not mean that everyone has to use it. You can have communities of people who have a very high level of trust, which is what an awful lot of this sort of thing runs on at the moment. All of these things are pretty well known. In my personal opinion, they tend to screw up where they are done without respecting the domain. People do them thinking that one size fits all and they can do the same thing as they move between different areas. That is where you get all sorts of sociological and cultural problems.

Stephen Metcalfe: Thank you.

Q144 Stephen Mosley: While the focus of the last question was very much on trust, it started to move on to things like security and safety of the data as well. Is it possible for the public ever to be sure that their data are secure?

Professor Robertson: No. I suppose any of us could probably expand on that a little, but let me be the first to do so. I suppose the more detailed answer is "not on any scale that we would care about." Pretty much as soon as the data start to contain significant content, it is very difficult to guard against that content being able to be used to breach various aspects of your privacy and security. That is it; that is the basic problem we are all fighting against.

Q145 Stephen Mosley: I will assume the answer is pretty much the same for all of you, but could I expand it a bit? Have the Government done enough to make sure that individuals' data are secure?

Professor van Zoonen: What we find is that people are very concerned about things like identity fraud and privacy breaches. They do not have a real clue about how that works, but it is a big concern. In addition, saying that their data will never be secure is a recipe for quite a lot of distrust. The issue is not making the data more secure but making sure that, once your data have been stolen or mistakes have been made with them, there are mechanisms to correct that. You bring your car to the garage for a yearly check-up to see whether it is still all right. You need to have something like that for your personal data as well, so that they are checked on a yearly basis to see whether they are still all right everywhere and have not been stolen. We hardly ever talk about that kind of mechanism and whether it is a way to repair identity loss, identity fraud and privacy breaches.

Q146 Chair: That would have to be predicated on the citizen having the right of access to data.

Professor van Zoonen: Exactly.

Q147 Chair: And on having some sort of ombudsman role that presumes the citizen is right, rather than waits for—

Professor van Zoonen: Yes, it would be the other way round—trusting the citizen to come up with the right personal data, instead of distrusting—

Chair: I have been arguing that for years.

Professor van Zoonen: Yes, but an ombudsman is much more on a collective level. There should be individual mechanisms as well. In that sense, we know very little about what kind of crime identity theft is, for instance—who are the most vulnerable people, who are the most likely perpetrators and so on. That would be part of bringing back confidence in Government doing stuff about this, because it is one thing that is very high on the agenda of the people we spoke to, at least.

Dr d'Aquin: There is an element that relates to transparent traceability—putting in mechanisms that make citizens able almost to audit by themselves how their data are being handled. It is not about having unnecessary technical complexity, but simply about being able to demonstrate the level of security that is being given to the data.

You must also respond directly to any possible threat or issues. We had a discussion before about whether or not Government were leaking more data than private companies. I disagree respectfully with the answer that was given. I think private companies leak personal data much more. Almost every month one of the very large companies is shown to have been hacked and to have had data stolen from its system. The difference is that those companies are starting to get better at putting on a direct response by locking out all the accounts that may have been affected and getting in contact directly with the users who may have been affected.

It requires very complex and demanding mechanisms to be able to trace back any security issue and how it may have affected users directly. That is the sort of mechanism that could make trust in the security of Government systems with respect to personal data more effective. You can never achieve complete trust in the security of a system, but you can improve trust in the people who are trying to make it secure and the traceability and auditability of the system.

Professor van Zoonen: It is interesting that in the last two years many of the big companies and consultative groups have issued reports about gaining online trust from their customers. Basically, they have the same discussion that we are now having, about how to assure trust and the secure handling of personal data. For them, it has become one of the key issues in maintaining a good business case. They recognise that if that trust is not gained from their customers they will lose business. It is a very simple economic argument.

Emma Carr: Going back to your original question, not only can you never be sure that anyone will keep your data safe but it would be completely dangerous to try and allude to that. Sometimes both the public and the private sector can overpromise about how safely data can be kept, whether those are being stored in the UK or outside.

Going back to your question about the safeguards, the regularity framework around the Data Protection Act is pretty good, certainly for data in this country, despite its being fairly old. The one thing I would say about it is that there could be far harsher penalties available for people who breach section 55 of the Data Protection Act by knowingly disclosing information. That is something that could very easily be rectified. You could have custodial sentences available. Under section 77 of the Criminal Justice and Immigration Act, the Home Secretary has the ability to amend the Data Protection Act to include those. If people saw harsher penalties were available, rather than just very small fines—sometimes we see fines of less than £1 per piece of information that has been lost—that would give them more faith that the safeguards were there.

We also should not rely just on the Information Commissioner to enforce the Data Protection Act. Too often, they have had to prosecute under the Computer Misuse Act or for the offence of misconduct in public office, because harsh penalties are available and it is easier to prosecute. We could allow civil cases to be put if I know my data have been lost, whether it is by the Government or a private sector company, as is possible in the US.

Q148 Stephen Mosley: You have mentioned the Information Commissioner. He has noted that individuals lose data rights once the data have become anonymised. Do you think that should be reconsidered? It depends on the situation, doesn't it, but is it getting easier to un-anonymise data?

Professor Robertson: In narrow technical terms—to give you that kind of answer—this is a very difficult one. It is not as if data just sit there, you then do something with them, they go through and you can easily spot the routes by which they go through. Typically, very early on in the analysis of data they will be transformed into something else, with the same identifiers, different identifiers or a different structure. It is a bit like taking two numbers, multiplying them together and then trying to guess what the original two numbers were—it is not always easy. It is quite difficult to disentangle that.

You can go the other way, of course, and take data where it would not have been obvious that somebody was identified, or maybe not obvious even that the structured data were there, because they were obtained from mining Twitter feeds or from unstructured text. All of those algorithms are getting really quite good now. All of that is generating data that in some sense maybe you never put in—or certainly that you never put in intentionally. That is being pushed in all sorts of different directions, because a lot of those data were already on the open web. That makes it quite a complicated topology. That is just technical. As far as I am aware, I never said anything that remotely verged on the sociological—it only gets worse. My point is that it would be very difficult to think of some gold standard or single route for doing it. Basically, it is about continuous vigilance and continuous work.

Emma Carr: I would certainly like to see the ICO address further anonymisation, the potential of re-identification and pseudonymised and de-identified information. From my awareness, as time goes on and more datasets are created, it becomes easier to re-identify information that has been de-identified. That is something that I would like to see the ICO address before we come up against any potential problems in the future.

Professor Robertson: I also do not want to give the impression that it is completely hopeless. One thing that you want is an appropriate level of de-identification. You want to reduce the probability of identification to some reasonable level, given the benefits involved. It is a bit like safety cases for aircraft. Occasionally an aircraft will indeed fall out of the sky, but not so often that typically it feels like we should ground every aircraft.

Q149 Chair: We are practical as well. We have suggested in the context of some health data, where sharing for research is valuable, that the data could be held in a safe repository and released under controlled circumstances, case by case. Those sorts of mechanisms are feasible, aren't they?

Professor Robertson: Yes—those sorts of proportionate mechanisms are. The property of these domains is that it always looks really bad in the general case, but it is possible to make progress just by shifting the parameters a little bit, if culture allows. That is the crucial point.

Emma Carr: All too often getting the terminology right here is very important. When I have entered meetings, people have been talking about anonymised data when they mean pseudonymised, and pseudonymised data when they mean de-identified or statistical. Far too often they say "anonymised" when that is not what they mean, when we are talking about data sharing in practice. There needs to be a greater awareness within Government Departments, if they want to use data, of what the terminology actually is, so that when they try to inform people of what is going to happen they are informing them correctly.

Dr d'Aquin: That is exactly the point I wanted to make, but formulated in a slightly different way. I agree with the Information Commissioner that truly anonymised data should not be subject to data protection, but I say that knowing that anonymisation can mean an awful lot of different things. There are other examples; the ones that have been mentioned are completely right. I have just read one about a city in the US releasing anonymised data about the taxi trips done during the year in that city. Those anonymised data were entirely and fully de-anonymised within two hours. It was not because the technology is better or because people are getting better at re-identifying data, but simply because the anonymisation mechanism that was used was provenly and very ignorantly weak. It was a big technical mistake.

I like the idea of anonymised data; it can be extremely useful, if it is done properly. However, if you are talking about data that are personal or useful, I do not agree. What is needed is not only a proper definition but a proper framework—standardisation of what anonymisation means and what an anonymisation level is. There are mechanisms that exist in which single datasets can be assessed with respect to anonymity. Any discourse or release of anonymised data should come with a proper assessment of the level of anonymity and should be considered as having to follow data protection or not on that basis. That is a very important point. Basically, anonymisation can mean an awful lot of different things.

Q150 Graham Stringer: Not everybody uses social media, and there are parts of the country that are not covered by these sorts of electronic communications. Is it sensible for the Government to use information from social media as a basis for forming policy, because those datasets will always be inadequate?

Professor Robertson: While they are thinking, I will give you my academic's answer. There is a different weight. There are some demographics, for example, that use less social media, but there will be pretty strong drivers for some of those to do so soon. For example, when you look at the size of the problem that is emerging with the elderly, the frail and the infirm, it is very hard to imagine some notion of technical social networking not being used. That will have to happen, because otherwise it is not going to work. If that does happen, you will see it shift. It would be brilliant to be prepared for that, rather than to try to exclude people from that particular ethic. We should be approaching it the other way round and trying to push it in, with the aim of making the other problems to do with demographic shift rather less difficult. That would be a more positive way of putting it. I realise that, being an academic, I have not really answered your question.

Graham Stringer: It was interesting.

Dr d'Aquin: The most naive answer would be, "No, it is unreasonable." At least, it is unreasonable to do it in a completely naive way. It is obvious that any data collection mechanism, including social media analysis, includes a bias, but the proper use of social science methodology includes assessing the bias and the effect it may have, and including any other complementary mechanism that can help to reduce that bias. While David's answer might be to try to reduce it by making sure that people join in, there is also an element of simply assessing what is the effect of the bias introduced by the data collection.

Professor van Zoonen: In addition to what has been said about vulnerable groups not being on social media, it is important to realise that there is an increasing movement of people to getting off the grid altogether. That has nothing to do with vulnerability and everything to do with people not wanting to be traceable by energy companies, Governments or whatever and wanting to live off the grid. It is a serious social movement that would no longer be captured by social media analytics.

Q151 Graham Stringer: That is right. Where do you think the data on social media could give us most benefit in extra information and as an extra base for policy making?

Professor van Zoonen: There is a fantastic example in the health context, although it comes from the Netherlands, of failure of a policy because of social media. It relates to a disease for which girls have to get an inoculation when they are 13. I cannot remember what it is called, but when they are 13 years old girls need to get a vaccine against a certain type of cancer. In the Netherlands that policy failed because of a social media campaign by parents, who were very concerned about what was going to happen to their children. The year after that, the Dutch Government started their own social media campaign. Whenever there was a story about the vaccine being dangerous, they would put in their expert and explain what was going on. Since a lot of social media analytics is about marketing, advertising and campaigning, it is there that the Government could probably gain most—not by using the data but by being on there to explain their policy.

Q152 Graham Stringer: Is social media best—or only useful—when it is used in relation to information from other sources, such as CCTV?

Emma Carr: It depends on what you want to achieve. One suggestion is that social media data could potentially help the police with information that is usually gathered by the census or something like that. However, as has been said, the accuracy of social media data has been questioned in academic studies. I know that the ONS is looking at Twitter to see what it can ascertain from that that it could not necessarily ascertain from the census and the questions it has to ask every 10 years—or in addition to that. It depends on what you want to do with social media. If you want to inform people via social media, that is one thing; if you want to analyse people via social media, the safeguards around that are obviously different.

Professor Robertson: It is starting to develop enough outside the public sector that there may be something to be learned from looking at applications that were not developed there. It would take too long to go through all of the examples, but there are canonical examples to do with emergency response, where you very quickly build up maps and so on from just the population, because there is nothing else. There are commercial examples to do with navigating roads and so on. You want to get really fast, up-to-the-minute information about where there are blockages on roads, so you do it through the social network rather than by trying to model the road network. In the commercial world and that kind of funny world of people just doing stuff, there are an awful lot of those kinds of examples. The bigger question for this kind of audience is, how many of those would translate and survive were they to be used in Government? Government is a big lever, because it could potentially use data that it has acquired to jump-start some of these things or to provide a bit of an incentive for people to be involved. However, as you may be aware, there is rather less expertise on how you do that.

Dr d'Aquin: I cannot talk generally, but from conversations we have with local government, with the police about crime and with people working on disaster management and things of that sort, I can say that there is an increasing awareness that social media data have two extremely good qualities: first, there are a lot of them; and, secondly, they are fast. In many cases—I take the example of crime reporting—that is a double-edged sword. It comes with needing the ability to process the data in very large amounts and to make sense of them, to reduce noise and to enable early signals of something to look at in more detail to emerge. In many cases it cannot replace longer-term, richer and more sophisticated ways of data collection.

What I see emerging as a pattern is the use of social media to detect in a very rapid way early signals of a phenomenon that then has to be validated through the use of other data. In crime reporting, especially, the example you give of CCTV is an absolutely obvious one. Looking at Twitter to detect that an issue may be happening somewhere at that moment means that someone can go and watch a real stream on a CCTV camera to validate that something is indeed happening and then take appropriate action. In many cases false reporting, noisy data or simply misinterpretation of what in the end is a very large amount of unstructured data will show that such validation mechanisms are absolutely necessary.

Graham Stringer: Thank you.

Chair: Stephen has a couple more quick questions.

Q153 Stephen Metcalfe: I want to go back to the legislative framework that is currently in place to control the way our data are used. The EU is proposing some changes to the Data Protection Act. There seems to be a mixed reaction to that. TechUK said it would have a "negative impact" and Big Brother Watch said it would improve individuals' control over the amount of data. Briefly, what is the purpose of data protection legislatively? What is your view on these changes? Is there a balance to be struck somewhere that will mitigate against the negativity while allowing individuals some control over what is done with their data?

Emma Carr: One of the things that Big Brother Watch quite liked about the new EU regulation was that it shifts the burden of evidence away from the individual to the researcher. No longer will the individual have to explain why those data should be kept by the researcher—the researcher will have to justify why they should be kept. That is one of the safeguards. It gives individuals more control over their information, which is always a good thing.

EU data protection is incredibly important because of the online world we now live in. I would like to see some kind of global principle for how our data are used and shared, because the internet has completely changed the way in which those data transfer across borders. This new regulation is also about additional safeguards for when data are moved outside the EU. I believe there is a process where the national data regulator has to get involved and to sign up to some safeguards saying, "If these data enter our country, we will be responsible for them." That higher threshold is important and is reflected in the individual nations, which tend to adopt the same principles. Anything that can be done to give individuals more control and to have a higher level of safeguards is a good thing.

Professor van Zoonen: In this respect it is interesting that, when we asked our respondents what they feared most in the future about data sharing, their biggest fear was that they would have to pay for privacy and would always lose out to big business interests. If you contrast the EU policy and citizen protection with what techUK is saying about the economy suffering if we do not allow this data sharing, that is exactly where this public fear is located—that the citizen interest will lose out to big business interests. It is necessary to take on board that kind of sensitivity and what people are thinking about these issues.

Dr d'Aquin: I am not profoundly familiar or at least as familiar as my colleagues may be with the changes that are proposed, but one of the key arguments is that, while, in principle, giving more control to users, these changes will introduce a bigger need for companies to invest in data infrastructure—in mechanisms such as responding to requests for the right to be forgotten. These mechanisms of this investment might hamper the economic growth of Europe in this particular area and might simply introduce additional obstacles to the business and economic use of social media analysis. That is a valid concern, from a purely economic perspective. In response to that, the important aspect is to provide support for businesses to make this investment and, especially, to factorise, understand and show good practices and how to make this investment adequately so that the economic benefit of social media analysis is not reduced by the need to comply with the regulation.

Q154 Chair: There are other examples beyond the US, but my final question relates specifically to the US Patriot Act. You will be glad to know that the notes we have on you on these iPads are not out in the iCloud. The House took that decision consciously, because it is not held within the framework of legislation the UK has control over. Clearly, legislation like that is going to impact on ordinary citizens. Do you think the Government should be doing more to help people understand how their data are held in other countries and what rights they have in respect of those?

Professor Robertson: Yes.

Emma Carr: Yes.

Q155 Chair: I do not think I have caused any controversy there—I am just curious. I might have if the American ambassador were here. Looking at it from the British Government's perspective, should we adopt a Patriot Act kind of approach to gain access to social media data?

Professor Robertson: The things that are important are the evidence base and transparency. That is the difficulty with the Patriot Act. You can understand how it was rolled in at a very emotional time, why it was felt to be necessary and so on—that would be a very long conversation—but the way that it has played out has been very murky. The effects of that are pretty wide ranging. It is not just the Patriot Act but everything that went along with it. Even if you look at the research community in pure science, you will find that quite a lot of people working in security in that community feel personally undermined in their work by some of the things that are believed to have been going on. A lot of these things are beliefs—you do not know exactly what happened—but it has certainly demolished a huge amount of confidence. If you had to think of how to demolish confidence with public bodies in that area, that would be a good way to do it. I could not have made it up.

On the other hand, we have just talked about the EU directive and so on. In my personal opinion, that is another example of something that is quite broad-brush, so it is very difficult—in the other direction, but still a bit unwieldy. Both of those things feel slightly uncomfortable to me. Somewhere there must be a pragmatic middle route, because we have to do something. As you say, part of it has to involve simply explaining, "These are the things that we believe we do, to the best of our ability and our ability to tell you. Here's how they are laid out." That would be a tremendously positive thing to do.

Emma Carr: If the question is, "Do we need a Patriot Act to get access to communications?" the response is no. We already have things in place in this country to allow people to have access to them, such as the Regulation of Investigatory Powers Act, the Telecommunications Act and the mutual legal assistance treaty, for access to communications from other countries. We have to be very careful when we think about the Patriot Act and its repercussions for the United Kingdom. As has been said, it is not just about communications. We must remember that in this country we have a different legal framework from the US.

Dr d'Aquin: One thing that was quite misguided about the US Patriot Act was that it was realised in a slightly naive way, possibly because it was rolled out very quickly, at a time when it was felt to be necessary. Under the Act, the only approach to data collection to fight terrorism is the bulk collection of all information about every citizen and filtering that out a posteriori—after or when it has been collected—which sounds really naive.

The argument about the need to collect data for security is not one that I would feel confident to have myself, but it requires complex and sophisticated mechanisms to ensure that data are not misguidedly and naively collected, in a way that might have bad repercussions, that data are collected at the time when they are needed and that the conditions under which data are collected about specific individuals are made extremely clear to the public.

Chair: This is a debate that will go on. I was going to come back to Professor van Zoonen to point out that we pay for privacy already, in the sense that if you do not want a Clubcard you do not get the discounts and so on. There are plenty of examples like that. Thank you very much for a very good evidence session.