CHATHAM HOUSE

Chatham House, 10 St James's Square, London SW1Y 4LE
T: +44 (0)20 7957 5700 E: contact@chathamhouse.org
F: +44 (0)20 7957 5710 www.chathamhouse.org
Charity Registration Number: 208223

Transcript Q&A

Tackling Global Problems with Big Data

Viktor Mayer-Schönberger

Professor of Internet Governance and Regulation, Oxford Internet Institute, University of Oxford

Kenneth Cukier

Data Editor, *The Economist*

Chair: Professor Angela Sasse

Head of Information Security Research, University College London

25 March 2013

Question 1:

I think you hinted at this in your presentation, but I wonder if you could both say a little more about the implications for educational systems of what you have discussed.

Viktor Mayer-Schönberger:

We feel quite strongly that education is ripe for a big data revolution, in the sense that in many ways in the educational sector today, we make decisions that are based on small data. Take just two examples that come to mind here. One is with textbooks now being digitized in so many instances and accessed through tablets or the Amazon Kindle, the data of how fast we read something, how often we re-read a page and so forth, actually is captured – it is datafied. Because the Kindle or the iPad knows how often you looked at the same page or how long you spent on a page or what you scribbled in the margins and so forth. The problem is that this data currently stays in the tablet or stays with Amazon or stays with Apple. If we could feed it back to the authors, for example, the authors could actually make sense of it and then hopefully improve some of their works.

But in addition, school districts as well as school authorities could look at school textbooks and have much better data on the quality of textbooks. Right now many of these decisions, particularly in the United States where school boards are on the regional and local level, are made on hunches and very subjective biases. With big data we might be able to capture the data. That might not necessarily improve decision-making but it will certainly improve the empirical basis on which decisions could be made. So we believe that there is a real revolution. Ken is actually going to say more.

Kenneth Cukier:

We have a wonderful anecdote in the book about this very thing, where there was a professor of computer science in the United States, at Stanford, that in his class he was getting the homework in, and he noticed that a huge number of people in the class not only got an answer wrong, but they got it wrong with the exact same answer. So suddenly he thought: this is very interesting, and he realized what was happening was that they were inverting an algebraic equation where people thought that the sequence didn't matter but actually the sequence did. So they could send the students an alert to say: check your math. By watching how students are performing, it's providing feedback not only to the students to improve how they learn, but feedback to the professor

on how he or she teaches. And so I think you're going to see a real revolution as we datafy more aspects of education.

Question 2:

I wonder if you could move onto privacy framework which you mentioned, because the algorithmists would have to be a very special set of people. They won't just be any algorithmists, and that's one of our problems with managing big data: it's how 'any' can preface anything you do with it. The 'any' is really sort of an elusive animal at the moment in this great jungle. There's a Research Integrity Forum coming up next month and it's going to be looking at various aspects of research integrity, using academic bases mainly. So it's the previous questioner, to some extent. But where do you see this fitting into a bigger picture of how the overall governance of big data is going?

Viktor Mayer-Schönberger:

There are two aspects of this that I'd like to comment on. One is the aspect of privacy – I apologize for my American pronunciation, I should say privacy. With respect to informational privacy, we do suggest that we need to switch away from individual consent and notice at the level of collection to an accountability – and dare I say, liability – of data users at the use stage. It's incredibly important to do that. It's going to be tough for data users because it means that they become accountable. Many of them today are not, or are only sort of perfunctorily accountable. But by the same token they gain a lot – they gain the ability to reuse data, to reuse the treasure chest of information that they have amassed. So we believe that that is actually something of a balance that they would find appealing.

You also mentioned data integrity and that's really a very dear subject to me. One advantage of big data is that small data can be faked much more easily than big data. If you have 100 data points of a sample, you only need to fake 20 or 25 of them to make a dent in your analysis. 'Let's throw out these 15 outliers here and we'll be fine' – right? If you have 10 million data points, you actually need to throw out many more, or fake or change many more. In that sense, big data is a little – given where we are technologically today – a little more persistent. That might help on the data integrity side as well.

Angela Sasse:

Can I just ask a quick follow-up question to that? In your book you're quite optimistic that the data users will behave in the long run. You basically speak about the necessity of the right kind of regulation and oversight. But we have just over the past few weeks seen that several of the big data users were lobbying the European Commission to basically water down its proposed new privacy laws. Myself and dozens of professors across Europe were basically writing this open letter saying don't give in to these arguments that they're saying, that privacy is anti-business – it's just not right.

Kenneth Cukier:

Just because they want to have simpler privacy laws doesn't on its face seem wrong. It doesn't seem like the EU does a particularly good job of protecting privacy and it doesn't seem like their rules are always sensible. Wouldn't reforming the law make sense? Why should we think that there's something wrong with that, *prima facie*?

Angela Sasse:

I think what was wrong with it is that essentially they wanted a lot of the accountability removed from it. It wasn't just about wanting simpler laws, it was they definitely wanted to have a lot less accountability.

Viktor Mayer-Schönberger:

It seems to me that what is at stake here is a relatively straightforward – or what is happening here is a relatively straightforward battle, a lobbyist battle. Brussels is coming up with new rules. I happen to think that they are good, I happen to endorse all of them. I think that this is a smart move forward given the circumstances that we live in. But it is not embracing big data. Brussels stays within this old small data mindset of collection, of notice and consent, of agreement at the point of collection. That is not the kind of privacy framework we need for big data. But then, a big data privacy framework might not have been acceptable at the level of Brussels and the European Union at this point.

So it's a pragmatic kind of little step forward, but it will not help us in the big data age. We are not optimistic in the book that privacy can be protected in the big data age. Rather, we think very strongly that without clear, government-imposed safeguards, our privacy will be compromised. That is

why we push for very clear, very strict, government safeguards. But in return we also want to enable the ability of data users to do big data analysis.

Kenneth Cukier:

I'll also explain why I questioned the presumption – it's just sort of a natural journalistic inclination, but it's also born of despair. Many European news sites have lost money over the last 18 months. We are poorer because of Brussels. The reason why is that we've had to implement a policy that says we need to disclose to every user when there's a cookie. Cookies have been around on the internet for about 15 years, and American news sites don't have to do this. But what it means is that when search engines come to our site to index us so that people can find our content, the first thing that they tend to see is not, for example, Syria or a story about British politics, but actually that there's a cookie site. So we're not indexed, so we get actually less traffic. And this is dollars and cents.

The European news sites are receiving less traffic. I don't think there's a single person on planet Earth that feels safer in terms of their privacy being protected because they've been given notice that a website is collecting cookies. However, this is how public policy not only interferes with the natural commerce but how it undermines, in this small area, European industry.

Angela Sasse:

Actually, I can point you to a paper that says that this is what the ordinary citizen wants. But there are other questions.

Question 3:

I have a very specific question. I know you mentioned a bit about the government usage of data when you introduced the subject. My question actually refers to… you hear all the stories with regards to the Department of Homeland Security and counterterrorism and money laundering, and you hear these stories about government just being swarmed with big data and not being able to appropriately analyse it. My question is: is there a way to start instigating more sharing of policies with regards to the corporate and the government world? How can governments become more accountable with regard to big data usage and commodifying their data?

Kenneth Cukier:

So the idea is that governments are collecting more data than ever before and they're not able to process it, but what can we do as citizens to make sure that we have the societal safeguards that it's not going to be abused?

Question 3:

Or, just so that they are actually able to utilize the data appropriately.

Kenneth Cukier:

Okay. We're hamstrung in responding to that because there is what's on the surface and then what's private. We don't have a lens into how the National Security Agency is handling data. We do know, from lawsuits that have taken place and from what's been publicly disclosed over the last decade, that they're collecting a lot of data. By 'a lot' I mean most international telecommunications traffic, both email and voice, and all social media is being collected. We have good reason to believe that it's all being stored.

Now, how it is being used is a black box. We don't know.

Just by dint of collecting the information, is privacy being abused? I think we need to have a really serious debate about this, as a society. Anyone who has a strong view on it at the outset is speaking from an ideological basis and not from a thoughtful one, because the first step in is to say: is the dint of collecting it abusive? Well, maybe not. Maybe the abuse would be in the misuse of that information. Maybe if the misuse is so easy to grasp for that the proverbial madman given the keys to the kingdom – whether it's the United Kingdom or the United States – could do wanton damage, which is quite possible, we might want to create the building of the safeguards. But we also have to recognize we live in a very dangerous world, that there are well-meaning public officials who want to protect us, and that information is the tool by which they protect people.

So if that were to be the universe, we could run a thought experiment simply to say that we might allow a vast capture of information – might, a vast capture of information – in return for very toothsome, muscular safeguards – that if it was used abusively, those that misused it, their lives would be ruined and their careers would be ruined. In that it empowered government to crunch this information and learn from it to spot the next terrorist attack or the next whatever, the next bad thing that could happen, I don't want to say at the outset, to reach for the idea that this is a calamity and terrible – partially

because I'm not certain it is, but secondly because what we already know is it's being collected. So even if you were going to be the most pessimistic person, you would want to look for the solution that is going to be the most feasible, and that would be layering on the right safeguards.

What we do mention in the book is that it seems that national security, and in this case surveillance, has changed in the big data age. In the past if you had suspicion, you would sort of drill down on the single individual and try to create a dossier of that person. Today, when we are the sum of all of our connections through Facebook and our interaction with content and others, you want to create a network map and it needs to extend out further and further to identify people who need to go under suspicion. To do that, you can at least appreciate why security agencies would want to collect all the information, not to do a dragnet but because when someone does fall under suspicion they can immediately create a network map and try to identify the penumbra of who this person is.

What is actually happening? We don't really know. So maybe we need more scrutiny on that through institutions like Chatham House and others, to create a very serious societal discussion – because the tools are there for this to be used in this way.

Question 4:

I want to start by saying I think that big data is obviously really, really exciting, and the way you can use it is exciting. But I'm wondering if you could say a little bit more about spurious correlations, because if you have such a lot of data – you look at risk algorithms in 2007–08 and things they missed. I'd like to hear more about how we can avoid the spurious correlations.

Viktor Mayer-Schönberger:

You're absolutely right. There is the danger of spurious correlations. By the way, there is a danger of spurious correlations, one could argue, even greater in small data than in big data. Because you might not have just spurious correlations, you might have collection biases in the data sample that even if you were perfectly fine on the methodology of analysing it, you might have collected it the wrong way and then you have the wrong view of the world. So there's no question that we need to be quite thoughtful about what we believe and what we take out of the data or not.

There is also no question that a lot of the methodologies that we have developed – statistical, network analysis and so forth – methodologies that we have developed based on the small data age need to be adjusted to the big data age. There is a very robust debate in statistical and econometric journals about what are the equivalents in the big data age of some of the methods and tools of the small data age. Take DR^2 as a measurement of confidence that you have – what would be the R^2 equivalent in the big data age? The need to analyse or to go beyond linear regressions, to understand non-linear relationships and so forth.

So it's early days. There is really a push on so many fronts, in academia and elsewhere, to come up with these additional tools that might be necessary to provide us with deeper insights. But we see a lot of that coming online and on stream relatively quickly, and so we are reasonably optimistic that innovation will happen there. That doesn't undo the danger of spurious correlations, by any chance, or it doesn't undo the problem of using the wrong tool to apply to the data. But with more comprehensive collections of data respective to a particular phenomenon you want to study, some, for example, of the selection biases and the collection biases will go away.

Certainly when we compare it to causal relationships that we are trying to investigate, we are far better off than with some of the causal relationships we have. There is no mathematical way of expressing a causal relationship in the first place. You might be surprised but that's the truth. So we will go looking for causes but we have no mathematical, quantitative way to express that. The way by which we try to prove causality today is through very carefully designed, double-blind experiments, in the hope that the outer parameters can be kept equal. In the medical sciences we run drug trials in the United States and the Europe with a couple of dozen at the beginning and then a couple of hundred people. We think that we can keep everything, all of the boundary conditions, the same in order to be able to do an A/B test and provide some causal inference here. I would suggest that our causal understanding of the world today is far more prone to error and to mistake than big data correlations are prone to spurious correlations.

Question 5:

Mine was more of a sort of comment first of all, and then a question. It seems that what you're proposing is quite a huge paradigm change to the application of technology. During perhaps the 19th and 20th centuries we applied it more

to industry and I don't think we've ever thought about really applying it to human behaviour as such. So that's sort of a comment.

The question was: if you're breaking with the correlation between causality, that's another big paradigm shift for Western scientific thought. I found it quite interesting that the book is the number one bestseller in China. Perhaps they tend to focus more on networks and how everything sort of fits together. Are they perhaps in a better position to benefit from big data, from their sort of philosophical tradition, than we are?

Kenneth Cukier:

I just came back from Beijing this past weekend. To be honest with you, I think it's much more pedestrian. They see big data as a chance to increase and fuel economic growth. They see it as a potential for leapfrogging the West once again. And in sometimes more subtle terms – I'll go on the record here on this – sometimes more subtly, when technocrats talk to me in the government in China, they want to use big data and the big data argument to open up more government data to the society. So they see it as a way by which data could play a bigger role in society, and empirical decisions rather than autocratic fiat could play a bigger role in society. Some of the public sector technocrats think that that actually is a good thing. I think we should applaud that, I think it is a good thing. I think their hope is that big data might be a vehicle by which they can achieve some of that.

Question 6:

I'd like to say thanks for giving the caveats to your talk and talking about the limitations of big data. It just strikes me that all the examples you've given so far have been sort of in support of the possibilities. I'd like to hear your favourite big data disasters. I can get the ball rolling – my example would be the Google Flu Trends that you referred to in the beginning, which though they gave a very nice backwards correlation, led to a very erroneous prediction.

Kenneth Cukier:

By erroneous prediction, you're referring to last December's – the fact that they overcounted? There's a sort of flippant answer to that – well, let me explain what it was. Last December, Google Flu Trends showed that there was a huge spike in flu cases, but it turns out that the CDC's reported data

was much more modest. They saw a huge spike, but it was nowhere near the spike that Google Flu Trends search queries had identified.

So there's a few flippant answers to that. The first one is: who's to say that actually the CDC is right and Google is wrong? You could imagine a case whereby the situation is always changing and so therefore at the time of a great recession, when people are a lot more wary of taking the day off to go to a clinic from their jobs, because they feel like they need to show up at work, that they power through even though they're running a temperature. So in fact the CDC's reported data is undercounting and Google Flu Trends is actually true. That's one possibility. There are other simple answers to that particular case as well.

In terms of great big data blow-ups, it's true we haven't looked at many yet, because we're really at the outset of big data. So usually they're fizzles, they're just not success stories. But if you want to point to one that is enormous and tragic, you just need to look at the global economic crisis that was set off not because, if you will, pure and simple big data, but it was about investment banks having computer models that didn't reflect reality, that were crunching lots of data trying to manage risk and didn't do it because they were throwing out – because the data itself was faulty, because the methods were faulty. They didn't actually listen to it when it did speak to them.

Question 7:

You've talked a lot about the uses of big data. In your answer to the previous question you alluded to the accessibility, when you're talking about China. But it seems to me like at the moment a lot of the data is sort of concentrated in the hands of, as you mentioned, Google, Facebook, Twitter, LinkedIn or government institutions. I was wondering if you could offer some thoughts on whether that should be made more accessible to, say, the everyday public or to a commercial organization wanting to use it, or if it's okay for it to be concentrated.

Viktor Mayer-Schönberger:

Let me pass the policy question around to Ken but just begin by saying: it is tempting to think the data is concentrated, but perhaps it isn't. Perhaps a lot of the data treasures that we are not aware of existing are actually much more dispersed. Think about predictive maintenance with cars and so forth –

who has that data? Google doesn't have that yet, or Twitter or Facebook. Other companies might have that. So data manifests itself in very odd places.

Kenneth Cukier:

When we were touring Silicon Valley two weeks ago, the view that we were hearing from people was actually just the opposite. When we were mentioning this idea of the data barons and this concentration of data, from their point of view they saw it just in the inverse. They said the means by which people can collect information now is just so vast that no one person is really going to own all the information or have the sort of great benefit with it. So they did acknowledge that Amazon, in terms of shopping behaviour, yes. But they are thinking in terms of the smartphone, how a single app developer can become popular, and that both individuals as well as others can datafy location. We didn't need a large centralized organization to do it anymore.

So I think the way that we see it is that we're in the first inning of big data. It's day one. We're just coming at the outset of something special. So there's still a lot of room to grow and you're going to see a lot of players ebb and fall, or wax and wane. So I'm not too worried about that, and I think that what we will see is – or, I wouldn't want to see some sort of 'must share' provision imposed on industry to share their data. It would be far too early for that.

Question 8:

There was a presentation at Chatham House a couple weeks ago talking about the ability to opt out of the internet. What are the implications for people who do want to opt out? Will society or business penalize them because they won't be contributing to big data in the future?

Viktor Mayer-Schönberger:

I think the answer to that is quite obvious. It seems to me that going forward, opting out of big data is not going to be an option anymore if you want to survive, simply. That is because of individualized health care. Individualized health care will become so much better using big data analysis than what we have with the sort of standardized health care today that the additional years of life expectancy most people won't want to forgo by not opting in to big data. I don't think therefore that our children will want to opt out of big data. In fact our children, when they go to the doctor and the doctor tells them a diagnosis without giving them an empirical basis, they will request one.

Angela Sasse:

Thank you very much for an excellent presentation and a very spirited question-and-answer session. Can we thank the speakers again.