Big data, artificial intelligence, machine learning and data protection

ico.
Information Commissioner's Office

Information Commissioner's Office

Contents

Information Commissioner's foreword 3

Chapter 1 – Introduction 5
What do we mean by big data, AI and machine learning? 6
What's different about big data analytics? 9
What are the benefits of big data analytics? 15

Chapter 2 – Data protection implications 19
Fairness 19
Effects of the processing 20
Expectations 22
Transparency 27
Conditions for processing personal data 29
Consent 29
Legitimate interests 32
Contracts 35
Public sector 35
Purpose limitation 37
Data minimisation: collection and retention 40
Accuracy 43
Rights of individuals 46
Subject access 46
Other rights 47
Security 49
Accountability and governance 51
Data controllers and data processors 56

Chapter 3 – Compliance tools 58
Anonymisation 58
Privacy notices 62
Privacy impact assessments 70
Privacy by design 72
Privacy seals and certification 75
Ethical approaches 77
Personal data stores 84
Algorithmic transparency 86

Chapter 4 – Discussion 90

Chapter 5 – Conclusion 94

Chapter 6 – Key recommendations 97

Annex 1 – Privacy impact assessments for big data analytics 99

Information Commissioner's foreword

Big data is no fad. Since 2014 when my office's first paper on this subject was published, the application of big data analytics has spread throughout the public and private sectors. Almost every day I read news articles about its capabilities and the effects it is having, and will have, on our lives. My home appliances are starting to talk to me, artificially intelligent computers are beating professional board-game players and machine learning algorithms are diagnosing diseases.

The fuel propelling all these advances is big data – vast and disparate datasets that are constantly and rapidly being added to. And what exactly makes up these datasets? Well, very often it is personal data. The online form you filled in for that car insurance quote. The statistics your fitness tracker generated from a run. The sensors you passed when walking into the local shopping centre. The social-media postings you made last week. The list goes on…

So it's clear that the use of big data has implications for privacy, data protection and the associated rights of individuals – rights that will be strengthened when the General Data Protection Regulation (GDPR) is implemented. Under the GDPR, stricter rules will apply to the collection and use of personal data. In addition to being transparent, organisations will need to be more accountable for what they do with personal data. This is no different for big data, AI and machine learning.

However, implications are not barriers. It is not a case of big data 'or' data protection, or big data 'versus' data protection. That would be the wrong conversation. Privacy is not an end in itself, it is an enabling right. Embedding privacy and data protection into big data analytics enables not only societal benefits such as dignity, personality and community, but also organisational benefits like creativity, innovation and trust. In short, it enables big data to do all the good things it can do. Yet that's not to say someone shouldn't be there to hold big data to account.

In this world of big data, AI and machine learning, my office is more relevant than ever. I oversee legislation that demands fair, accurate and non-discriminatory use of personal data; legislation that also gives me the power to conduct audits, order corrective action and issue monetary penalties. Furthermore, under the GDPR my office will be working hard to improve standards in the use of personal data through the implementation of privacy seals and certification schemes. We're uniquely placed to provide the right framework for the regulation of big data, AI and machine learning, and I strongly believe that our efficient, joined-up and co-regulatory approach is exactly what is needed to pull back the curtain in this space.

So the time is right to update our paper on big data, taking into account the advances made in the meantime and the imminent implementation of the GDPR. Although this is primarily a discussion paper, I do recognise the increasing utilisation of big data analytics across all sectors and I hope that the more practical elements of the paper will be of particular use to those thinking about, or already involved in, big data.

This paper gives a snapshot of the situation as we see it. However, big data, AI and machine learning is a fast-moving world and this is far from the end of our work in this space. We'll continue to learn, engage, educate and influence – all the things you'd expect from a relevant and effective regulator.

Elizabeth Denham
Information Commissioner

Chapter 1 – Introduction

1. This discussion paper looks at the implications of big data, artificial intelligence (AI) and machine learning for data protection, and explains the ICO's views on these.

2. We start by defining big data, AI and machine learning, and identifying the particular characteristics that differentiate them from more traditional forms of data processing. After recognising the benefits that can flow from big data analytics, we analyse the main implications for data protection. We then look at some of the tools and approaches that can help organisations ensure that their big data processing complies with data protection requirements. We also discuss the argument that data protection, as enacted in current legislation, does not work for big data analytics, and we highlight the increasing role of accountability in relation to the more traditional principle of transparency.

3. Our main conclusions are that, while data protection can be challenging in a big data context, the benefits will not be achieved at the expense of data privacy rights; and meeting data protection requirements will benefit both organisations and individuals. After the conclusions we present six key recommendations for organisations using big data analytics. Finally, in the paper's annex we discuss the practicalities of conducting privacy impact assessments in a big data context.

4. The paper sets out our views on the issues, but this is intended as a contribution to discussions on big data, AI and machine learning and not as a guidance document or a code of practice. It is not a complete guide to the relevant law. We refer to the new EU General Data Protection Regulation (GDPR), which will apply from May 2018, where it is relevant to our discussion, but the paper is not a guide to the GDPR. Organisations should consult our website ico.org.uk for our full suite of data protection guidance.

5. This is the second version of the paper, replacing what we published in 2014. We received useful feedback on the first version and, in writing this paper, we have tried to take account of it and new developments. Both versions are based on extensive desk research and discussions with business, government and other stakeholders. We're grateful to all who have contributed their views.

What do we mean by big data, AI and machine learning?

6. The terms 'big data', 'AI' and 'machine learning' are often used interchangeably but there are subtle differences between the concepts.

7. A popular definition of big data, provided by the Gartner IT glossary, is:

> "…high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making."[1]

Big data is therefore often described in terms of the 'three Vs' where volume relates to massive datasets, velocity relates to real-time data and variety relates to different sources of data. Recently, some have suggested that the three Vs definition has become tired through overuse[2] and that there are multiple forms of big data that do not all share the same traits[3]. While there is no unassailable single definition of big data, we think it is useful to regard it as data which, due to several varying characteristics, is difficult to analyse using traditional data analysis methods.

8. This is where AI comes in. The Government Office for Science's recently published paper on AI provides a handy introduction that defines AI as:

> "…the analysis of data to model some aspect of the world. Inferences from these models are then used to predict and anticipate possible future events."[4]

[1] Gartner IT glossary Big data. http://www.gartner.com/it-glossary/big-data Accessed 20 June 2016
[2] Jackson, Sean. Big data in big numbers - it's time to forget the 'three Vs' and look at real-world figures. Computing, 18 February 2016. http://www.computing.co.uk/ctg/opinion/2447523/big-data-in-big-numbers-its-time-to-forget-the-three-vs-and-look-at-real-world-figures Accessed 7 December 2016 Accessed 7December 2016
[3] Kitchin, Rob and McArdle, Gavin. What makes big data, big data? Exploring the ontological characteristics of 26 datasets. Big Data and Society, January-June 2016 vol. 3 no. 1. Sage, 17 February 2016.
[4] Government Office for Science. Artificial intelligence: opportunities and implications for the future of decision making. 9 November 2016.

This may not sound very different from standard methods of data analysis. But the difference is that AI programs don't linearly analyse data in the way they were originally programmed. Instead they learn from the data in order to respond intelligently to new data and adapt their outputs accordingly[5]. As the Society for the Study of Artificial Intelligence and Simulation of Behaviour puts it, AI is therefore ultimately about:

> "…giving computers behaviours which would be thought intelligent in human beings."[6]

9. It is this unique ability that means AI can cope with the analysis of big data in its varying shapes, sizes and forms. The concept of AI has existed for some time, but rapidly increasing computational power (a phenomenon known as Moore's Law) has led to the point at which the application of AI is becoming a practical reality.

10. One of the fasting-growing approaches[7] by which AI is achieved is machine learning. iQ, Intel's tech culture magazine, defines machine learning as:

> "…the set of techniques and tools that allow computers to 'think' by creating mathematical algorithms based on accumulated data."[8]

Broadly speaking, machine learning can be separated into two types of learning: supervised and unsupervised. In supervised learning, algorithms are developed based on labelled datasets. In this sense, the algorithms have been trained how to map from input to output by the provision of data with 'correct' values already assigned to them. This initial 'training' phase creates models of the world on which predictions can then be made in the second 'prediction' phase.

---

[5] The Outlook for Big Data and Artificial Intelligence (AI). IDG Research, 11 November 2016 https://idgresearch.com/the-outlook-for-big-data-and-artificial-intelligence-ai/ Accessed 7 December 2016.

[6] The Society for the Study of Artificial Intelligence and Simulation of Behaviour. What is Artificial Intelligence. AISB Website. http://www.aisb.org.uk/public-engagement/what-is-ai Accessed 15 February 2017

[7] Bell, Lee. Machine learning versus AI: what's the difference? Wired, 2 December 2016. http://www.wired.co.uk/article/machine-learning-ai-explained Accessed 7 December 2016

[8] Landau, Deb. Artificial Intelligence and Machine Learning: How Computers Learn. iQ, 17 August 2016. https://iq.intel.com/artificial-intelligence-and-machine-learning/ Accessed 7 December 2016.

Conversely, in unsupervised learning the algorithms are not trained and are instead left to find regularities in input data without any instructions as to what to look for.[9] In both cases, it's the ability of the algorithms to change their output based on experience that gives machine learning its power.

11. In summary, big data can be thought of as an asset that is difficult to exploit. AI can be seen as a key to unlocking the value of big data; and machine learning is one of the technical mechanisms that underpins and facilitates AI. The combination of all three concepts can be called 'big data analytics'. We recognise that other data analysis methods can also come within the scope of big data analytics, but the above are the techniques this paper focuses on.

[9] Alpaydin, Ethem. Introduction to machine learning. MIT press, 2014.

What's different about big data analytics?

12. Big data, AI and machine learning are becoming part of business as usual for many organisations in the public and private sectors. This is driven by the continued growth and availability of data, including data from new sources such as the Internet of Things (IoT), the development of tools to manage and analyse it, and growing awareness of the opportunities it creates for business benefits and insights. One indication of the adoption of big data analytics comes from Gartner, the IT industry analysts, who produce a series of 'hype cycles', charting the emergence and development of new technologies and concepts. In 2015 they ceased their hype cycle for big data, because they considered that the data sources and technologies that characterise big data analytics are becoming more widely adopted as it moves from hype into practice[10]. This is against a background of a growing market for big data software and hardware, which it is estimated will grow from £83.5 billion worldwide in 2015 to £128 billion in 2018[11].

13. Although the use of big data analytics is becoming common, it is still possible to see it as a step change in how data is used, with particular characteristics that distinguish it from more traditional processing. Identifying what is different about big data analytics helps to focus on features that have implications for data protection and privacy.

14. Some of the distinctive aspects of big data analytics are:

    - the use of algorithms
    - the opacity of the processing
    - the tendency to collect 'all the data'
    - the repurposing of data, and
    - the use of new types of data.

---

[10] Sharwood, Simon. Forget big data hype says Gartner as it cans its hype cycle. The Register, 21 August 2015. http://www.theregister.co.uk/2015/08/21/forget_big_data_hype_says_gartner_as_it_cans_its_hype_cycle/ and Heudecker, Nick. Big data isn't obsolete. It's normal. Gartner Blog Network, 20 August 2015. http://blogs.gartner.com/nick-heudecker/big-data-is-now-normal/ Both accessed 12 February 2016

[11] Big data market to be worth £128bn within three years. DataIQ News, 24 May 2016. http://www.dataiq.co.uk/news/big-data-market-be-worth-ps128bn-within-three-years Accessed 17 June 2016

In our view, all of these can potentially have implications for data protection.

15. **Use of algorithms**. Traditionally, the analysis of a dataset involves, in general terms, deciding what you want to find out from the data and constructing a query to find it, by identifying the relevant entries. Big data analytics, on the other hand, typically does not start with a predefined query to test a particular hypothesis; it often involves a 'discovery phase' of running large numbers of algorithms against the data to find correlations[12]. The uncertainty of the outcome of this phase of processing has been described as 'unpredictability by design'[13]. Once relevant correlations have been identified, a new algorithm can be created and applied to particular cases in the 'application phase'. The differentiation between these two phases can be regarded more simply as 'thinking with data' and 'acting with data'[14]. This is a form of machine learning, since the system 'learns' which are the relevant criteria from analysing the data. While algorithms are not new, their use in this way is a feature of big data analytics.

16. **Opacity of the processing**. The current 'state of the art' in machine learning is known as deep learning[15], which involves feeding vast quantities of data through non-linear neural networks that classify the data based on the outputs from each successive layer[16]. The complexity of the processing of data through such massive networks creates a 'black box' effect. This causes an inevitable opacity that makes it very difficult to understand the reasons for decisions made as a result of deep learning[17]. Take, for instance, Google's AlphaGo, a

[12] Centre for Information Policy Leadership. Big data and analytics. Seeking foundations for effective privacy guidance. Hunton and Williams LLP, February 2013 http://www.hunton.com/files/Uploads/Documents/News_files/Big_Data_and_Analytics_February_2013.pdf Accessed 17 June 2016

[13] Edwards, John and Ihrai, Said. Communique on the 38th International Conference of Data Protection and Privacy Commissioners. ICDPPC, 18 October 2016.

[14] Information Accountability Foundation. IAF Consultation Contribution: "Consent and Privacy" – IAF response to the "Consent and Privacy" consultation initiated by the Office of the Privacy Commissioner of Canada. IAF Website, July 2016. http://informationaccountability.org/wp-content/uploads/IAF-Consultation-Contribution-Consent-and-Privacy-Submitted.pdf Accessed 16 February 2017

[15] Abadi, Martin et al. Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. ACM, October 2016.

[16] Marr, Bernard. What Is The Difference Between Deep Learning, Machine Learning and AI? Forbes, 8 December 2016. http://www.forbes.com/sites/bernardmarr/2016/12/08/what-is-the-difference-between-deep-learning-machine-learning-and-ai/#f7b7b5a6457f Accessed 8 December 2016.

[17] Castelvecchi, Davide. Can we open the black box of AI? Nature, 5 October 2016. http://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731 Accessed 8 December 2016

computer system powered by deep learning that was developed to play the board game Go. Although AlphaGo made several moves that were evidently successful (given its 4-1 victory over world champion Lee Sedol), its reasoning for actually making certain moves (such as the infamous 'move 37') has been described as 'inhuman'[18]. This lack of human comprehension of decision-making rationale is one of the stark differentials between big data analytics and more traditional methods of data analysis.

17. **Using all the data**. To analyse data for research, it's often necessary to find a statistically representative sample or carry out random sampling. But a big data approach is about collecting and analysing all the data that is available. This is sometimes referred to as 'n=all'[19]. For example, in a retail context it could mean analysing all the purchases made by shoppers using a loyalty card, and using this to find correlations, rather than asking a sample of shoppers to take part in a survey. This feature of big data analytics has been made easier by the ability to store and analyse ever-increasing amounts of data.

18. **Repurposing data**. A further feature of big data analytics is the use of data for a purpose different from that for which it was originally collected, and the data may have been supplied by a different organisation. This is because the analytics is able to mine data for new insights and find correlations between apparently disparate datasets. Companies such as DataSift[20] enable the analysis of data taken from social media services (via Twitter's GNIP service for example) for marketing and other purposes. The Office for National Statistics (ONS) has experimented with using geolocated Twitter data to infer people's residence and mobility patterns, to supplement official population estimates[21]. Geotagged photos on Flickr, together with the profiles of contributors, have been used as a reliable proxy for estimating visitor numbers at tourist sites and where the visitors have come from[22]. Mobile-phone presence data can be used to

[18] Wood, Georgie. How Google's AI viewed the move no human could understand. Wired, 14 March 2016. https://www.wired.com/2016/03/googles-ai-viewed-move-no-human-understand/ Accessed 8 December 2016.
[19] Mayer-Schönberger, Viktor and Cukier, Kenneth, in Chapter 2 of Big data. A revolution that will transform how we live, work and think. John Murray, 2013
[20] http://datasift.com
[21] Swier, Nigel; Komarniczky, Bence and Clapperton, Ben. Using geolocated Twitter traces to infer residence and mobility. GSS Methodology Series no. 41. ONS, October 2015. http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/the-Data form smart meters ons-big-data-project/index.html Accessed 19 February 2016
[22] Wood, Spencer A et al. Using social media to quantify nature-based tourism and recreation. Nature Scientific Reports, 17 October 2013 http://www.nature.com/articles/srep02976 Accessed 26 February 2016

analyse the footfall in retail centres[23]. Data about where shoppers have come from can be used to plan advertising campaigns. And data about patterns of movement in an airport can be used to set the rents for shops and restaurants.

19. **New types of data**. Developments in technology such as IoT, together with developments in the power of big data analytics mean that the traditional scenario in which people consciously provide their personal data is no longer the only or main way in which personal data is collected. In many cases the data being used for the analytics has been generated automatically, for example by tracking online activity, rather than being consciously provided by individuals. The ONS has investigated the possibility of using data from domestic smart meters to predict the number of people in a household and whether they include children or older people[24]. Sensors in the street or in shops can capture the unique MAC address of the mobile phones of passers-by[25].

20. The data used in big data analytics may be collected via these new channels, but alternatively it may be new data produced by the analytics, rather than being consciously provided by individuals. This is explained in the taxonomy developed by the Information Accountability Foundation[26], which distinguishes between four types of data – provided, observed, derived and inferred:

    - **Provided data** is consciously given by individuals, eg when filling in an online form.

    - **Observed data** is recorded automatically, eg by online cookies or sensors or CCTV linked to facial recognition.

    - **Derived data** is produced from other data in a relatively simple and straightforward fashion, eg calculating customer

[23] Smart Steps increase Morrisons new and return customers by 150%. Telefonica Dynamic Insights, October 2013 http://dynamicinsights.telefonica.com/1158/a-smart-step-ahead-for-morrisons Accessed 20 June 2016

[24] Anderson, Ben and Newing, Andy. Using energy metering data to support official statistics: a feasibility study. Office for National Statistics, July 2015. http://www.ons.gov.uk/aboutus/whatwedo/programmesandprojects/theonsbigdataproject Accessed 26 February 2016

[25] Rice, Simon. How shops can use your phone to track your every move and video display screens can target you using facial recognition. Information Commissioner's Office blog, 21 January 2016. https://iconewsblog.wordpress.com/2016/01/21/how-shops-can-use-your-phone-to-track-your-every-move/ Accessed 17 June 2016

[26] Abrams, Martin. The origins of personal data and its implications for governance. OECD, March 2014. http://informationaccountability.org/wp-content/uploads/Data-Origins-Abrams.pdf Accessed 17 June 2016

profitability from the number of visits to a store and items bought.

- **Inferred data** is produced by using a more complex method of analytics to find correlations between datasets and using these to categorise or profile people, eg calculating credit scores or predicting future health outcomes. Inferred data is based on probabilities and can thus be said to be less 'certain' than derived data.

IoT devices are a source of observed data, while derived and inferred data are produced by the process of analysing the data. These all sit alongside traditionally provided data.

21. Our discussions with various organisations have raised the question whether big data analytics really is something new and qualitatively different. There is a danger that the term 'big data' is applied indiscriminately as a buzz word that does not help in understanding what is happening in a particular case. It is not always easy (or indeed useful) to say whether a particular instance of processing is or is not big data analytics. In some cases it may appear to be simply a continuation of the processing that has always been done; for example, banks and telecoms companies have always handled large volumes of data and credit card issuers have always had to validate purchases in real time. Furthermore, as noted at the start of this section, the technologies and tools that enable big data analytics are increasingly becoming a part of business as usual.

22. For all these reasons, it may be difficult to draw a clear line between big data analytics and more conventional forms of data use. Nevertheless, we think the features we have identified above represent a step change. So it is important to consider the implications of big data analytics for data protection.

23. However, it is also important to recognise that many instances of big data analytics do not involve personal data at all. Examples of non-personal big data include world climate and weather data; using geospatial data from GPS-equipped buses to predict arrival times; astronomical data from radio telescopes in the Square Kilometre Array[27]; and data from sensors on containers carried on ships. These are all areas where big data analytics enable new discoveries and improve services and business processes, without using personal data. Also, big data analytics may not involve personal data for other reasons; in particular it may be possible to successfully anonymise

[27] Square Kilometre Array website https://www.skatelescope.org/ Accessed 17 June 2016

what was originally personal data, so that no individuals can be identified from it. We discuss this in more detail in the section on anonymisation in chapter 3.

24. Still, it is obvious that other examples of big data analytics do involve personal data. The data may directly identify individuals, or they may be identified by apparently anonymous datasets being combined. In such cases, the question of whether the processing complies with data protection principles is unavoidable.

What are the benefits of big data analytics?

25. In 2012 the Centre for Economics and Business Research estimated that the cumulative benefit to the UK economy of adopting big data technologies would amount to £216 billion over the period 2012-17, and £149 billion of this would come from gains in business efficiency[28].

26. There are obvious commercial benefits to companies, for example in being able to understand their customers at a granular level and hence making their marketing more targeted and effective. Consumers may benefit from seeing more relevant advertisements and tailored offers and from receiving enhanced services and products. For example, the process of applying for insurance can be made easier, with fewer questions to answer, if the insurer or the broker can get other data they need through big data analytics.

27. Big data analytics is also helping the public sector to deliver more effective and efficient services, and produce positive outcomes that improve the quality of people's lives. This is shown by the following examples:

> **Health**. In 2009, Public Health England (PHE) was aware that cancer survival rates in the UK were poor compared to Europe, suspecting this might be due to later diagnosis. After requests from Cancer Research UK to quantify how people came to be diagnosed with cancer, the Routes to Diagnosis project was conceived to seek answers to this question.
>
> This was a big data project that involved using complex algorithms to analyse 118 million records on 2 million patients from several data sources. The analysis revealed the ways in which patients were diagnosed with cancer from 2006 to 2013. A key discovery (from results published in 2011) was that in 2006 almost 25% of cancer cases were only diagnosed in an emergency when the patient came to A&E. Patients diagnosed via this route have lower chances of survival compared to other routes. So PHE was able to put in place initiatives to increase diagnosis through other routes. The

[28] Centre for Economics and Business Research Ltd. Data equity: unlocking the value of big data. CEBR, April 2012. http://www.sas.com/offices/europe/uk/downloads/data-equity-cebr.pdf Accessed 17 June 2016

latest results (published in 2015) show that by 2013 just 20% of cancers were diagnosed as an emergency[29].

The understanding gained from this study continues to inform public health initiatives such as PHE's Be Clear on Cancer campaigns, which raise awareness of the symptoms of lung cancer and help people to spot the symptoms early.[30]

Education. Learning analytics in higher education (HE) involves the combination of 'static data' such as traditional student records with 'fluid data' such as swipe card data from entering campus buildings, using virtual learning environments (VLEs) and downloading e-resources. The analysis of this information can reveal trends that help to improve HE processes, benefiting both staff and students. Examples include the following:

- Preventing drop-out via early intervention with students who are identified as disengaged from their studies by analysing VLE login and campus attendance data.

- The ability for tutors to provide high-quality, specific feedback to students at regular intervals (as opposed to having to wait until it is 'too late' – after an exam for instance). The feedback is based on pictures of student performance gleaned from analysis of data from all the systems used by a student during their study.

- Increased self-reflection by students and a desire to improve their performance based on access to their own performance data and the class averages.

- Giving students shorter, more precise lecture recordings based on data analysis that revealed patterns regarding the parts of full lecture recordings that were repeatedly watched (assessment requirements, for example).

[29] Elliss-Brookes, Lucy. Big data in action: the story behind Routes to Diagnosis. Public health matters blog, 10 November 2015. https://publichealthmatters.blog.gov.uk/2015/11/10/big-data-in-action-the-story-behind-routes-to-diagnosis/ Accessed 19 February 2016

[30] Public Health England Press Release, 10 November 2015. https://www.gov.uk/government/news/big-data-driving-earlier-cancer-diagnosis-in-england Accessed 8 December 2016

Such benefits have been seen by HE institutions including Nottingham Trent University, Liverpool John Moores University, the University of Salford and the Open University[31].

Transport. Transport for London (TfL) collects data on 31 million journeys every day including 20 million ticketing system 'taps', location and prediction information for 9,200 buses and traffic-flow information from 6,000 traffic signals and 1,400 cameras. Big data analytics are applied to this data to reveal travel patterns across the rail and bus networks. By identifying these patterns, TfL can tailor its products and services to create benefits to travellers in London such as:

- more informed planning of closures and diversions to ensure as few travellers as possible are affected
- restructuring bus routes to meet the needs of travellers in specific areas of London; for instance, a new service pattern for buses in the New Addington neighbourhood was introduced in October 2015[32]
- building new entrances, exits and platforms to increase capacity at busy tube stations – as at Hammersmith tube station in February 2015[33].

28. It is clear therefore that big data analytics can bring benefits to business, to society and to individuals as consumers and citizens. By recognising these benefits here, we do not intend to set up this paper as a contest between the benefits of big data and the rights given by data protection. To look at big data and data protection through this lens can only be reductive. Although there are implications for data protection (which we discuss in chapter 2), there are also solutions (discussed in chapter 3). It's not a case of big data *or* data

[31] Shacklock, Xanthe. From bricks to clicks. The potential of data and analytics in higher education. Higher Education Commission, January 2016. http://www.sas.com/offices/europe/uk/downloads/data-equity-cebr.pdf Accessed 17 June 2016

[32] Weinstein, Lauren. How TfL uses 'big data' to plan transport services. Eurotransport, 20 June 2016. http://www.eurotransportmagazine.com/19635/past-issues/issue-3-2016/tfl-big-data-transport-services/ Accessed 9 December 2016

[33] Alton, Larry. Improved Public Transport for London, Thanks to Big Data and the Internet of Things. London Datastore, 9 June 2015. https://data.london.gov.uk/blog/improved-public-transport-for-london-thanks-to-big-data-and-the-internet-of-things/ Accessed 9 December 2016

protection, it's big data *and* data protection; the benefits of both can be delivered alongside each other.

Chapter 2 – Data protection implications

Fairness

In brief…

- Some types of big data analytics, such as profiling, can have intrusive **effects** on individuals.
- Organisations need to consider whether the use of personal data in big data applications is within people's reasonable **expectations**.
- The complexity of the methods of big data analysis, such as machine learning, can make it difficult for organisations to be **transparent** about the processing of personal data.

29. Under the first DPA principle, the processing of personal data must be fair and lawful, and must satisfy one of the conditions listed in Schedule 2 of the DPA (and Schedule 3 if it is sensitive personal data as defined in the DPA). The importance of fairness is preserved in the GDPR: Article 5(1)(a) says personal data must be "processed fairly, lawfully and in a transparent manner in relation to the data subject".

30. By contrast, big data analytics is sometimes characterised as sinister or a threat to privacy or simply 'creepy'. This is because it involves repurposing data in unexpected ways, using complex algorithms, and drawing conclusions about individuals with unexpected and sometimes unwelcome effects[34].

31. So a key question for organisations using personal data for big data analytics is whether the processing is fair. Fairness involves several elements. *Transparency* – what information people have about the processing – is essential. But assessing fairness also involves looking

[34] For example, Naughton, John Why big data has made your privacy a thing of the past Guardian online, 6 October 2013 http://www.theguardian.com/technology/2013/oct/06/big-data-predictive-analytics-privacy; Richards Neil M. and King, Jonathan H. Three paradoxes of big data 66 Stanford Law Review Online, 41 3 September 2013 http://www.stanfordlawreview.org/online/privacy-and-big-data/three-paradoxes-big-data; Leonard, Peter. Doing big data business: evolving business models and privacy regulation. August 2013. International Data Privacy Law, 18 December 2013. http://idpl.oxfordjournals.org/content/early/2013/12/18/idpl.ipt032.short?rss=1 All accessed 17 June 2016

at the *effects* of the processing on individuals, and their *expectations* as to how their data will be used[35].

Effects of the processing

32. How big data is used is an important factor in assessing fairness. Big data analytics may use personal data purely for research purposes, eg to detect general trends and correlations, or it may use personal data to make decisions affecting individuals. Some of those decisions will obviously affect individuals more than others. Displaying a particular advert on the internet to an individual based on their social media 'likes', purchases and browsing history may not be perceived as intrusive or unfair, and may be welcomed if it is timely and relevant to their interests. However, in some circumstances even displaying different advertisements can mean that the users of that service are being profiled in a way that perpetuates discrimination, for example on the basis of race[36]. Research in the USA suggested that internet searches for "black-identifying" names generated advertisements associated with arrest records far more often than those for "white-identifying" names[37]. There have also been similar reports of discrimination in the UK, for instance a female doctor was locked out of a gym changing room because the automated security system had profiled her as male due to associating the title 'Dr' with men[38].

33. Profiling can also be used in ways that have a more intrusive effect upon individuals. For example, in the USA, the Federal Trade Commission found evidence of people's credit limits being lowered based on an analysis of the poor repayment histories of other people who shopped at the same stores[39] as them. In that scenario, people are not being discriminated against because

---

[35] Information Commissioner's Office. Guide to data protection. ICO, May 2016. http://ico.org.uk/for_organisations/data_protection/~/media/documents/library/Data_Protection/Practical_application/the_guide_to_data_protection.pdf Accessed 12 December 2016

[36] Rabess, Cecilia Esther. Can big data be racist? The Bold Italic, 31 March 2014. http://www.theboldItalic.com/articles/4502-can-big-data-be-racist Accessed 20 June 2016

[37] Sweeney, Latanya. Discrimination in online ad delivery. Data Privacy Lab, January 2013. http://dataprivacylab.org/projects/onlineads/1071-1.pdf Accessed 20 June 2016

[38] Fleig, Jessica. Doctor locked out of women's changing room because gym automatically registered everyone with Dr title as male. Mirror, 18 March 2015. http://www.mirror.co.uk/news/uk-news/doctor-locked-out-womens-changing-5358594 Accessed 16 December 2016

[39] Federal Trade Commission. Big data: a tool for inclusion or exclusion. FTC, January 2016 https://www.ftc.gov/reports/big-data-tool-inclusion-or-exclusion-understanding-issues-ftc-report Accessed 4 March 2016

they belong to a particular social group. But they are being treated in a certain way based on factors, identified by the analytics, that they share with members of that group.

34. The GDPR includes provisions dealing specifically with profiling, which is defined in Article 4 as:

> "Any form of automated processing of personal data consisting of using those data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements."

35. Recital 58 of the GDPR also refers to examples of automated decision making "like automatic refusal of an on-line credit application or e-recruiting practices without any human intervention". The wording here reflects the potentially intrusive nature of the types of automated profiling that are facilitated by big data analytics. The GDPR does not prevent automated decision making or profiling, but it does give individuals a qualified right not to be subject to purely automated decision making[40]. It also says that the data controller should use "appropriate mathematical or statistical procedures for the profiling" and take measures to prevent discrimination on the basis of race or ethnic origin, political opinions, religion or beliefs, trade union membership, genetic or health status or sexual orientation[41].

36. The 1995 Data Protection Directive and the DPA already contained provisions on automated decision making. Instances of decision making by purely automated means, without human intervention, were hitherto relatively uncommon. But the new capabilities of big data analytics to deploy machine learning mean it is likely to become more of an issue. The more detailed provisions of the GDPR reflect this.

37. Yet not all processing that has an unlooked-for or unwelcome effect on people is necessarily unjustified. In insurance, big data analytics can be used for micro-segmentation of risk groups; it may be possible to identify people within a high-risk (and therefore high-premium) group who actually represent a slightly

[40] GDPR Article 22
[41] GDPR Recital 71

lower risk compared to others in that group. Their premiums can be adjusted accordingly in their favour. In this case big data is being used to give a more accurate assessment of risk that benefits those individuals as well as the insurer. The corollary of this, given that insurance is about pooling risk, is that the remaining high-risk group members may find they have to pay higher premiums. Arguably this is a fair result overall, but inevitably there are winners and losers. And to the losers the process may seem 'creepy' or unfair.

38. This means that if big data organisations are using personal data, then as part of assessing fairness they need to be aware of and factor in the effects of their processing on the individuals, communities and societal groups concerned. Given the sometimes novel and unexpected ways in which data is used in the analytics, this may be less straightforward than in more conventional data-processing scenarios. Privacy impact assessments provide a structured approach to doing this, and we discuss their use in the section on privacy impact assessments in chapter 3.

Expectations

39. Fairness is also about expectations; would a particular use of personal data be within the reasonable expectations of the people concerned? An organisation collecting personal data will generally have to provide a privacy notice explaining the purposes for which they need the data, but this may not necessarily explain the detail of how the data will be used. It is still important that organisations consider whether people could reasonably expect their data to be used in the ways that big data analytics facilitates.

40. There is also a difference between a situation where the purpose of the processing is naturally connected with the reason for which people use the service and one where the data is being used for a purpose that is unrelated to the delivery of the service. An example of the former is a retailer using loyalty card data for market research; there would be a reasonable expectation that they would use that data to gain a better understanding of their customers and the market in which they operate. An example of the latter is a social-media company making its data available for market research; when people post on social media, is it reasonable to expect this information could be used for unrelated purposes? This does not mean that such use is necessarily unfair; it depends on various factors that make up people's overall expectations of reasonableness, such

as what they are told when they join and use the social-media service.

41. Deciding what is a reasonable expectation is linked to the issue of transparency and the use of privacy notices, and also to the principle of purpose limitation, ie whether any further use of the data is incompatible with the purpose for which it was obtained. We discuss both transparency and purpose limitation below, but it is also important for an organisation to consider in general terms whether the use of personal data in a big data application is within people's reasonable expectations.

42. This inevitably raises the wider question of people's attitudes to the use of their personal data. The view is often put forward that people are becoming less concerned about how organisations use their personal data. This is said to be particularly true of 'digital natives', younger people who have grown up with ubiquitous internet access and who are happy to share personal information via social media with little concern for how it may be used. For example, the Direct Marketing Association commissioned the Future Foundation to look into attitudes to use of personal data in 2012 and 2015[42]. They found that the percentage of 'fundamentalists' who won't share their data fell from 31% to 24% and the percentage of 'not concerned' increased from 16% to 22%.

43. If it were true that people are simply unconcerned about how their personal data is used, this would mean their expectations about potential data use are open-ended, leaving a very wide margin of discretion for big data organisations. However, research suggests that this view is too simplistic; the reality is more nuanced:

The International Institute of Communications (IIC). Research commissioned by the IIC[43] showed that people's willingness to give personal data, and their attitude to how that data will be used, is context-specific. The context depends on a number of variables, eg how far an individual trusts the organisation and what information is being asked for.

[42] Combemale, Chris. Taking the leap of faith. DataIQ, 15 September 2015. http://www.dataiq.co.uk/blog/taking-leap-faith Accessed 18 March 2016
[43] International Institute of Communications. Personal data management: the user's perspective. International Institute of Communications, September 2012.

The Boston Consulting Group (BCG). The BCG[44] found that for 75% of consumers in most countries, the privacy of personal data remains a top issue, and that young people aged 18-24 are only slightly less cautious about the use of personal online data than older age groups.

KPMG. A global survey by KPMG[45] found that, while attitudes to privacy varied (based on factors such as types of data, data usage and consumer location), on average 56% of respondents reported being "concerned" or "extremely concerned" about how companies were using their personal data.

44. Some studies have pointed to a 'privacy paradox': people may express concerns about the impact on their privacy of 'creepy' uses of their data, but in practice they contribute their data anyway via the online systems they use. In other words they provide the data because it is the price of using internet services. For instance, findings from Pybus, Coté and Blanke's study of mobile phone usage by young people in the UK[46] and two separate studies by Shklovski et al. [47], looking at smartphone usage in Western Europe, supported the idea of the privacy paradox. It has also been argued that the prevalence of web tracking means that, in practice, web users have no choice but to enter into an 'unconscionable contract' to allow their data to be used[48]. This suggests that people may be resigned to the use of their data because they feel there is no alternative, rather than being indifferent or positively welcoming it. This was the finding of a study of US consumers by the Annenberg School for

[44] Rose, John et al. The trust advantage: how to win with big data. Boston Consulting Group, November 2013. https://www.bcgperspectives.com/content/articles/information_technology_strategy_consumer_products_trust_advantage_win_big_data/ Accessed 17 June 2016

[45] KPMG. Crossing the line: Staying on the right side of consumer privacy. KPMG, November 2016. https://assets.kpmg.com/content/dam/kpmg/xx/pdf/2016/11/crossing-the-line.pdf Accessed 23 January 2017

[46] Pybus, Jennifer; Cote, Mark; Blanke, Tobias. Hacking the social life of Big Data Big Data & Society, July-December 2015 vol. 2 no. 2 http://m.bds.sagepub.com/content/2/2/2053951715616649 Accessed 18 March 2016

[47] Shklovski, Irina et al. Leakiness and creepiness in app space: Perceptions of privacy and mobile app use. In Proceedings of the 32nd annual ACM conference on Human factors in computing systems, pp. 2347-2356. ACM, 2014

[48] Peacock, Sylvia E. How web tracking changes user agency in the age of Big Data; the used user. Big data and society, July-December 2014 Vol 1 no 2. http://m.bds.sagepub.com/content/1/2/2053951714564228 Accessed 23 March 2016

Communication[49]. The study criticised the view that consumers continued to provide data to marketers because they are consciously engaging in trading personal data for benefits such as discounts; instead, it concluded that most Americans believe it is futile to try to control what companies can learn about them. They did not want to lose control over their personal data but they were simply resigned to the situation.

45. In some cases the fact that people continue to use services that extract and analyse their personal data may also mean they invest a certain level of trust in those organisations, particularly those that are major service providers or familiar brands; they trust that the organisation will not put their data to bad use. Given the practical difficulty of reading and understanding terms and conditions and of controlling the use of one's data, this is at least pragmatic. At the same time, it obliges the organisation to exercise proper stewardship of the data, so as not to exploit people's trust. We return to this point in the section on ethical approaches in chapter 3.

46. In the UK, a survey for Digital Catapult[50] showed a generally low level of trust. The public sector was the most trusted to use personal data responsibly, by 44% of respondents; financial services was the next most trusted sector, but only by 29% of respondents. Other sectors had a much lower rating. On the other hand, the survey found that a significant proportion of people were happy for their data to be shared for purposes such as education and health. These themes – a feeling of resignation despite a general lack of trust, combined with a willingness for data to be used for socially useful purposes – were reflected in a report from Sciencewise[51] which summarised several recent surveys on public attitudes to data use.

47. A previous US study[52] suggested that if people had concerns about data use, particularly by companies, these were really

[49] Turow, Joseph; Hennessy, Michael and Draper, Nora. The tradeoff fallacy: how marketers are misrepresenting American consumers and opening them up to exploitation. University of Pennsylvania. Annenberg School for Communication, June 2015. https://www.asc.upenn.edu/sites/default/files/TradeoffFallacy_1.pdf Accessed 31 March 2016

[50] Trust in personal data: a UK review. Digital Catapult, 29 July 2015. http://www.digitalcatapultcentre.org.uk/pdtreview/ Accessed 30 March 2016

[51] Big data. Public views on the collection, sharing and use of big data by governments and companies. Sciencewise, April 2014. http://www.sciencewise-erc.org.uk/cms/public-views-on-big-data/ Accessed 30 March 2016

[52] Forbes Insights and Turn. The promise of privacy. Respecting consumers' limits while realizing the marketing benefits of big data. Forbes Insights, 2013

about the security of the data, not about privacy. However, this was not borne out by the UK and European surveys referred to in the Sciencewise report. These identified several concerns to do with:

Surveillance. A 2013 study by the Wellcome Trust, consisting of focus groups and telephone interviews, found there to be a "widespread wariness" about being spied on by government, corporations and criminals.

Discrimination. The same study revealed concerns about possible discrimination against people based on medical data, for instance where such data is shared with employers who might make discriminatory decisions about people because of mental health issues.

Consent. An online survey conducted by Demos in 2012 found that people's top concern for personal data use was about companies using it without their permission.

Data sharing. The Institute for Insight in the Public Services conducted a telephone survey in 2008 which revealed that while people are generally happy for their personal data to be held by one organisation, they are concerned when it is shared with others. These concerns centred on loss of control over personal data and fears that errors in the data would be perpetuated through sharing.

48. There is also evidence of people trying to exercise a measure of privacy protection by deliberately giving false data. A study by Verve found that 60% of UK consumers intentionally provide incorrect information when submitting their personal details online, which is a problem for marketers[53]. Even in younger people, attitudes seem to be changing, with a trend among 'Generation Z' towards using social media apps that appear to be more privacy friendly[54].

---

http://images.forbes.com/forbesinsights/StudyPDFs/turn_promise_of_privacy_report.pdf Accessed 20 June 2016

[53] Chahal, Mindi. Consumers are 'dirtying' databases with false details. Marketing Week 8 July 2015. https://www.marketingweek.com/2015/07/08/consumers-are-dirtying-databases-with-false-details/ Accessed 18 March 2016

[54] Williams, Alex. Move over millennials, here comes Generation Z. New York Times, 18 September 2015. http://www.nytimes.com/2015/09/20/fashion/move-over-millennials-here-comes-generation-z.html?_r=0 Accessed 18 March 2016

49. Microsoft's Digital Trends report 2015[55] noted a trend it called Right to My Identity. This means that, rather than simply wishing to preserve privacy through anonymity, a significant percentage of global consumers now want to be able to control how long information they have shared stays online, and are also interested in services that help them manage their digital identity. This suggests consumers have increasing expectations of how organisations will use their data and want to be able to influence it.

50. That people continue to provide personal data and use services that collect data from them does not necessarily mean they are happy about how their data is used or simply indifferent. Many people may be resigned to a situation over which they feel they have no real control, but there is evidence of people's concerns about data use, and also of their desire to have more control over how their data is used. This leads to the conclusion that expectations are a significant issue that needs to be addressed in assessing whether a particular instance of big data processing is fair.

Transparency

51. The complexity of big data analytics can mean that the processing is opaque to citizens and consumers whose data is being used. It may not be apparent to them their data is being collected (eg, their mobile phone location), or how it is being processed (eg, when their search results are filtered based on an algorithm – the so-called "filter bubble" effect[56]). Similarly, it may be unclear how decisions are being made about them, such as the use of social-media data for credit scoring.

52. This opacity can lead to a lack of trust that can affect people's perceptions of and engagement with the organisation doing the processing. This can be an issue in the public sector, where lack of public awareness can become a barrier to data sharing. Inadequate provision of information to the public about data use has been seen as a barrier to the roll-out of the care.data project in the NHS[57]. A study for the Wellcome Trust into public

[55] Digital trends 2015. Microsoft, March 2015. http://fp.advertising.microsoft.com/en/wwdocs/user/display/english/insights/Microsoft-Advertising-Digital-Trends.pdf Accessed 31 March 2016
[56] Pariser, Eli. Beware online "filter bubbles". TED Talk, March 2011. http://www.ted.com/talks/eli_pariser_beware_online_filter_bubbles/transcript?language=en Accessed 1 April 2016
[57] House of Commons Science and Technology Committee. The big data dilemma. Fourth report of session 2015-16 HC468. The Stationery Office, 12 February 2016.

attitudes to the use of data in the UK[58] found a low level of understanding and awareness of how anonymised health and medical data is used and of the role of companies in medical research. People had some expectations about the use of the data when they were dealing with a company (though they were unaware of some of the uses of their social-media data), and these differed from their expectations when using public-health services. However, they were not aware of how their health data might also be used by companies for research. The report referred to this as an example of "context collapse".

53. In the private sector, a lack of transparency can also mean that companies miss out on the competitive advantage that comes from gaining consumer trust. The BCG[59] stresses the importance of "informed trust", which inevitably means being more open about the processing:

> "Personal data collected by businesses cannot be treated as mere property, transferred once and irrevocably, like a used car, from data subject to data user. Data sharing will succeed only if the organizations involved earn the informed trust of their customers. Many such arrangements today are murky, furtive, undisclosed; many treat the data subject as a product to be resold, not a customer to be served. Those businesses risk a ferocious backlash, while their competitors are grabbing a competitive advantage by establishing trust and legitimacy with customers."

54. While the use of big data has implications regarding the transparency of the processing of personal data, it is still a key element of fairness. The DPA contains a specific transparency requirement, in the form of a 'fair processing notice', or more simply a privacy notice. Privacy notices are discussed in more detail in chapter 3 as a tool that can aid compliance with the transparency principle in a big data context.

---

http://www.publications.parliament.uk/pa/cm201516/cmselect/cmsctech/468/468.pdf Accessed 8 April 2016

[58] Ipsos MORI Social Research Institute. The one-way mirror: public attitudes to commercial access to health data. Ipsos MORI, March 2016. http://www.wellcome.ac.uk/stellent/groups/corporatesite/@msh_grants/documents/web_document/wtp060244.pdf Accessed 8 April 2016

[59] Evans, Philip and Forth, Patrick. Borges' map: navigating a world of digital disruption. Boston Consulting Group, 2 April 2015. https://www.bcgperspectives.com/content/articles/borges-map-navigating-world-digital-disruption/ Accessed 8 April 2016

Conditions for processing personal data

In brief...

- Obtaining meaningful **consent** is often difficult in a big data context, but novel and innovative approaches can help.
- Relying on the **legitimate interests** condition is not a 'soft option'. Big data organisations must always balance their own interests against those of the individuals concerned.
- It may be difficult to show that big data analytics are strictly necessary for the performance of a **contract**.
- Big data analysis carried out in the **public sector** may be legitimised by other conditions, for instance where processing is necessary for the exercise of functions of a government department.

55. Under the first DPA principle, the processing of personal data must not only be fair and lawful, but must also satisfy one of the conditions listed in Schedule 2 of the DPA (and Schedule 3 if it is sensitive personal data as defined in the DPA). This applies equally to big data analytics that use personal data. The Schedule 2 conditions that are most likely to be relevant to big data analytics, particularly in a commercial context, are consent, whether processing is necessary for the performance of a contract, and the legitimate interests of the data controller or other parties. Our Guide to data protection[60] explains these conditions in more detail. Here we consider how they relate to big data analytics specifically.

Consent

56. If an organisation is relying on people's consent as the condition for processing their personal data, then that consent must be a freely given, specific, and informed indication that they agree to the processing[61]. This means people must be able to understand what the organisation is going to do with their data ("specific

[60] Information Commissioner's Office. Guide to data protection. ICO, May 2016. http://ico.org.uk/for_organisations/data_protection/~/media/documents/library/Data_Protection/Practical_application/the_guide_to_data_protection.pdf Accessed 20 June 2016

[61] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and of the free movement of such data. Article 2(h)

and informed") and there must be a clear indication that they consent to it.

57. The GDPR makes it clearer that the consent must also be "unambiguous" and that it must be a "clear affirmative action" such as ticking a box on a website or choosing particular technical settings for "information society services"[62] (services delivered over the internet, eg a social-networking app). Furthermore, the data controller must be able to demonstrate that the consent was given, and the data subject must be able to withdraw that consent[63].

58. It has been suggested that the so-called 'notice and consent' model, where an organisation tells data subjects what it is going to do with their data, is not practical in a big data context. The opaque nature of analysis using AI techniques can make it difficult for meaningful consent to be provided[64], but consent has also been criticised because it is 'binary', ie it only gives people a yes/no choice at the outset. This is seen as incompatible with big data analytics due to its experimental nature and its propensity to find new uses for data, and also because it may not fit contexts where data is observed rather than directly provided by data subjects[65].

59. However, there are new approaches to consent that go beyond the simple binary model. It may be possible to have a process of graduated consent, in which people can give consent or not to different uses of their data throughout their relationship with a service provider, rather than having a simple binary choice at the start. This can be linked to 'just in time' notifications. For example, at the point when an app wants to use mobile phone location data or share data with a third party, the user can be asked to give their consent.

60. A recent report by the European Union Agency for Network and Information Security (ENISA) found positive developments in the way consent is obtained and that it is not a real barrier to usability. It called for more technical innovation in the methods of obtaining consent:

[62] GDPR Article 4(11) and Recital 32
[63] GDPR Article 7
[64] Buttarelli, Giovanni. A smart approach: counteract the bias in artificial intelligence. European Data Protection Supervisor, 8 November 2016. https://secure.edps.europa.eu/EDPSWEB/edps/pid/696 Accessed 13 December 2016.
[65] Nguyen, M-H Carolyn et al. A user-centred approach to the data dilemma: context, architecture and policy. Digital Enlightenment Forum Yearbook, 2013.

> "Practical implementation of consent in big data should go beyond the existing models and provide more automation, both in the collection and withdrawal of consent. Software agents providing consent on user's behalf based on the properties of certain applications could be a topic to explore. Moreover, taking into account the sensors and smart devices in big data, other types of usable and practical user positive actions, which could constitute consent (e.g. gesture, spatial patterns, behavioral patterns, motions), need to be analysed."[66]

61. The Royal Academy of Engineering looked at the benefits of big data analytics in several sectors, and the risks to privacy. In the health sector, they suggested that in cases where personal data is being used with consent and anonymisation is not possible, consent could be time limited so that the data is no longer used after the time limit has expired[67]. This is in addition to the principle that, if people have given consent, they can also withdraw it at any time. They said that when seeking consent, the government and the NHS should take a patient-centric approach and explain the societal benefits and the effect on privacy.

62. These examples suggest that the complexity of big data analytics need not be an obstacle to seeking consent. If an organisation can identify potential benefits from using personal data in big data analytics, it should be able to explain these to users and seek consent, if that is the condition it chooses to rely on. It must find the point at which to explain the benefits of the analytics and present users with a meaningful choice – and then respect that choice when processing their personal data.

63. If an organisation buys a large dataset of personal data for analytics purposes, it then becomes a data controller regarding that data. The organisation needs to be sure it has met a condition in the DPA for the further use of that data. If it is relying on the original consent obtained by the supplier as that

[66] D'Acquisito, Giuseppe et al. Privacy by design in big data. An overview of privacy enhancing technologies in the era of big data analytics. ENISA, December 2015. https://www.enisa.europa.eu/activities/identity-and-trust/library/deliverables/big-data-protection Accessed 19 April 2016

[67] Royal Academy of Engineering. Connecting data: driving productivity and innovation. Royal Academy of Engineering, 16 November 2015. http://www.raeng.org.uk/publications/reports/connecting-data-driving-productivity Accessed 19 April 2016

condition, it should ensure this covers the further processing it plans for the data. This issue often arises in the context of marketing databases. Our guidance on direct marketing[68] explains how the DPA (and the Privacy and Electronic Communications regulations) apply to the issue of indirect, or 'third party' consent.

64. Just because people have put data onto social media without restricting access does not necessarily legitimise all further use of it. The fact that data can be viewed by all does not mean anyone is entitled to use it for any purpose or that the person who posted it has implicitly consented to further use. This is particularly an issue if social-media analytics is used to profile individuals, rather than for general sentiment analysis (the study of people's opinions[69]). If a company is using social-media data to profile individuals, eg for recruitment purposes or for assessing insurance or credit risk, it needs to ensure it has a data protection condition for processing the data. Individuals may have consented to this specifically when they joined the social-media service, or the company may seek their consent, for example as part of a service to help people manage their online presence. If the company does not have consent, it needs to consider what other data protection conditions may be relevant.

65. The processing of personal data has to meet only one of the conditions in the DPA or the GDPR. Consent is one condition for processing personal data. But it is not the only condition available, and it does not have any greater status than the others. In some circumstances consent will be required, for example for electronic marketing calls and messages[70], but in others a different condition may be appropriate.

Legitimate interests

66. The condition in schedule 2 condition 6 of the DPA is that the processing is necessary for the legitimate interests of the organisation collecting the data (or others to whom it is made available). However, the processing must not be "unwarranted" because of prejudice to the rights, freedoms or legitimate

[68] Information Commissioner's Office. Direct marketing. ICO, May 2016. https://ico.org.uk/media/for-organisations/documents/1555/direct-marketing-guidance.pdf Accessed 20 June 2016
[69] Liu, Bing, and Zhang, Lei. A survey of opinion mining and sentiment analysis. In Mining text data, pp. 415-463. Springer US, 2012
[70] See our Direct marketing guidance for more detail on this point

interests of the data subjects. In the GDPR, this condition is expressed as follows:

> "Processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data, in particular where the data subject is a child."[71]

67. An organisation may have several legitimate interests that could be relevant, including profiling customers in order to target its marketing; preventing fraud or the misuse of its services; and physical or IT security. However, to meet this condition the processing must be "necessary" for the legitimate interests. This means it must be more than just potentially interesting. The processing is not necessary if there is another way of meeting the legitimate interest that interferes less with people's privacy.

68. Having established its legitimate interest, the organisation must then do a balancing exercise between those interests and the rights and legitimate interests of the individuals concerned. So organisations seeking to rely on this condition must pay particular attention to how the analytics will affect people's privacy. This can be a complex assessment involving several factors. The opinion of the Article 29 Working Party[72] on legitimate interests under the current Data Protection Directive sets out in detail how to assess these factors and do the balancing exercise.

69. The legitimate interests condition is one alternative to seeking data subjects' active consent. If an organisation is relying on it to legitimise its big data processing, it need not seek the consent of the individuals concerned, but it still has to tell them what it is doing, in line with the fairness requirement. Furthermore, the European Data Protection Supervisor has suggested[73] that in big data cases where it is difficult to strike a

[71] GDPR Article 6(1)(f)
[72] Article 29 Data Protection Working Party. Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC. European Commission 9 April 2014. http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp217_en.pdf Accessed 20 June 2016
[73] European Data Protection Supervisor. Meeting the challenges of big data. Opinion 7/2015. EDPS, 19 November 2015.

balance between the legitimate interests of the organisation and the rights and interests of the data subject, it may be helpful to also give people the opportunity of an opt-out. While an opt-out would not necessarily satisfy all the DPA requirements for valid consent, this 'belt and braces' approach could help to safeguard the rights and interests of the data subjects.

70. The legitimate interests condition is not a soft option for the organisation; it means it takes on more responsibility. Under the consent condition, while the organisation must ensure its processing is fair and satisfies data protection principles, the individual is responsible for agreeing (or not) to the processing, which may not proceed without their consent. By contrast, the legitimate interests condition places the responsibility on the organisation to carry out an assessment and proceed in a way that respects people's rights and interests.

71. This means a big data organisation will have to have a framework of values against which to test the proposed processing, and a method of carrying out the assessment and keeping the processing under review. It will also have to be able to demonstrate it has these elements in place, in case of objections by the data subjects or investigations by the regulator. It should also be noted that under the GDPR if a data controller is relying on legitimate interests, it will have to explain what these are in its privacy notice[74]. Larger organisations at least may need to have some form of ethics review board to make this assessment. This form of internal regulation is in line with a trend we have noted in business and government towards developing ethical approaches to big data. We discuss this further in the section on ethical approaches in chapter 3.

72. Given some of the difficulties associated with consent in a big data context, legitimate interests may provide an alternative basis for the processing, which allows for a balance between commercial and societal benefits and the rights and interests of individuals. For example, a paper by the Information Accountability Foundation[75] on a holistic governance model for big data gives examples of the different interests at play in IoT scenarios. It suggests that while consent is important for some

---

https://secure.edps.europa.eu/EDPSWEB/webdav/site/mySite/shared/Documents/Consultation/Opinions/2015/15-11-19_Big_Data_EN.pdf Accessed 22 April 2016

[74] GDPR Article 13(1)(d) and 14(2)(b)

[75] Cullen, Peter; Glasgow, Jennifer and Crosley, Stan. Introduction to the HGP framework. Information Accountability Foundation, 29 October 2015. http://informationaccountability.org/wp-content/uploads/HGP-Overview.pdf Accessed 22 April 2016

uses of the data, for others it may not be appropriate; in these cases the legitimate interests condition, with its inherent balancing test, may be an alternative.

Contracts

73. Condition 2 of Schedule 2 of the DPA is that the processing is necessary for the performance of a contract to which the data subject is a party. The GDPR also contains a similar provision for contracts[76]. This is relevant, for example, when someone makes a purchase online, and the website has to process their name, address and credit-card details to complete the purchase. Specific consent is not required for this. The problem in applying this in a big data context is that the processing must be "necessary". Big data analytics, by its nature, is likely to represent a level of analysis that goes beyond what is required simply to sell a product or deliver a service. It often takes the data that is generated by the basic provision of a service and repurposes it. So it may be difficult to show that the big data analytics are strictly necessary for the performance of a contract.

Public sector

74. Issues regarding conditions for processing may be different in the case of public authorities. Where administrative data is being used for big data analytics, it is unlikely to have been collected on the basis of consent or legitimate interests. Other conditions are available, for example that the processing is necessary for the exercise of functions conferred by law or functions of government departments or other public functions in the public interest[77]; these provisions are also reflected in the GDPR[78]. Furthermore, under the GDPR the legitimate interests condition will not be available to public authorities, since it will not apply to processing they carry out "in performance of their tasks".[79]

75. HMRC's Connect system[80] is an example of big data analytics in the public sector, based on statutory powers rather than consent. It is used to identify potential tax fraud by bringing

[76] GDPR Article 6(1) (b)
[77] DPA Schedule 2(5)
[78] GDPR Article 6(1)(e)
[79] GDPR Recital 47 and Article 6(1)
[80] BDO. HMRC's evolution into the digital age. Implications for taxpayers. BDO, March 2015. http://www.bdo.co.uk/__data/assets/pdf_file/0011/1350101/BDO_HMRC_DIGITAL_AGE.pdf Accessed 22 April 2016

together over a billion items of data from 30 sources, including self-assessment tax returns, PAYE, interest on bank accounts, benefits and tax credit data, the Land Registry, the DVLA, credit card sales, online marketplaces and social media.

76. In some cases the further use of data by the public sector may require consent. The Administrative Data Research Network makes large volumes of public-sector data available for research. It has systems in place to ensure that the data used for analysis is anonymised. The Task Force report[81] that led to this said that if administrative data is being linked to survey data supplied voluntarily by individuals, then consent would normally be required for the linkage even if the linked data is de-identified before analysis. This is another 'belt and braces' approach we would support in the interest of safeguarding the rights and freedoms of data subjects.

[81] Administrative Data Taskforce. The UK Administrative Data Research Network: improving access for research and policy. Administrative Data Taskforce, December 2012. https://www.statisticsauthority.gov.uk/wp-content/uploads/2015/12/images-administrativedatataskforcereportdecember201_tcm97-43887.pdf Accessed 22 April 2016

Purpose limitation

In brief…

- The purpose limitation principle does not necessarily create a barrier for big data analytics, but it means an **assessment of compatibility** of processing purposes must be done.
- **Fairness** is a key factor in determining whether big data analysis is incompatible with the original processing purpose.

77. The second data protection principle creates a two-part test: first, the purpose for which the data is collected must be specified and lawful (the GDPR adds 'explicit'[82]); and second, if the data is further processed for any other purpose, it must not be incompatible with the original purpose.

78. Some suggest[83] that big data challenges the principle of purpose limitation, and that the principle is a barrier to the development of big data analytics. This reflects a view of big data analytics as a fluid and serendipitous process, in which analysing data using many different algorithms reveals unexpected correlations that can lead to the data being used for new purposes. Some suggest that the purpose limitation principle restricts an organisation's freedom to make these discoveries and innovations. Purpose limitation prevents arbitrary re-use but it need not be an insuperable barrier to extracting the value from data. The issue is how to assess compatibility.

79. The Article 29 Working Party's Opinion on purpose limitation[84] under the current Directive says:

> "By providing that any further processing is authorised as long as it is not incompatible (and if the requirements of lawfulness are simultaneously also fulfilled), it would appear that the

[82] GDPR Article 5(1)(b)
[83] For example, in World Economic Forum Unlocking the value of personal data; from collection to usage. World Economic Forum, February 2013 http://www3.weforum.org/docs/WEF_IT_UnlockingValuePersonalData_CollectionUsage_Report_2013.pdf Accessed 20 June 2016
[84] Article 29 Data Protection Working Party. Opinion 03/2013 on purpose limitation. European Commission, 2 April 2013. http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf Accessed 1 June 2016

> legislators intended to give some flexibility with regard to further use. Such further use may fit closely with the initial purpose or be different. The fact that the further processing is for a different purpose does not necessarily mean that it is automatically incompatible: this needs to be assessed on a case-by-case basis …" (p. 21)

80. The Opinion sets out a detailed approach to assessing whether any further processing is for an incompatible purpose. It also addresses directly the issue of repurposing data for big data analytics. It identifies two types of further processing: first, where it is done to detect trends or correlations; and second, where it is done to find out about individuals and make decisions affecting them. In the first case, it advocates a clear functional separation between the analytics operations. In the second, it says that "free, specific, informed and unambiguous 'opt-in' consent would almost always be required, otherwise further use cannot be considered compatible"[85]. It also emphasises the need for transparency, and for allowing people to correct and update their profiles and to access their data in a portable, user-friendly and machine-readable format.

81. In our view, a key factor in deciding whether a new purpose is incompatible with the original purpose is whether it is fair. In particular, this means considering how the new purpose affects the privacy of the individuals concerned and whether it is within their reasonable expectations that their data could be used in this way. This is also reflected in the GDPR, which says that in assessing compatibility it is necessary to take account of any link between the original and the new processing, the reasonable expectations of the data subjects, the nature of the data, the consequences of the further processing and the existence of safeguards[86].

82. If, for example, information that people have put on social media is going to be used to assess their health risks or their credit worthiness, or to market certain products to them, then unless they are informed of this and asked to give their consent, it is unlikely to be fair or compatible. If the new purpose would be otherwise unexpected, and it involves making decisions about them as individuals, then in most cases the organisation concerned will need to seek specific consent, in addition to assessing whether the new purpose is incompatible with the original reason for processing the data.

[85] *Ibid* p.46
[86] GDPR Recital 50

83. If an organisation is buying personal data from elsewhere for big data analytics, it needs to practise due diligence. It will need to assess whether the new processing is incompatible with the original purpose for which the data was collected, as well as checking whether it needs to seek further consent or provide a new privacy notice.

Data minimisation: collection and retention

In brief…

- Big data analytics can result in the collection of personal data that is **excessive** for the processing purpose.
- Organisations may be encouraged to **retain** personal data for longer than necessary because big data applications are capable of analysing large volumes of data.

84. Data protection legislation embodies the concept of data minimisation – that is, organisations should minimise the amount of data they collect and process, and the length of time they keep the data. Principle 3 of the DPA says "personal data shall be adequate, relevant and not excessive in relation to the purpose or purposes for which they are processed", while principle 5 says "personal data processed for any purpose or purposes shall not be kept for longer than is necessary for that purpose or those purposes". The GDPR says personal data shall be "adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed."[87]

85. By contrast, big data analytics tends to involve collecting and analysing as much data as possible, and in many cases all the data points in a particular set, rather than a sample ("n=all"). The issue regarding data minimisation is not simply the amount of data being used, but whether it is necessary for the purposes of the processing, or excessive. This is not a hypothetical problem: in a study of businesses in the UK, France and Germany, 72% said they had gathered data they did not subsequently use[88]. Excessive data collection is a data protection issue, but it can also make it more difficult for businesses to locate and work on the data they actually need.

86. Big data is also about the variety of data sources used in the analysis. In terms of principle 3, this raises questions as to whether it is relevant. Big data analytics may discover unexpected correlations, for example between data about people's lifestyles and their credit worthiness, but that does not mean any information that can be

[87] GDPR Article 5(1)(c)
[88] Pure Storage. Big data's big failure. The struggles businesses face in accessing the information they need. Pure Storage, December 2015. http://info.purestorage.com/rs/225-USM-292/images/Big%20Data%27s%20Big%20Failure_UK%281%29.pdf?aliId=64921319 Accessed 25 April 2016

obtained about them is necessarily relevant to the purpose of assessing credit risk. Finding the correlation does not retrospectively justify obtaining the data in the first place.

87. The principle 5 requirement, that personal data shall not be kept longer than necessary for the purpose for which it is being processed, supports the data privacy of individuals and also reflects good practice in records management. However, in the world of big data this may be problematic for two reasons. First, the capacity to store data increases all the time and the cost of storing it is falling; in the words of the technology historian George Dyson, "big data is what happened when the cost of storing information became less than the cost of throwing it away."[89] Second, the ability of big data analytics to process huge volumes of data may encourage data controllers to keep long runs of historical data beyond the period required for normal business purposes.

88. The GDPR introduces the 'right to be forgotten'[90]. Data subjects will have the right for their data to be erased in several situations. This will apply, for example, where the data is no longer necessary for the purpose for which it was collected, or where it is processed on the basis of consent and the data subject withdraws that consent. This is particularly an issue for business, rather than the public sector, since the right to be forgotten does not apply if the processing is necessary for a legal obligation or for the exercise of official authority. It may be practically difficult for a business to find and erase someone's data if it is stored across several different systems[91].

89. Organisations therefore need to be able to articulate at the outset why they need to collect and process particular datasets. They need to be clear about what they expect to learn or be able to do by processing that data, and thus satisfy themselves that the data is relevant and not excessive, in relation to that aim. The challenge is to define the purposes of the processing and establish what data will be relevant.

90. Big data organisations also need to adopt good information governance and in particular enforce appropriate retention schedules.

---

[89] Warner, Andrew. George Dyson seminar media. The Long Now Foundation, 28 March 2013. http://blog.longnow.org/02013/03/28/george-dyson-seminar-media/ Accessed 25 April 2016

[90] GDPR Article 17

[91] Becoming an analytics-driven organisation to create value. Ernst Young, January 2015. http://www.ey.com/Publication/vwLUAssetsPI/Becoming_an_analytics_driven_organisation_to_create_value/$FILE/ey%20big%20data%20report_low-res.pdf Accessed 27 April 2016

Retention periods may be specified in general records management and accounting standards, and in some sectors there are regulatory requirements as to how long records should be kept. In some research contexts, such as clinical trials, long-term studies are important but in a commercial context businesses are more likely to be interested in analysing the most current data rather than historical records.

91. We understand that following these principles may be harder in a big data context, but we would argue that having well-managed, up-to-date and relevant data – and not acquiring and keeping data just in case it may be useful – helps to improve data quality and assists the analytics[92].

[92] This view is supported by the IBM white paper: Information lifecycle governance in a big data environment. IBM, January 2015. http://public.dhe.ibm.com/common/ssi/ecm/wv/en/wvw12356usen/WVW12356USEN.PDF Accessed 27 April 2016

Accuracy

In brief...

- There are implications regarding the **accuracy** of personal data at all stages of a big data project: collection, analysis and application.
- Results of data analysis may not be **representative** of the population as a whole.
- **Hidden biases** in datasets can lead to inaccurate predictions about individuals.

92. The fourth principle of the DPA requires that personal data is accurate and, where necessary, kept up to date. This is obviously good practice in terms of information management but it is also linked to the rights of the individual: people have a right to have inaccurate data corrected. By contrast, it has been suggested that big data analytics can tolerate a certain amount of 'messy' (ie, inaccurate) data, because the volumes of data being processed are generally so large[93]. A certain level of 'messiness', such as an incorrect name or address, may not be a problem when the analytics is used to detect general trends. But it is much more likely to be problematic when the processing is used to profile particular individuals.

93. It may be thought that being able to analyse vast amounts of data to identify correlations will inevitably lead to better analysis of trends and more accurate predictions. However, even if the data itself is recorded accurately (eg, the content of a tweet or the location of a mobile phone), this does not necessarily mean the conclusions that can be drawn from it are accurate.

94. Using all the available data (n=all) is a feature of big data analytics, but the 'all' may itself exclude or under-represent certain groups. There is growing interest in using data from social media to study behaviour during natural disasters. For example, an analysis was made of tweets and postings on Foursquare in New York as indicators of people's behaviour around the time of Hurricane Sandy in 2012. However, most of the social-media messages originated from Manhattan rather than the most storm-affected areas, where power

[93] Mayer-Schönberger, Viktor and Cukier, Kenneth. Big data. A revolution that will transform how we live, work and think. John Murray, 2013

blackouts limited mobile-phone access[94]. This starkly illustrates the obvious fact that analysis based on social media is not necessarily representative of the whole population. Even if there are no such obvious gaps and the data collected does represent a particular population, there may be a discrepancy between how people express themselves on social media and how they behave in the real world.

95. Issues of how accurately big data represents the population as a whole are not limited to social-media sources. The City of Boston makes available a Street Bump app for mobile phones. On a car journey, the app uses the phone's accelerometer and GPS positioning to record unusual movements due to problems with the road such as potholes, and transmits the data to the council for analysis. The council is aware that varying levels of mobile-phone ownership among different socioeconomic groups means more data may be collected from more affluent areas, rather than those with the worst roads[95].

96. Machine learning itself may contain hidden bias. A common phrase used in the discussion of machine learning is "garbage in garbage out"[96] Essentially, if the input data contains errors and inaccuracies, so will the output data. While supervised machine learning in particular often involves a pre-processing stage to improve the quality of the input data[97], the human-labelling of a training dataset can create a further opportunity for inaccuracies or bias to creep in. Hypothetically, a predictive model used in recruitment may achieve an overall accuracy rate of 90%, but this may be because it is 100% accurate for a majority population who make up 90% of applicants but wholly inaccurate for minority groups who make up the other 10%. It would be necessary to test for this and build in corrective measures[98].

97. In the examples given here, even when the raw data used in the analysis is recorded accurately, there may be issues as to how representative the dataset is and whether the analytics contain hidden bias. This is potentially a data protection issue if, in the

[94] Crawford, Kate. The hidden biases in big data. Harvard Business Review, 1 April 2013. https://hbr.org/2013/04/the-hidden-biases-in-big-data Accessed 20 May 2016
[95] Crawford, Kate. The hidden biases in big data. Harvard Business Review, 1 April 2013. https://hbr.org/2013/04/the-hidden-biases-in-big-data Accessed 20 May 2016
[96] Marinov, Svetoslav. How to get the most out of machine learning systems. ITProPortal, 18 June 2016. http://www.itproportal.com/2016/06/18/how-to-get-the-most-out-of-machine-learning-systems/ Accessed 13 December 2016
[97] Kotsiantis, S. B.; Kanellopoulos, D and Pintelas, P. E. Data preprocessing for supervised leaning. International Journal of Computer Science 1, no. 2 (2006): 111-117.
[98] Ajunwa, Ifeoma et al. Hiring by algorithm; predicting and preventing disparate impact. SSRN, 10 March 2016. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2746078 Accessed 17 June 2016

application phase (when acting with data), the results of the analytics are used to profile individuals. Profiling people in this way involves creating derived or inferred data about them and, given the issues highlighted here, this data may be inaccurate and lead to incorrect predictions about their behaviour or their health, creditworthiness or insurance risk. This also raises questions about the general fairness of the processing.

Rights of individuals

In brief…

- The vast quantities of data used in big data analytics may make it more difficult for organisations to comply with the **right of access** to personal data.
- Organisations will need to have appropriate processes in place to deal with the GDPR's extension of rights regarding **decisions based on automated processing**.

Subject access

98. The DPA gives individuals powerful rights to access their personal data and these rights still apply in the world of big data. People have the right to be told what personal data about them is being processed, the purposes for which it is being processed and who it may be disclosed to. They have the right to receive a copy of the information that constitutes their personal data, as well as information about its sources. This right is only limited by exemptions set out in the DPA itself and in associated secondary legislation. Under the GDPR, the data controller will have to provide more information[99]; for example, they have to say how long they will keep the personal data, or at least explain their criteria for determining this. Furthermore, if the person makes their request electronically, the data controller must provide the information "in a commonly used electronic form" unless the requester specifies otherwise.

99. The volume and variety of big data and the complexity of the analytics could make it more difficult for organisations to meet this obligation. However, such reasons cannot be an excuse for disregarding legal obligations. The existence of the right of access compels organisations to practise good data management. They need adequate metadata, the ability to query their data to find all the information they have on an individual, and knowledge of whether the data they are processing has been truly anonymised or whether it can still be linked to an individual.

100. Historically, a common problem for organisations in dealing with subject access requests has been that the information is held in

[99] GDPR Article 15

different places. In our discussions with industry, it has been suggested that if an organisation's move to big data means that disparate data stores are brought together, this may make it easier to find all the data on an individual.

101. If an organisation is using and/or buying in a range of data sources, including unstructured data, it can be difficult to produce all the data on one individual. Moreover, the increasing use of observed, derived and inferred data means the data held may not all have been provided directly by the data subject. However, in other cases, the actual data held on an individual may not necessarily be extensive or hard to identify, even though the analytics applied to it are complex and extensive. For example, the data in question may be a record of phone calls which is also available to a customer on an itemised phone bill.

102. Some organisations already make this data available to their customers on request or proactively online, through a secure log-in. In the USA, the data broker Acxiom has a web portal that enables people to see the data held about them for marketing purposes and the sources of that data[100]. This is likely to become more common here, as the GDPR encourages this approach[101]:

> "Where possible, the controller should be able to provide remote access to a secure system which would provide the data subject with direct access to his or her personal data."

103. If personal data can be made available like this, it will help the organisation to meet its data protection obligations. It could also help to reassure people as to the amount and type of information being held about them.

Other rights

104. Under the DPA, individuals already have several other rights regarding the prevention of processing likely to cause damage or distress; the prevention of direct marketing[102]; the right not to be subject to purely automated decision making; and the

[100] https://aboutthedata.com
[101] GDPR Recital 63
[102] For guidance on direct marketing under the DPA and the Privacy and Electronic Communications Regulations, see: Information Commissioner's Office. Direct marketing. ICO, May 2016 https://ico.org.uk/media/for-organisations/documents/1555/direct-marketing-guidance.pdf Accessed 27 May 2016

rectification of inaccurate data. Organisations using big data need to ensure their systems can allow people to exercise these rights. The GDPR will extend these rights. In particular it deals specifically with profiling, defining it as:

> "…any form of automated processing of personal data consisting of using those data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements;"[103]

105. People have the right not to be subject to a decision based solely on automated processing, including profiling, if it "significantly affects" them, such as automated decisions made on online credit applications or e-recruitment[104]. This right is qualified in that it does not apply if the profiling is necessary for a contract between the data subject and the data controller, or is authorised by law or where people have given explicit consent. Nevertheless, given that a feature of big data is the ability to profile individuals and make decisions about them, by applying algorithms to large amounts of granular data, it is likely to significantly affect those individuals. This right may apply in those cases.

106. Linked with this is the right to an explanation of a decision based on automated processing. This is discussed in more detail in the section on accountability and governance below.

107. The GDPR also contains new rights to do with data portability, which are discussed in the section on personal data stores in chapter 3.

[103] GDPR Article 4
[104] GDPR Article 22 and recital 71

Security

In brief…

- There are several **information security risks** specific to big data analytics.
- Organisations need to recognise these new risks and put in place **appropriate security measures**.

108. Big data analytics can be used to improve information security. The ability to analyse huge volumes of data very quickly means it can be used to analyse network traffic, transactions and log files that are too big to handle with other technologies. The analytics can detect patterns and anomalies and rapidly identify security threats[105].

109. At the same time, questions have been raised as to whether big data itself creates new security risks. ENISA produces a regular report on the big data 'threat landscape'. The latest report[106] recognised that big data analytics can be a powerful tool in detecting security risks, but it also identified several potential security risks specific to big data processing. For example, the high level of replication in big data storage and the frequency of outsourcing the analytics increase the risk of breaches, data leakages and degradation; also, the creation of links between the different datasets could increase the impact of breaches and leakages. In a further study[107], ENISA looked at the security of big data in three sectors: financial services, energy and telecoms. ENISA identified threats to do with access controls, the ability to securely restore datasets, and validation of the data sources. ENISA proposed measures to mitigate these threats.

110. These reports clearly indicate that, in addition to the security issues associated with any IT system, specific security threats can arise

[105] Big Data Working Group Big data analytics for security intelligence. Cloud Security Alliance, September 2013 https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Big_Data_Analytics_for_Security_Intelligence.pdf Accessed 25 June 2014; Curry, Sam et al Big data fuels intelligence-driven security. RSA Security Brief. EMC, January 2013. http://www.emc.com/collateral/industry-overview/big-data-fuels-intelligence-driven-security-io.pdf Accessed 25 May 2016

[106] Damiani, Ernesto et al. Big data threat landscape and good practice guide. ENISA, January 2016. https://www.enisa.europa.eu/publications/bigdata-threat-landscape Accessed 27 May 2016

[107] Naydenov, Rossen et al. Big data security. Good practices and recommendations on the security of big data systems. ENISA, December 2015. https://www.enisa.europa.eu/publications/big-data-security Accessed 27 May 2016

from the nature of big data processing. These need to be addressed by data controllers as part of their risk assessment in order to meet the requirement, in both the DPA and the GDPR, to put in place appropriate security measures to protect personal data.

111. In some cases it will be more practical to carry out big data analytics in the cloud. In this case, organisations must obtain sufficient guarantees from the cloud provider as to the security measures it uses. Our guidance on the use of cloud computing[108] explains the data protection issues involved.

[108] Information Commissioner's Office. Guidance on the use of cloud computing. ICO, October 2012. https://ico.org.uk/media/for-organisations/documents/1540/cloud_computing_guidance_for_organisations.pdf Accessed 20 June 2016

Accountability and governance

In brief...

- Accountability is increasingly important for big data analytics and will become an explicit **requirement under the GDPR**.
- Big data organisations may need to make **changes** to their reporting structures, internal record keeping and resource allocation.
- Machine learning algorithms have the potential to make decisions that are **discriminatory**, **erroneous** and **unjustified**.
- **Data quality** is a key issue for those with information governance responsibilities in a big data context.

112. The concept of 'accountability' has always been an implicit requirement in the DPA. However, under the GDPR its importance is elevated by introducing an explicit accountability principle that requires organisations to demonstrate compliance with all the other principles in the regulation[109]. This is further emphasised by several provisions throughout the GDPR that promote accountability. The requirements of this new principle have several implications for organisations undertaking big data analytics.

113. One of the accountability requirements is that records of processing activities must be maintained in circumstances (among others) where organisations have more than 250 employees or where they are processing personal data that could result in a risk to individuals' rights and freedoms. It is likely that organisations involved in big data analytics may fall within one, or both, of the above situations. One of the records that must be maintained is the purposes of the processing of personal data[110]. This may be problematic in a big data context as the initial analysis of data is often experimental and without any predefined hypothesis or business need[111]. This unique feature of big data analytics (discussed in more detail in the privacy notices section of chapter 3) means the ultimate reasons for

[109] GDPR Article 5(2)
[110] GDPR Article 30(1)(b)
[111] Andreev, Alexei. The Death of the Hypothesis, or, Investing in Big Data Analytics and Deep Learning. Harris and Harris Group, 23 September 2014. http://www.hhvc.com/blog/death-hypothesis-investing-big-data-analytics-deep-learning/ Accessed 9 December 2016.

processing may be unclear; this may make it difficult to state the purposes as a part of internal record keeping. Furthermore, the purpose an organisation initially records may change as new correlations in the data are discovered which prompt different uses.

114. A further accountability provision under the GDPR is the requirement to appoint a data protection officer (DPO)[112]. This will be a necessity for almost all public authorities but also for organisations that systematically monitor individuals on a large scale, including those using big data analytics for purposes such as online behaviour tracking or profiling. Although this will mainly affect smaller organisations that may not currently have an appointed DPO, it also has implications for big data organisations that already have a DPO. This is because the GDPR places new obligations on organisations regarding the DPO's position and tasks, which may require changes in reporting structures and the provision of additional resources[113].

115. The importance of accountability is not limited to what is explicitly stated in the provisions of the GDPR; it extends to all aspects of an organisation's processing operations involving personal data. Given the growing popularity of some types of big data analytics, such as machine learning, there has recently been increasing discussion about the role of algorithms in decision making, often referred to as 'algorithmic accountability'[114]. In essence, this is about being able to check that the algorithms used and developed by machine learning systems are actually doing what we think they're doing and aren't producing discriminatory, erroneous or unjustified results.

116. As regards discrimination, this is associated with issues relating to the effects of the processing of personal data, as discussed in the fairness section above. The autonomous and opaque nature of machine learning algorithms can mean that decisions based on their output may only be identified as having been discriminatory afterwards – when the effects have already been felt by the people discriminated against. For instance, ProPublica analysed 7,000 'risk scores' produced by a machine learning tool used in some US states to predict the future criminal behaviour of defendants. The findings revealed discrimination based on race, with black defendants falsely classified as future criminals on nearly twice as many occasions as

[112] GDPR Article 37(1)
[113] GDPR Article 38
[114] Taneja, Hemant. The need for algorithmic accountability. TechCrunch, 8 September 2016. https://techcrunch.com/2016/09/08/the-need-for-algorithmic-accountability/ Accessed 12 December 2016

white defendants[115]. Detecting discriminatory decisions in hindsight will not be sufficient to comply with the accountability provisions of the GDPR[116]. Big data analysts will need to find ways to build discrimination detection into their machine learning systems to prevent such decisions being made in the first place.

117. As regards erroneous algorithmic decisions, there are clear implications for the data protection principle of accuracy, such as inaccurate predictions based on biased profiling, as discussed in the accuracy section above. However, it is not just profiling decisions that need to be held to account for their accuracy. Association algorithms find links between data that can then be applied to real-world situations. For instance, the autocomplete functionality of Google's search engine suggests words that its machine learning algorithms have determined are associated with the user-typed text[117]. This type of decision can be problematic regarding accuracy. Autocomplete associations have twice been contested in German courts. One case involved a businessman who discovered that Google linked him with 'scientology' and 'fraud'. Another involved the wife of a former German president who sued Google over autocomplete's suggestions of words that linked her with escort services[118]. In Nicholas Diakopoulos' article on accountability in algorithmic decision making, he makes the following point about associations:

> "One issue with the church of big data is its overriding faith in correlation as king. Correlations certainly do create statistical associations between data dimensions. But despite the popular adage, 'Correlation does not equal causation,' people often misinterpret correlational associations as causal."[119]

118. The distinction between correlation and causation is very important; organisations using machine learning algorithms to discover associations need to appropriately consider this distinction and the potential accuracy (or inaccuracy) of any resulting decisions.

---

[115] Angwin, Julia. Make Algorithms Accountable. The New York Times, 1 August 2016. http://www.nytimes.com/2016/08/01/opinion/make-algorithms-accountable.html?_r=1 Accessed 12 December 2016

[116] GDPR Recital 71

[117] Rangaswami, Shanta et al. Analysis of Optimized Association Rule Mining Algorithm using Genetic Algorithm. IJCA Proceedings on International Conference on Information and Communication Technologies ICICT(2):12-15, October 2014.

[118] BBC News. Germany tells Google to tidy up auto-complete. BBC, 14 May 2013. http://www.bbc.co.uk/news/technology-22529357 Accessed 12 December 2016.

[119] Diakopoulos, Nicholas. Accountability in algorithmic decision making. Communications of the ACM 59, no. 2 (2016): 56-62.

119. As regards the justification of decisions based on machine learning algorithms, and further to the issues of fairness discussed above, there are also implications for the new right under the GDPR to obtain an explanation of a decision based on automated processing[120]. Some have suggested this right can be easily circumvented as it is restricted to 'solely' automated processing and decisions that 'significantly' affect individuals[121]. However, there are still many situations where the right is very likely to apply, such as credit applications, recruitment and insurance. In such circumstances, it may be difficult to provide a meaningful response to an individual exercising their right to an explanation. This is because, as Jenna Burrell notes in her paper on the opacity of machine learning algorithms, when computers learn and make decisions, they do so "without regard for human comprehension"[122]. Big data organisations therefore need to exercise caution before relying on machine learning decisions that cannot be rationalised in human-understandable terms. If an insurance company cannot work out the intricate nuances that cause their online application system to turn some people away but accept others (however reasonable those underlying reasons may be), how can it hope to explain this to the individuals affected?

120. Related to the concept of accountability are data quality and information governance. Data quality is a key issue for organisations using big data analytics. This is linked to what is often seen as the 'fourth V' of big data: veracity, or in other words the reliability of the data[123]. Senior managers in big data organisations need to know whether they can trust what the data is apparently telling them. This can involve looking at, for example, the sources of the data, how accurate it is, whether is sufficiently up to date, how securely it is kept and whether there are restrictions on how it can be used. In 2013 Forrester Consulting[124] carried out a survey of senior executives in companies dealing with big data, asking them: "What best describes how you govern big data today?" The top issues they

[120] GDPR Recital 71
[121] Jaakonsaari, Liisa. Who sets the agenda on algorithmic accountability? EurActiv, 26 October 2016. https://www.euractiv.com/section/digital/opinion/who-sets-the-agenda-on-algorithmic-accountability/ Accessed 12 December 2016
[122] Burrell, Jenna. How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data & Society 3, no. 1 (2016)
[123] For example, IBM Institute for Business Value. Analytics: the real world use of big data. IBM, October 2012. https://www.ibm.com/smarterplanet/global/files/se__sv_se__intelligence__Analytics_-_The_real-world_use_of_big_data.pdf Accessed 15 June 2016
[124] Forrester Consulting. Big data needs agile information and integration governance. Forrester Research Inc, August 2013. http://www.ibmbigdatahub.com/whitepaper/big-data-needs-agile-information-and-integration-governance Accessed 15 June 2016

identified included the following (in each case we have also identified relevant data protection provisions):

Information governance issue	Data protection provision
Security and monitoring	Principle 7 – security
Protection and masking of sensitive data	Sensitive data definition and conditions for processing
Profiling data sources (lineage, traceability, format, etc.)	Anonymisation and definition of personal data; Principle 1 – fairness
Data lifecycle management: archiving data	Principle 5 – retention

121. It is clear that these information management issues all have implications for data protection. This means data protection cannot be seen as simply a matter of complying with external legal requirements. Addressing data protection issues supports good practice in information governance. The Forrester study found that mature information governance is linked to business success, because it is essential to realising the benefits of big data. Data protection should be seen as an enabler of that success, not a barrier to it.

122. The growth of big data has led some organisations to create a position of chief data officer[125]. The role typically includes overall responsibility for data quality and data security. We suggest this also involves a consideration of data protection, which may traditionally have been seen as a compliance function.

[125] Terlink, Marc et al. The new hero of big data and analytics: The Chief Data Officer. IBM, June 2014. http://www-935.ibm.com/services/us/gbs/thoughtleadership/chiefdataofficer/ Accessed 15 June 2016

Data controllers and data processors

In brief...

- Big data analytics can make it **difficult to distinguish** between **data controllers** and **data processors**.
- Organisations **outsourcing** analytics to companies specialising in AI and machine learning need to consider carefully who has **control** over the processing of any personal data.

123. Under the DPA, a distinction is made is between 'data controllers' and 'data processors'[126]. A data controller, as the name suggests, has overall control over the processing of personal data – it determines the purposes and manner of the processing, either on its own, jointly or in common with another data controller. However, a data processor does not have such control – it merely processes personal data on a data controller's behalf. When a data controller uses a data processor, the DPA says a written contact must be in place to ensure that the data processor only acts upon its instructions; ultimate responsibility for compliance with the DPA lies with the data controller. Our guidance on the differences between, and governance implications for, data controllers and data processors goes into more detail[127].

124. The definitions of data controllers and data processors are very similar under the GDPR. However, data processors will share more of the responsibility for compliance with the regulation, specifically regarding adoption of appropriate security measures and the possibility of regulatory action for failure to comply. The GDPR also sets out certain details that contracts between data controllers and data processors must contain, such as types of data; purpose (and duration of) processing; and categories of data subject[128].

125. Distinguishing between data controllers and data processors can be relatively straightforward in certain circumstances. For instance, if an organisation chooses to store its customer data in the cloud, the cloud provider is likely to be a data processor as it is simply acting on

[126] Data Protection Act 1998 Part I Section 1(1)
[127] Information Commissioner's Office. Data controllers and data processors: what the difference is and what the governance implications are. ICO, May 2014. https://ico.org.uk/media/for-organisations/documents/1546/data-controllers-and-data-processors-dp-guidance.pdf Accessed 16 February 2017
[128] GDPR Article 28(3)

the original organisation's behalf and is not determining the purposes of the processing.

126. However, when personal data is processed in the context of big data, AI and machine learning, this can make it more difficult to distinguish between data controllers and data processors. This is because, typically, big data analytics is about finding correlations, making predictions and aiding decision-making; all of which blur the lines between who is actually determining the purposes and manner of the processing when an organisation has chosen to outsource the analytics to another company – one that specialises in AI, for example.

127. Therefore, when outsourcing big data analytics to other companies, careful consideration should be given to where control over the processing of personal data actually lies – this will have implications for compliance and liability. If an organisation intends to conduct its big data outsourcing in a data controller-data processor relationship, it is important that the contract includes clear instructions about how the data can be used and the specific purposes for its processing.

128. However, the existence of such a contract would not automatically mean that the company doing the data analysis is a data processor. If that company has enough freedom to use its expertise to decide what data to collect and how to apply its analytic techniques, it is likely to be a data controller as well. For instance, in a forthcoming article on the transfer of data from the Royal Free London NHS Foundation Trust to Google DeepMind, Julia Powles argues that, despite assertions to the contrary, DeepMind is actually a joint data controller as opposed to a data processor[129].

[129] Powles, Julia. Google DeepMind and healthcare in an age of algorithms. Forthcoming in Journal of Health and Technology

Chapter 3 – Compliance tools

129. The previous chapter discussed several key data protection implications that can arise from the use of big data analytics. We now turn to some of the compliance tools that can help organisations meet their data protection obligations and protect people's privacy rights in a big data context.

Anonymisation

In brief…

- Often, big data analytics will not require the use of data that identifies individuals.
- **Anonymisation** can be a successful tool that takes processing out of the data protection sphere and mitigates the risk of loss of personal data.
- Organisations using anonymisation techniques need to make robust assessments of the **risk of re-identification**.

130. If personal data can be fully anonymised, it is no longer personal data. In this context, 'anonymised' means it is not possible to identify an individual from the data itself or from that data in combination with other data, taking account of all the means that are reasonably likely to be used to identify them. If the data is no longer personal data, it is not covered by data protection legislation. The GDPR makes this explicit[130]:

> "The principles of data protection should therefore not apply to anonymous information, namely information that does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable."

131. Therefore, a key question for big data organisations is whether they need to use data that identifies individuals. There are many examples of the use of anonymised data in big data analytics. Telefonica's

[130] GDPR Recital 26

Smart Steps[131] tool uses data on the location of mobile phones on its network to track the movement of crowds of people. This can be used by retailers to analyse footfall in a particular location. The data that identifies individuals is stripped out before the analysis and the anonymised data is aggregated to gain insights about the population as a whole and combined with market research data from other sources. Another example of this is in medical research, where data from clinical trials is rigorously anonymised before being made available for analysis.

132. In practice, anonymised data may be used in several scenarios: organisations may bring in anonymised data, or they may seek to irreversibly anonymise their own data before using it themselves or sharing it with others.

133. Some commentators have pointed to examples of where it has apparently been possible to identify individuals in anonymised datasets, and so concluded that anonymisation is becoming increasingly ineffective in the world of big data[132]. On the other hand, Cavoukian and Castro[133] have found shortcomings in the main studies on which this view is based. A recent MIT study looked at records of three months of credit card transactions for 1.1 million people and claimed that, using the dates and locations of four purchases, it was possible to identify 90 percent of the people in the dataset. However, Khalid El Emam has pointed out[134] that, while the researchers were able to identify unique patterns of spending, they did not actually identify any individuals. He also suggested that in practice access to a dataset such as this would be controlled and also that the anonymisation techniques applied to the dataset were not particularly sophisticated and could have been improved.

134. It may not be possible to establish with absolute certainty that an individual cannot be identified from a particular dataset, taken together with other data that may exist elsewhere. The issue is not

---

[131] Telefonica. Smart Steps http://dynamicinsights.telefonica.com/488/smart-steps Accessed 20 June 2016

[132] For example, President's Council of Advisors on Science and Technology. Big data and privacy. A technological perspective. White House, May 2014 http://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf Accessed 20 June 2016

[133] Cavoukian, Anne and Castro, Daniel. Big data and innovation, setting the record straight: de-identification does work. Office of the Information and Privacy Commissioner, Ontario, June 2014. https://www.ipc.on.ca/English/Resources/Discussion-Papers/Discussion-Papers-Summary/?id=1413 Accessed 1 June 2016.

[134] El Emam, Khaled. Is it safe to anonymise data? BMJ, February 2015. http://blogs.bmj.com/bmj/2015/02/06/khaled-el-emam-is-it-safe-to-anonymize-data/ Accessed 1 June 2016

about eliminating the risk of re-identification altogether, but whether it can be mitigated so it is no longer significant. Organisations should focus on mitigating the risks to the point where the chance of re-identification is extremely remote. The range of datasets available and the power of big data analytics make this more difficult, and the risk should not be underestimated. But that does not make anonymisation impossible or ineffective.

135. Organisations using anonymised data need to be able to show they have robustly assessed the risk of re-identification, and have adopted solutions proportionate to the risk. This may involve a range of technical measures, such as data masking, pseudonymisation and aggregation, as well as legal and organisational safeguards. The ICO's Anonymisation code of practice[135] explains these in more detail.

136. Anonymisation may be used when data is shared externally or within an organisation. For example, an organisation may hold a dataset containing personal data in one data store, and produce an anonymised version of it to be used for analytics in a separate area. Whether it remains personal data will depend on whether the anonymisation 'keys' and other relevant data that enable identification are kept by the organisation. Even if the data remains personal data, this is still a relevant safeguard to consider so that processing can comply with the data protection principles.

137. The Administrative Data Research Network is an example of how de-identified data can be used for research[136]. It is made up of four Administrative Data Research Centres (ADRCs), which provide a secure environment in which accredited researchers, working on approved projects, can access administrative data collected by government. The ADRCs do not hold personal identifiers with the data; instead, the personal identifiers are sent to trusted third parties. These third parties match records for the same individual across different datasets, and send the results of the matching to the ADRC (but they do not hold the research data about the individuals). This means the researchers see the data they need, including matchings across different datasets, but without any personal identifiers.

[135] Information Commissioner's Office. Anonymisation: managing data protection risk code of practice. ICO, November 2012. http://ico.org.uk/for_organisations/data_protection/topic_guides/~/media/documents/library/Data_Protection/Practical_application/anonymisation-codev2.pdf Accessed 25 June 2014

[136] Administrative Data Research Network https://adrn.ac.uk/protecting-privacy/ Accessed 1 June 2016

138. The UK Anonymisation Network (UKAN)[137], originally funded by the ICO, also has an important role in providing expert advice on anonymisation techniques. It is co-ordinated by a consortium of the Universities of Manchester and Southampton, the Open Data Institute and the ONS.

139. Anonymisation should not be seen merely as a means of reducing a regulatory burden by taking the processing outside the DPA. It is also a means of mitigating the risk of inadvertent disclosure or loss of personal data, so it is a tool that assists big data analytics and helps organisations to do research or develop products and services. It also enables organisations to provide an assurance to individuals that data identifying them is not being used for such analytics. This is part of the process of building trust, which is key to taking big data forward.

[137] UK Anonymisation Network website http://ukanon.net/ Accessed 20 June 2016

Privacy notices

In brief...

- There are several **innovative approaches to providing privacy notices** including the use of videos, cartoons, just-in-time notifications and standardised icons.
- Using a **combination of approaches** can help make complex information on big data analytics easier to understand.

140. The DPA says that, apart from in certain specified circumstances, the processing of personal data cannot be considered fair unless the data subject is given some basic information about it, including the data controller's identity, the purpose of the processing and any other information necessary for the processing to be fair. Our Privacy notices code of practice[138] explains this further with practical examples.

141. The GDPR expands on this and requires data controllers to provide more detailed information. This includes "the existence of automated decision-making including profiling"[139], in which case data controllers have to explain the logic involved and the "significance and envisaged consequences" of the profiling for the data subject. This refers to decisions based solely on automated processing, rather than on decisions made by people. This may be relevant in big data scenarios, where algorithms developed in the initial discovery phase are then applied to make decisions in individual cases, for example in credit scoring. In that case data controllers will have to find ways to describe, in meaningful terms, how the decision was made.

142. In a big data context these requirements can be problematic, and it has been suggested that privacy notices are not feasible regarding big data analytics. This is argued on several grounds:

    - People are unwilling to read lengthy privacy notices.
    - The context in which data is collected (eg from smartphone apps or IoT devices) can make it practically difficult to give the

[138] Information Commissioner's Office. Privacy notices code of practice. ICO, December 2010. http://ico.org.uk/for_organisations/data_protection/topic_guides/~/media/documents/library/Data_Protection/Detailed_specialist_guides/PRIVACY_NOTICES_COP_FINAL.ashx Accessed 8 April 2016

[139] GDPR Article 13(2)(f)

information.

- The analytics used in big data are too difficult to explain in terms that people can understand.

- Given that big data analytics often involves repurposing data, the data controller cannot foresee, at the outset, all the uses that may be made of the data.

People don't read privacy notices?

143. Ticking a box to say 'I have read and agreed to the terms and conditions' has been described as "the biggest lie on the web"[140]. Clearly, when people want to buy something online or download an app, they often tick 'I agree' without reading the privacy notice. Reluctance to read privacy notices has been well documented:

    - The House of Commons Science and Technology Committee heard evidence that to read all the terms and conditions encountered on the internet would take a month per year[141].

    - The White House report on big data[142] noted the phenomenon of 'privacy fatigue', and found that even though US advertisers provided information on their use of data, few people read or understood it.

    - A report for Ofcom by WIK Consult says there is little incentive for people to read privacy policies when using the internet since it would take significantly more time than they spend using the content or the app[143].

144. However, it would be wrong to infer that people are indifferent to how their data is used or that the requirement to provide privacy information is therefore irrelevant or inapplicable in a big

[140] http://biggestlie.com/
[141] House of Commons Science and Technology Committee. Responsible use of data. HC245. Stationery Office Ltd, November 2014. http://www.publications.parliament.uk/pa/cm201415/cmselect/cmsctech/245/245.pdf Accessed 17 June 2016
[142] Executive Office of the President. Big data: seizing opportunities, preserving values. White House, 1 May 2014. https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_5.1.14_final_print.pdf Accessed 11 April 2016
[143] Arnold, Rene; Hillebrand, Annette and Waldburger, Martin. Personal data and privacy. WIK Consult, 26 May 2015. http://stakeholders.ofcom.org.uk/binaries/internet/personal-data-and-privacy/Personal_Data_and_Privacy.pdf Accessed 11 April 2016

data context. As we have discussed in the section on expectations in chapter 2, people's attitudes to the use of their data are more complex than that simplistic approach would suggest. Rather, as the studies referred to in the previous paragraph suggest, the problem lies with the format of many privacy notices. It is understandable that people are unwilling to read privacy notices that are long, written in legalistic terms or intended primarily to protect the organisation using the data rather than informing the data subject.

145. In our view, rather than making privacy notices irrelevant, big data challenges organisations to be as innovative in this area as they are in using analytics, and to find new ways of conveying information concisely. Privacy notices should be written in plain English, with a person of an average reading age in mind. In this context it is important to recognise that the data subjects concerned may well be young people. Textual information can be accompanied by other ways of delivering information in a user-friendly form. Channel 4's use of a YouTube video[144] to accompany their 'Viewer Promise' is one example of an innovative approach. The Guardian[145] and O2[146] use cartoons to explain their privacy policies. A combination of different approaches can be used to make the information easier to understand.

146. In 2014 the House of Commons Science and Technology Committee called on the government to take the lead in encouraging clarity and simplicity in privacy notices[147]. Two years later, they noted that "a combination of consumer demand, reputational concerns and legislative pressure are beginning to have effect" and gave examples of improved practice by Google and Facebook[148]. A report for the British Standards Institution (BSI) said clearer and simpler privacy information would benefit consumers and businesses. The BSI is

[144] Channel 4 website http://www.channel4.com/4viewers/ Accessed 8 April 2016
[145] The Guardian website https://www.theguardian.com/info/privacy Accessed 17 June 2016
[146] O2 website http://www.o2.co.uk/termsandconditions/privacy-policy Accessed 8 April 2016
[147] House of Commons Science and Technology Committee. Responsible use of data. HC245. Stationery Office Ltd, November 2014. http://www.publications.parliament.uk/pa/cm201415/cmselect/cmsctech/245/245.pdf Accessed 17 June 2016
[148] House of Commons Science and Technology Committee. The big data dilemma. Fourth report of session 2015-16 HC468. The Stationery Office Ltd, February 2016. http://www.publications.parliament.uk/pa/cm201516/cmselect/cmsctech/468/468.pdf Accessed 17 June 2016

planning to develop a set of standards specifically for big data, including a Terms and Conditions standard[149].

147. Several organisations in the UK and abroad are currently working on practical ways to make privacy notices more understandable[150]. As well as advocating the use of plain language, these approaches include identifying and defining commonly used terms and creating a database so that they can be re-used in different contexts, with standard icons. It has been suggested that since nutrition information is conveyed in standardised ways on food packaging, a similar approach could be applied to privacy notices.

148. The GDPR encourages clarity and new approaches. It says that information addressed to the public or the data subject should be "concise, easily accessible and easy to understand, and that clear and plain language, and additionally, where appropriate, visualisation is used". It says that this is particularly relevant in situations such as online advertising, where "the proliferation of actors and the technological complexity of practice make it difficult for the data subject to know and understand if personal data relating to him or her are being collected, by whom and for what purpose."[151]. It also says standardised icons can be used to explain the processing[152].

New methods of data collection

149. When big data includes data that is collected or observed by apps or devices rather than provided directly by individuals, it can be more challenging to provide a privacy notice, but it is possible to address this issue, as the following examples show:

    - Our guidance on Privacy in mobile apps[153] gives examples of a notice within an app store and of in-app notices, including a layered privacy notice, where brief summary information is given along with a link to a more detailed

[149] Circle Research. Big data and standards market research. BSI Standards Ltd, January 2016. http://shop.bsigroup.com/forms/The-Big-Data-and-market-research-report/ Accessed 11 April 2016
[150] For example, the Biggest Lie http://biggestlie.com/; Common Terms http://commonterms.org/; the Meaningful Consent Project http://www.meaningfulconsent.org/
[151] GDPR Recital 58
[152] GDPR Recital 60 and Article 12(7)
[153] Information Commissioner's Office. Privacy in mobile apps. ICO, December 2013. https://ico.org.uk/media/for-organisations/documents/1596/privacy-in-mobile-apps-dp-guidance.pdf Accessed 20 June 2016

policy, and a just-in-time notification at the point where data is collected.

- The EU's Article 29 Working Party looked at data protection issues associated with IoT devices, focusing on wearable computing such as watches and glasses, 'quantified self' devices such as activity trackers, and devices in the home such as smart thermostats. They suggested that privacy information could be provided on the device itself, or by broadcasting the information via Wi-Fi or making it available through a QR code[154].

- We have published guidance on Wi-Fi location analytics[155] (ie, the collection of MAC addresses of devices trying to connect to Wi-Fi networks, such as those provided by businesses for customers to use on their premises). This recommends giving privacy information through signage in the building, in the location(s) where the data is processed, or at the portal page of the Wi-Fi network.

150. We recognise that in these contexts it may be difficult to provide the information required by the DPA and the GDPR, and to do so in way that encourages people to read it. But the advice in these documents (and our Privacy notices code of practice) suggests it is possible. It is important to consider at an early stage of development how this information will be provided, and to look at the relationship between usability and privacy by design[156].

Big data is too hard to explain?

151. It may be argued that it is too difficult to explain the processing in a privacy notice, because big data tends to rely on complex analytics and algorithms. However, this would be to misunderstand the purpose of a privacy notice. The DPA does not require the privacy notice to describe *how* the data is processed (ie, the technical details of how the algorithms work), but the *purposes* for which it is processed. The DPA is also clear

[154] Article 29 Data Protection Working Party. Opinion 8/2014 on the on Recent Developments on the Internet of Things. European Commission, September 2014. http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp223_en.pdf Accessed 20 June 2016

[155] Information Commissioner's Office. Wi-Fi location analytics. ICO, February 2016. https://ico.org.uk/media/for-organisations/documents/1560691/wi-fi-location-analytics-guidance.pdf Accessed 20 June 2016

[156] Information and Privacy Commissioner of Ontario. Privacy by Design and User Interfaces. http://www.ipc.on.ca/images/Resources/pbd-user-interfaces_Yahoo.pdf Accessed 20 June 2016

that processing cannot be fair if people are deceived or misled about those purposes[157]. Even if it is difficult to explain in simple terms how the analytics works, it should be possible to explain the purposes in a way that does not deceive or mislead.

Unforeseen purposes

152. Given the propensity of big data to find unexpected correlations between datasets and therefore to find new uses for that data, it may be difficult for an organisation to foresee at the outset all the uses it may make of the data it collects. Big data may involve a 'discovery' phase (thinking with data) in which the organisation analyses the data to find useful correlations. This implies a 'bottom up' approach, starting with the data rather than with a specific business need. This was the recommendation in a consultants' report on big data for the telecoms industry:

> "… operators seeking to make initial inroads with big data are advised to avoid the usual top-down approach, which sets up a business problem to be solved and then seeks out the data that might solve it…
>
> Instead, operators should begin with the data itself, experimenting with what they have on hand to see what kinds of connections and correlations it reveals."[158]

153. Such an approach may be a feature of big data but does not remove the requirement to provide a privacy notice if personal data is being processed. Big data organisations must identify the purposes of the processing at the earliest possible stage and communicate this to data subjects, so that people are not misled about how their data is used. Alternatively, if individuals need not be identified in the initial discovery phase, the organisation should consider using anonymised data instead.

154. The ability to analyse data for different purposes, such as using the location of mobile phones to plot movements of people or traffic is an important characteristic – and a benefit – of big data analytics. If an organisation has collected personal data for one

[157] Data Protection Act 1998 Schedule I Part II 1(1)
[158] Acker, Olaf; Blockus, Adrian and Pötscher, Florian. Benefiting from big data: a new approach for the telecom industry. Strategy&, 12 April 2013 http://www.strategyand.pwc.com/reports/benefiting-big-data Accessed 14 April 2016

purpose and wants to start using it for a completely different purpose, it needs to update its privacy notice accordingly, ensure that people are aware of this, and obtain consent to use the data for the new purpose.

155. If an organisation buys in personal data from another organisation to use it for big data analytics, it also needs to ensure that the original privacy notice given to the individuals by the seller covers this further use of the data. If it does not, the buyer will need to give the individuals concerned its own privacy notice, making clear the new purpose for which the data will be processed. In limited circumstances this is not needed, chiefly if to do so would involve a "disproportionate effort"[159]. Furthermore, if there is a difference between what people were told originally and what the buyer intends to do with the data, then the buyer will have to obtain their consent to their data being used for the new purpose.

156. Additionally, privacy notices are also an important tool in situations where an organisation is merged with or acquired by another organisation. This is especially relevant in the technology sector, where mergers and acquisitions often occur[160], notably Facebook's acquisition of WhatsApp[161] and Microsoft's acquisition of LinkedIn[162]. In such circumstances, the data protection obligations follow the data. So organisations need to ensure that individuals are made aware of what is happening and are reassured their personal data will only be used in line with their reasonable expectations (i.e. for the same purposes for which the data was originally obtained). Providing a copy of the original privacy notice, along with further information to identify the new organisation and explain what is happening, can help meet this requirement. The ICO's Data sharing code of practice[163] says more about mergers.

---

[159] DPA Schedule 1 Part II paragraph 3 and GDPR Article 14(5)
[160] Winkler, Rolfe and Steele, Anne. Technology Busiest Sector in Merger Deals This Year. The Wall Street Journal, 13 June 2016. http://www.wsj.com/articles/technology-busiest-sector-in-merger-deals-this-year-1465861867 Accessed 20 December 2016
[161] Olson, Parmy. Facebook closes $19 Billion WhatsApp Deal. Forbes, 6 October 2014. http://www.forbes.com/sites/parmyolson/2014/10/06/facebook-closes-19-billion-whatsapp-deal/#5bf1b7cf179e Accessed 20 December 2016
[162] Lunden, Ingrid. Microsoft officially closes its $26.2B acquisitions of LinkedIn. TechCrunch, 8 December 2016. https://techcrunch.com/2016/12/08/microsoft-officially-closes-its-26-2b-acquisition-of-linkedin/ Accessed 20 December 2016
[163] Information Commissioner's Office. Data sharing code of practice. ICO, May 2011. https://ico.org.uk/media/1068/data_sharing_code_of_practice.pdf Accessed 20 December 2016

157. The use of social-media data is a growing area in which data collected by one organisation is used by another. Social media platforms such as Facebook and Twitter make the data that subscribers have posted available to third parties under certain conditions. A social-media company may make the data available via an application protocol interface (API), as with Twitter, or in some cases the third party may gather the information themselves by 'web scraping'[164]. The data might be used for sentiment analysis or identifying general trends, or in some cases for profiling individuals, eg for assessing credit risk. It may be difficult to anonymise this data when it is transferred, so the third party may be processing personal data. In that case it should consider whether it is necessary to provide a privacy notice to the individuals concerned. While the social-media company's terms of service may include reference to use by third parties, in reality people may not be aware of how their data is used. Research by Ipsos Mori showed that fewer than two in five adults were aware that their social-media data could be shared with companies or government for research, and more than three in five thought it should not happen[165].

[164] Oxford Internet Institute. The use of social media for research and analysis: a feasibility study. Department for Work and Pensions, December 2014. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/387591/use-of-social-media-for-research-and-analysis.pdf Accessed 15 April 2016

[165] Evans, Harry; Ginnis, Steve and Bartlett, Jamie. #SocialEthics. A guide to embedding ethics in social media research. Ipsos MORI, December 2015. https://www.ipsos-mori.com/Assets/Docs/Publications/im-demos-social-ethics-in-social-media-research-summary.pdf Accessed 15 April 2016

Privacy impact assessments

In brief…

- A **privacy impact assessment** is an important tool that can help to **identify and mitigate privacy risks** before the processing of personal data.
- Under the **GDPR**, it is highly likely that doing a privacy impact assessment – known as a **'data protection impact assessment'** – will be a **requirement** for big data analytics involving the processing of personal data.
- The unique features of big data analytics can make some steps of a privacy impact assessment more **difficult**, but these **challenges** can be **overcome**.

158. Big data analytics can involve novel, complex and sometimes unexpected uses of personal data. To establish whether the processing is fair, it is particularly important to assess, before processing begins, to what extent it is likely to affect the individuals whose data is being used and to identify possible mitigation measures. The tool for such an analysis is a privacy impact assessment (PIA). Our code of practice on conducting privacy impact assessments[166] gives practical advice on how to do PIAs, and links the PIA to standard risk-management methodologies.

159. Currently, it is good practice to do a PIA where projects involve new uses of data, but the GDPR[167] will require a PIA – referred to as a 'data protection impact assessment' (DPIA) – in certain instances that are likely to create a high risk to people's rights and freedoms, particularly when the processing uses new technologies. It is highly likely that most big data applications involving the processing of personal data will fall into this category. The GDPR refers to a "systematic and extensive"[168] evaluation of individuals based on automated processing, including profiling, where decisions based on this significantly affect individuals.

[166] Information Commissioner's Office. Conducting privacy impact assessments code of practice. ICO, February 2014, http://ico.org.uk/for_organisations/data_protection/topic_guides/~/media/documents/library/Data_Protection/Practical_application/pia-code-of-practice-final-draft.pdf Accessed 20 June 2016
[167] GDPR Article 35
[168] GDPR Article 35(3)(a)

160. In our discussions with certain industry sectors, some potential privacy risks associated with big data have been identified, for example from the use of inferred data and predictive analytics. Some organisations have said there are no significant differences between using our existing PIA methodology to address these big data risks and using it in their normal data processing.

161. However, others have suggested there may be particular issues to do with big data that make it more challenging to follow the methodology. For example, one of the early PIA stages is to describe the information flows, ie how the data will be used or shared. DPIAs under the GDPR have a similar requirement. As we have seen, big data is likely to involve a discovery phase, when new uses of existing data or new data sources are investigated. This is often about finding new and unexpected correlations. So it may not be clear at the outset what data will be useful or how it will be used, and this can make it more difficult to map the information flows. Furthermore, as we have noted above in the discussion of privacy notices and consent, it may be practically difficult to legitimise the processing by seeking consent.

162. Annex 1 of this paper discusses these and other issues, giving some guidance on how to overcome the inherent challenges of conducting PIAs in a big data context. We would still recommend using our privacy impact assessments code of practice as a full and detailed guide to conducting PIAs. But we hope Annex 1 will help organisations consider how to conduct a PIA for big data analytics and understand how the DPIA requirements in the GDPR affect this.

Privacy by design

In brief…

- The benefits of big data need not come at the cost of privacy.
- Embedding **privacy by design** solutions into big data analytics can help to protect privacy through a range of technical and organisational measures.
- Under the **GDPR**, privacy by design – known as **'data protection by design and by default'** – will become a **legal requirement**.

163. The basis of the privacy by design approach is that if a privacy risk with a particular project is identified, this can be an opportunity to find creative technical solutions that can deliver the real benefits of the project while protecting privacy. As stated in the benefits section of chapter 1, the ICO firmly supports this approach. Data protection and privacy rights are the foundations on which big data analytics can be successfully built. Implementing privacy by design solutions can be mutually beneficial for individuals, big data organisations and society.

164. The concept of privacy by design is often associated with the implementation of techniques to anonymise or pseudonymise personal data, as discussed in the anonymisation section above. One technique in this area is 'differential privacy'. Originally conceived in 2006[169] (but gaining momentum now due to the advancing capabilities of AI), differential privacy involves injecting 'noise' into the answers of dataset queries. The noise should be great enough to provide anonymity at an individual level, but not enough to affect the utility of the answer[170]. Some have criticised differential privacy on the basis that an appropriate trade-off between privacy and utility is unachievable in most circumstances[171]. However, it is quickly

[169] Dwork, Cynthia. Differential Privacy. In 33rd International Colloquium on Automata, Languages and Programming, part II. Springer Verlag, July 2006. https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/dwork.pdf Accessed 24 January 2017

[170] Lipton, Zachary Chase. Differential Privacy: How to make Privacy and Data Mining Compatible. KDnuggets, January 2015. http://www.kdnuggets.com/2015/01/differential-privacy-data-mining-compatible.html Accessed 24 January 2017

[171] Bambauer, Jane; Muralidhar, Krishnamurty and Sarathy, Rathindra. Fool's gold: an illustrated critique of differential privacy. Vanderbilt Journal of Entertainment and Technology Law, Vol 16, No 4, 2014.

becoming a popular privacy by design technique among large technology companies such as Apple[172] and Google[173].

165. However, privacy by design solutions involve not only anonymisation techniques, but a range of other measures both technical and organisational, including:

- security measures to prevent data misuse, such as access controls, audit logs and encryption
- data minimisation measures, to ensure that only the personal data that is needed for a particular analysis or transaction (such as validating a customer) is processed at each stage
- purpose limitation and data segregation measures so that, for example, personal data is kept separately from data used for processing intended to detect general trends and correlations, and
- 'sticky policies' that record individual's preferences and corporate rules within the metadata that accompanies data[174].

166. Much of the initial work on privacy by design was done by the Office of the Information and Privacy Commissioner of Ontario, Canada[175] . More recently, ENISA has produced a wide-ranging report on the use of privacy by design techniques in big data[176]. This includes examples of how the privacy by design approach could be applied in various 'smart city' use cases, such as smart-parking apps, smart metering and citizen platforms. They called for a conceptual shift from "big data versus privacy" to "big data with privacy", a viewpoint strongly shared by the ICO. They concluded that achieving this is not easy,

[172] Greenberg, Andy. Apple's 'Differential Privacy' is about collecting your data – but not your data. Wired, June 2016. https://www.wired.com/2016/06/apples-differential-privacy-collecting-data/ Accessed 24 January 2017

[173] Erlingsson, Úlfar; Pihur, Vasyl and Korolova, Aleksandra. RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. Proceedings of the 21st ACM Conference on Computer and Communications Security. ACM, 2014. https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/42852.pdf Accessed 24 January 2017

[174] Nguyen, Caroline et al. A user-centred approach to the data dilemma: context, architecture and policy. In Digital Enlightenment Yearbook 2013 – The value of personal data. Digital Enlightenment Forum September 2013. http://www.digitalenlightenment.org/publication/def-yearbook-2013-value-personal-data Accessed 7 June 2016

[175] http://www.privacybydesign.ca/ Accessed 7 June 2016

[176] D'Acquisito, Giuseppe et al. Privacy by design in big data. An overview of privacy enhancing technologies in the era of big data analytics. ENISA, December 2015. https://www.enisa.europa.eu/activities/identity-and-trust/library/deliverables/big-data-protection Accessed 7 June 2016

and that more work is needed on privacy-enhancing technologies (PETs) but that "the concept of privacy by design is key in identifying the privacy requirements early at the big data analytics value chain and in subsequently implementing the necessary technical and organizational measures".

167. The concept of privacy by design has now been included in the GDPR, under the heading 'data protection by design and by default'. It will therefore become a legal requirement, as data controllers will be obliged to take "appropriate technical and organisational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed."[177]

[177] GDPR Article 25

Privacy seals and certification

In brief…

- **Certification schemes** can be used to help demonstrate the data protection compliance of big data processing operations.
- The **GDPR** will encourage the establishment of such schemes.

168. In recent years the idea of having a system for certifying that a particular instance of personal data processing complies with data protection requirements, often described as a trust mark or privacy seal, has gained support. It has been suggested that this could be helpful in a big data context to promote consumer trust in the processing[178]. It has been reported that Huawei obtained a form of certification offered by the German company ePrivacy for its Hadoop-based FusionInsight product[179].

169. Certification has been included in the GDPR. It encourages the "establishment of data protection certification mechanisms and of data protection seals and marks" to demonstrate that processing operations comply with the Regulation. These would be awarded by data protection authorities or by accredited certification bodies[180].

170. The ICO has been looking into the feasibility of setting up a privacy seals scheme for data protection. Our idea was that the seal would certify a particular service, product or process (rather than an organisation as a whole) to show that it complies with data protection requirements. The award would be based on rigorous testing and follow-up by an established certification body, applying guidelines laid down by the ICO. We are considering how our original idea relates to the new provisions in the GDPR.

171. The Information Accountability Foundation has argued that big data organisations need to make their own ethical assessments of their processing and they have proposed a framework for this. They

[178] eg Taylor, Simon Data: the new currency. European Voice, June 2014 http://www.masquenegocio.com/wp-content/uploads/2014/07/20140710_InformeBigData.pdf Accessed 8 June 2016
[179] Nunns, James. Compliance with data regulation: how big data analytics vendors are tackling data protection. Computer Business Review, 13 January 2016 http://www.cbronline.com/news/big-data/analytics/compliance-with-data-regulation-how-big-data-analytics-vendors-are-tackling-data-protection-4784023 Accessed 8 June 2016
[180] GDPR Articles 42-43 and Recital 100

suggest that ”accountability agents”, such as certification bodies, could have an important role in monitoring this[181]. In this respect they would supplement the work of data protection authorities. This would need to be considered in the light of the new GDPR provisions that require data controllers to consult with the data protection authority before undertaking processing operations likely to result in high risk, in cases where the proposed mitigating measures do not reduce this risk to an acceptable level[182].

[181] Abrams, Martin. Time for an accountability agents summit. Information Accountability Foundation blog, 11 June 2015. http://informationaccountability.org/category/accountability-agents/ Accessed 10 June 2016

[182] GDPR Article 36

Ethical approaches

In brief...

- An **ethical approach** to the processing of personal data in a big data context is a very **important compliance tool**.
- **Ethics boards** at organisational and national level can help to assess issues and ensure the application of ethical principles.
- Ethical approaches to the use of personal data can help to build **trust** with individuals.
- There is a role for the setting of **big data standards** to encourage best practice across industries.

172. There has been a recent trend towards developing ethical approaches to the use of personal data in big data processing. Several commentators who are concerned about the privacy impact of big data have advocated the need for an ethical approach that supports and goes beyond compliance with legal requirements. For example, the European Data Protection Supervisor has said that, "In today's digital environment, adherence to the law is not enough; we have to consider the ethical dimension of data processing."[183] The Information Accountability Foundation has been working on a Big Data Ethics Initiative[184], which proposes a set of ethical values for assessing big data initiatives, summarised below:

    - Organisations should define the benefits of the analytics.
    - They should not incur the risks of big data analytics if the benefits could be achieved by less risky means.
    - The insights should be sustainable.
    - The processing should respect the interests of stakeholders.
    - The outcomes of the processing should be fair to individuals and avoid discriminatory impacts.

[183] European Data Protection Supervisor. Towards a new digital ethics. Opinion 4/2015. EDPS, September 2015. https://secure.edps.europa.eu/EDPSWEB/webdav/site/mySite/shared/Documents/Consultation/Opinions/2015/15-09-11_Data_Ethics_EN.pdf Accessed 10 June 2016

[184] Information Accountability Foundation. Big data ethics initiative. IAF website. http://informationaccountability.org/big-data-ethics-initiative/ Accessed 10 June 2016

173. We are now seeing examples of both public and private sector organisations developing their own sets of ethical principles. These are essentially ground rules setting out how the organisation will use people's data. They typically stress values of fairness and transparency. They are usually intended both as a guide for employees when they are using data and developing new projects, and as a means of reassuring customers and building a relationship of trust with them.

174. Sometimes these principles are condensed into a simple 'litmus test' to remind employees to think about them when planning new uses of data; for example, would you want the data of a member of your family to be used in this way? The US company Caesar's Entertainment applies a 'sunshine test': if the details of how we use data were made public, would it strengthen or threaten customer relationships?[185]

Ethical approaches in the private sector

Aimia. Aimia is a global company in the field of loyalty management and runs Nectar and other loyalty programmes. It has developed a set of data values with the acronym TACT: Transparency, Added value, Control and Trust[186]. Transparency means telling customers what data is being collected, how it is being collected and how it is being used. Added value means making customers aware they will receive rewards for their participation. Control is about giving customers control over the data they provide and enabling them to share it and to opt out. Trust means giving customers confidence that the data will only be used in ways that you say you will use it and only share it with partners you have identified.

IBM. IBM has published an ethical framework for big data analytics[187]. This takes account of the context in which the

[185] Etlinger, Susan and Groopman, Jessica. The trust imperative: a framework for ethical data use. Altimeter Group, June 2015. http://www.altimetergroup.com/2015/06/new-report-the-trust-imperative-a-framework-for-ethical-data-use/ Accessed 10 June 2016

[186] Johnson, David and Henderson-Ross, Jeremy The new data values Aimia, 2012. http://www.aimia.com/content/dam/aimiawebsite/CaseStudiesWhitepapersResearch/english/WhitepaperUKDataValuesFINAL.pdf Accessed 10 June 2016

[187] Chessell, Mandy. Ethics for big data and analytics. IBM Big Data and Analytics Hub, 2014 http://www.ibmbigdatahub.com/sites/default/files/whitepapers_reports_file/TCG%20Study%20Report%20-%20Ethics%20for%20BD%26A.pdf Accessed 10 June 2016

data will be collected and used; whether people will have a choice in giving their data; whether the amount of data and what will be done with it is reasonable in terms of the application; the reliability of the data; who owns the insights to be gained from the data; whether the application is fair and equitable; the consequences of processing; people's access to the data; and accountability for mistakes and unintended consequences.

Vodafone. Vodafone publishes a set of privacy commitments[188]. These cover respect for people's data; openness and honesty with customers; giving people meaningful choices; applying privacy by design; minimising privacy impacts when balancing privacy rights against other obligations; complying with privacy laws; and accountability.

International developments. In the USA, the Alliance of Automobile Manufacturers and the Global Alliance of Automakers has produced a set of privacy principles for the consumer data derived from new vehicle technologies[189]. At an international level, the GSMA, which represents mobile operators worldwide, produced guidelines on the use of mobile phone data in responding to the Ebola outbreak[190].

175. The trend towards spelling out ethical principles is evident not just among private-sector companies but also in the public sector. Ethical principles for research have been in place for some time, for example in universities, but the growth of big data means they are having to be applied in more challenging situations, particularly where data is not collected directly from individual participants, such as the use of social-media data.

[188] Vodafone. Privacy and security. Vodafone website, June 2015. https://www.vodafone.com/content/sustainabilityreport/2015/index/operating-responsibly/privacy-and-security.html Accessed 10 June 2016

[189] Consumer privacy protection principles. Privacy principles for vehicle technologies and services. Alliance of Automobile Manufacturers Inc and Association of Global Automakers Inc. November 2014. http://www.autoalliance.org/?objectid=865F3AC0-68FD-11E4-866D000C296BA163 Accessed 10 June 2016

[190] GSMA guidelines on the protection of privacy in the use of mobile phone data for responding to the Ebola outbreak. GSMA, November 2014. http://www.gsma.com/mobilefordevelopment/wp-content/uploads/2014/11/GSMA-Guidelines-on-protecting-privacy-in-the-use-of-mobile-phone-data-for-responding-to-the-Ebola-outbreak-_October-2014.pdf Accessed 10 June 2016

Cabinet Office. The Cabinet Office has published a Data Science Ethical Framework[191]. This aims to help researchers, as big data methods are starting to be used in research in the public sector. It puts forward six principles:

- Start with clear user need and public benefit.
- Use data and tools which have the minimum intrusion necessary.
- Create robust data science methods.
- Be alert to public perceptions.
- Be as open and accountable as possible.
- Keep data secure.

176. There is a role in this for councils or boards of ethics, both within organisations and at a national level. A large organisation may have its own board of ethics, which could ensure that its ethical principles are applied, and could make assessments of difficult issues such as the balance between legitimate interests and privacy rights. The use of internal ethics boards is advocated by the Council of Europe's Consultative Committee of Convention 108 in its recently published big data guidelines[192]. An important issue in this scenario is the organisational relationship between the ethics board and employees with responsibilities for data and analytics, such as the chief data officer and the data protection officer.

177. The Royal Statistical Society[193] and the House of Commons Science and Technology Committee[194] (the STC) have called for a UK Council

[191] Cabinet Office. Data science ethical framework. Cabinet Office, May 2016. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/524298/Data_science_ethics_framework_v1.0_for_publication__1_.pdf Accessed 10 June 2016
[192] Consultative Committee of the Convention for the protection of individuals with regard to automatic processing of personal data. Guidelines on the protection of individuals with regard to the processing of personal data in a world of big data. Council of Europe, 23 January 2017. https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000016806ebe7a Accessed 16 February 2017
[193] Royal Statistical Society. The opportunities and ethics of big data. RSS, February 2016. http://www.rss.org.uk/Images/PDF/influencing-change/2016/rss-report-opps-and-ethics-of-big-data-feb-2016.pdf Accessed 10 Jun 2016
[194] House of Commons Science and Technology Committee. The big data dilemma. Fourth report of session 2015-16 HC468. The Stationery Office, February 2016.

of Data Ethics to give a national lead and guidance on these issues. The government has agreed to consider how this can be established, probably under the auspices of the Alan Turing Institute[195]. In a subsequent report, the STC has also called for the creation of a standing Commission on Artificial Intelligence to work closely with the Council of Data Ethics on the social, ethical and legal implications of the application of AI techniques[196]. The government's response stopped short of agreeing to set up a Commission, but detailed some similar work under way by the Royal Society and the British Academy on the implications of machine learning and data governance[197].

178. An example of an advisory board dealing with privacy issues in the public sector is in Seattle, in the USA[198]. Seattle has set up a Privacy Advisory Board that oversees how the city uses personal data, particularly in the context of 'smart city' initiatives. It publishes a set of privacy principles and encourages the use of PIAs.

179. There seem to be several factors pushing the adoption of ethical principles. In the public sector, evidence of a lack of public awareness about data use and suspicions about data sharing have led to calls for ethical policies to be made explicit[199]. In the private sector there is a commercial imperative to mitigate risk. It would harm a company's reputation if it was the subject of media stories about the misuse of personal data, and consumers can also publicise

---

http://www.publications.parliament.uk/pa/cm201516/cmselect/cmsctech/468/468.pdf Accessed 10 June 2016

[195] House of Commons Science and Technology Committee. The big data dilemma. Government response to the Committee's fourth report of session 2015-16. HC992. Stationery Office, April 2016. http://www.publications.parliament.uk/pa/cm201516/cmselect/cmsctech/992/992.pdf Accessed 10 June 2016

[196] House of Commons Science and Technology Committee. Robotics and artificial intelligence. Fifth report of session 2016-17 HC145. The Stationery Office, October 2016. http://www.publications.parliament.uk/pa/cm201617/cmselect/cmsctech/145/145.pdf Accessed 20 December 2016

[197] House of Commons Science and Technology Committee. Robotics and artificial intelligence: Government response to the Committee's fifth report of session 2016-17. HC145. Stationary Office, December 2016. http://www.publications.parliament.uk/pa/cm201617/cmselect/cmsctech/896/89602.htm Accessed 17 January 2017

[198] Kitchin, R. Getting smarter about smart cities: Improving data privacy and data security. Data Protection Unit, Department of the Taoiseach, Dublin, Ireland, January 2016. http://www.taoiseach.gov.ie/eng/Publications/Publications_2016/Smart_Cities_Report_January_2016.pdf Accessed 17 June 2016

[199] Evans, Harry; Ginnis, Steve and Bartlett, Jamie. #SocialEthics. A guide to embedding ethics in social media research. Ipsos MORI, December 2015. https://www.ipsos-mori.com/Assets/Docs/Publications/im-demos-social-ethics-in-social-media-research-summary.pdf Accessed 15 April 2016

their views to the world instantly. This is an important consideration in a competitive environment.

180. More positively, companies may also seek to develop their relationship with the customer, so that they trust the company with their data and are happy to provide more data in return for enhanced services or other benefits. ICO-commissioned research shows that the more people trust businesses with their personal data, the more appealing they find new product offers such as smart thermostats and telematics devices in cars[200]. An international study of consumer attitudes to personal data reported in the Harvard Business Review[201] found that people are willing to accept potentially intrusive uses of their data, such as profiling, in return for the enhanced benefits of a service like Google Now. However, it is not simply about value; trust is also important. If two organisations offer a service with the same value, people are more likely to allow their data to be used by the one they trust more. The study found that being transparent with customers, teaching them about data use and giving them control over their data are key elements in building trust. This suggests there is a business case for developing an approach that aims to build trust and is based on transparency and fairness.

181. It is notable that these ethical frameworks have been developed not by regulators but by companies and other organisations themselves. Nevertheless, many aspects of these frameworks echo key data protection principles and requirements. They reflect the importance of telling people what is being done with their data and who it will be shared with, considering whether the uses of that data are within people's expectations, giving people access to their data and some control over the use of it, and considering the impact of the analytics on the people to whom the data relates. Developing ethical principles and frameworks for big data is a job for data controllers rather than data protection authorities. But we welcome this development because it helps organisations to ensure that their use of big data complies with data protection principles. In particular, it helps to meet what we see as the key issues of fairness and transparency.

182. In addition to ethical frameworks, there is also a role for developing common standards for big data analytics. Organisations such as the

[200] Citizenme. Annual track 2016. ICO, June 2016 https://ico.org.uk/about-the-ico/our-information/research-and-reports/information-rights-research/ Accessed 20 June 2016
[201] Morey, Timothy; Forbath, Theodore and Schoop, Allison. Customer data: designing for transparency and trust. Harvard Business Review, May 2015. https://hbr.org/2015/05/customer-data-designing-for-transparency-and-trust Accessed 14 June 2016

BSI[202], the International Telecommunications Union[203] and the International Organisation for Standardization[204] have been working towards a set of big data standards to help establish best practice and reduce risk for organisations involved in big data processing. The ICO supports the idea of big data standards and encourages their development, especially among trade associations and industry groups that can strongly influence their members.

[202] Circle Research. Big data and standards market research. BSI Standards Ltd, January 2016. http://shop.bsigroup.com/forms/The-Big-Data-and-market-research-report/ Accessed 20 December 2016

[203] Acharya, Sanjay. ITU members agree international standard for Big Data. International Telecommunications Union, 18 December 2015. http://www.itu.int/net/pressoffice/press_releases/2015/66.aspx#.WFo5xcpvjAU Accessed 21 December 2016

[204] ISO/IEC JTC 1 Information technology. Big Data Preliminary Report 2014. International Organisation for Standardization, January 2015. http://www.iso.org/iso/big_data_report-jtc1.pdf Accessed 21 December 2016

Personal data stores

In brief...

- The use of personal data stores can address issues of fairness and lack of transparency by giving individuals **greater control** over their personal data.
- Personal data stores can support the concept of **data portability** (which will become law under the GDPR in certain conditions) regarding the re-use of an individual's personal data under their control.

183. It has been suggested that one way to increase an individual's control over the use of their data is through what are usually called personal data stores, or sometimes personal information management services. These are third-party services that hold people's personal data on their behalf and make it available to organisations as and when the individuals wish to do so. Rubinstein[205], an early proponent of this concept, saw it as a way of embedding privacy controls by managing organisations' access to personal data and building in "fine-grained" privacy preferences. The European Data Protection Supervisor also sees personal data stores as a potential way of tackling concerns about individuals' loss of control of their data[206].

184. More recently, Obar[207] has criticised the idea that individuals can effectively control how their personal data is used in a big data environment as the "fallacy of data privacy self-management". He says people are not aware that their data is being collected or how it is used, and don't have the time to read privacy notices. Instead, he argues for "representative data management", ie a system of intermediaries who would manage a person's data on his or her behalf.

---

[205] Rubinstein, Ira S. Big data. The end of privacy or a new beginning? International Data Privacy Law, 25 January 2013. http://idpl.oxfordjournals.org/content/early/2013/01/24/idpl.ips036.full.pdf+html Accessed 13 June 2016

[206] European Data Protection Supervisor. Meeting the challenges of big data. Opinion 7/2015. EDPS, November 2015. https://secure.edps.europa.eu/EDPSWEB/webdav/site/mySite/shared/Documents/Consultation/Opinions/2015/15-11-19_Big_Data_EN.pdf Accessed 13 June 2016

[207] Obar, Jonathan A. Big data and the phantom public. Walter Lippmann and the fallacy of data privacy self-management. Big Data & Society, July-December 2015 vol. 2 no. 2. http://m.bds.sagepub.com/content/2/2/2053951715608876 Accessed 16 June 2016

185. There is some evidence of public support for personal data stores in the UK. In a survey for the Digital Catapult[208], 30% of the people they surveyed said they would welcome a service to help them collect, manage and preserve their personal data. Some services of this type already exist, for example Mydex[209], which provides free encrypted personal data stores. This enables individuals to share data from their personal data store when they apply for a service or make a purchase online. It also enables them to provide a verified digital identity. There have been proposals for setting up personal data stores on a co-operative basis, so that the individuals who keep their data in the store could benefit financially when it is used[210].

186. The growth of personal data stores can also support the government's midata initiative. This initiative currently focuses on the banking, telecoms and energy sectors. It enables individuals to receive the data that organisations in these sectors hold about them in a machine-readable electronic form. They can then use this, for example to find better deals using comparison websites[211].

187. The GDPR puts the concept of data portability into law. If a data controller is processing personal data on the basis of consent or contract, the data subject will have the right to receive the data they have provided in a "structured, commonly used and machine readable format". They also have the right to transmit this to another data controller[212].

188. This is a developing area but personal data stores can offer individuals a degree of control over the re-use of their personal data across different services. This can at least help to address the issues of fairness and lack of transparency that we have identified as potentially problematic in big data.

---

[208] Digital Catapult. Trust in personal data: a UK review. Digital Catapult, July 2015. http://www.digitalcatapultcentre.org.uk/pdtreview/ Accessed 13 June 2016
[209] https://mydex.org
[210] Open Data Manchester. Open data co-operation – building a data co-operative. Open Data Manchester, April 2015. https://opendatamanchester.org.uk/2015/04/14/open-data-cooperation-building-a-data-cooperative/ Accessed 13 June 2016
[211] Davies, Sean et al. Why midata will change personal banking forever. Gocompare, March 2015. http://www.gocompare.com/current-accounts/midata/ Accessed 13 June 2016
[212] GDPR Article 20

Algorithmic transparency

In brief…

- **Auditing techniques** can be used to identify the factors that influence an algorithmic decision.
- **Interactive visualisation systems** can help individuals to understand why a recommendation was made and give them control over future recommendations.
- **Ethics boards** can be used to help shape and improve the transparency of the development of machine learning algorithms.
- A **combination of technical and organisational approaches** to algorithmic transparency should be used.

189. One of the implications discussed in the accountability and governance section of chapter 2 was the need for algorithmic accountability to ensure and demonstrate the data protection compliance of 'black box' big data processing activities such as machine learning. There is no consensus on a simple 'one-size-fits-all' type solution. But there is increasing debate about how best to achieve algorithmic transparency, from which several approaches have emerged.

190. A popular approach, championed by several commentators, is algorithmic auditing. In an MIT Technology Review article, 'auditability' is one of the five principles suggested for accountable algorithms[213]. The idea is that auditability should be 'baked in' to algorithms in the development stage to enable third parties to check, monitor, review and critique their behaviour. For companies in the private sector the concept of an algorithmic audit is likened to an accounting audit which, while carried out in confidence to protect proprietary information, can still provide public assurance.

191. A lack of technical capability and computational resources has been cited as a potential barrier to the auditing of algorithms[214], yet there is also evidence of good progress on its implementation. A paper

[213] Diakopoulos, Nicholas and Friedler, Sorelle. How to Hold Algorithms Accountable. MIT Technology Review, 17 November 2016. https://www.technologyreview.com/s/602933/how-to-hold-algorithms-accountable/ Accessed 16 December 2016

[214] Mittelstadt, Brent. Automation, Algorithms, and Politics: Auditing for Transparency in Content Personalization Systems. International Journal of Communication, October 2016

presented at the 2016 International Conference on Data Mining showed a technique for algorithmic auditing that was evidenced as being effective at identifying discrete factors that influence the decisions made by algorithms[215]. Furthermore, in the USA, consultant companies are already being set up that specialise in providing algorithmic auditing services to their clients[216].

192. One of the applications of big data analytics is Natural Language Generation (NLG). This is the ability to create a human-understandable narrative from the analysis of various sources of data. IBM's Slamtracker system is an example of NLG in practice; it converts data on tennis matches at Wimbledon into real-time automated stories and Twitter messages[217]. NLG is commonly associated with this type of use (news and weather reports), but it may also be possible to apply it to algorithmic decision making with a view to increasing the transparency of how such decisions are made. According to an article written for the Association for Computing Machinery:

> "Other methods are being developed in natural language generation (NLG) to output text that explains why or how a decision was reached. Imagine if your favorite machine learning library, say scikit-learn, could explain in a sentence why a particular input case was classified the way it was. That would be useful for debugging, if nothing else."[218]

193. So organisations may wish to look into ways of integrating NLG into their new and existing big data processing activities. This would enable them to give individuals explanations of decisions based on automated processing.

194. A different approach to algorithmic transparency is to combine visualisation and interactivity. Much research has been done on this, particularly on the use of big data analytics for recommendation systems. Studies have found, for instance, that visualisation tools

[215] Adler, Philip et al. Auditing Black-box Models for Indirect Influence. IEEE International Conference on Data Mining. December 2016
[216] O'Neil Risk Consulting and Algorithmic Auditing. http://www.oneilrisk.com/ Accessed 16 December 2016
[217] Marr, Bernard. Can Big Data Algorithms Tell Better Stories Than Humans? Forbes, 22 July 2015. http://www.forbes.com/sites/bernardmarr/2015/07/22/can-big-data-algorithms-tell-better-stories-than-humans/#65d89fd242ba Accessed 19 December 2016
[218] Diakopoulos, Nicholas. Accountability in algorithmic decision making. Communications of the ACM 59, no. 2 (2016): 56-62

such as TalkExplorer[219] and SetFusion[220] allowed individuals to better understand why recommendations had been made for them and enabled them to create more accurate recommendations for themselves. The visual explanations and interactivity in these systems were facilitated through features such as bookmark interrelation charts, Venn diagrams and adjustable sliders to change the importance of recommenders' methods.

195. Where interactive visualisation systems might prove difficult to implement for certain types of big data analytics, a similar but more simplified approach could be adopted whereby individuals are given the opportunity to check and correct the outputs of machine learning algorithms. For example, if an organisation undertakes automated profiling on its customers to determine insurance premiums, it could allow them to inspect those profiles and correct any inaccurate labels assigned to them. Beyond demonstrating compliance with the data protection principle of accuracy, this could also lead to an overall improvement in the precision of the machine learning application[221].

196. Another non-technical approach to algorithmic transparency is the use of ethics boards. These are discussed in the ethical approaches section above as part of a general approach to data protection compliance, but they can also be used more specifically to appraise and make decisions on the development and application of machine learning algorithms. For instance, a European study[222] on the implementation of an algorithmic video surveillance system used an ethics board to which the researcher reported regularly regarding the algorithm. This allowed the board to understand the development of the algorithm and raise relevant questions so that matters of concern could be taken back to the project team for adjustments. To make transparent the algorithm's step-by-step development, the ethics board's minutes were made public. A well-known example of the use of an ethics board in an algorithmic context is Google's AI ethics board, set up when it originally acquired the UK company DeepMind in 2014. However, Google has been criticised by some for its lack of

[219] Verbert, Katrien et al. Visualizing recommendations to support exploration, transparency and controllability. In Proceedings of the 2013 International Conference on Intelligent User Interfaces, pp. 351-362. ACM, 2013
[220] Parra, Denis; Brusilovsky, Peter and Christoph Trattner. See what you want to see: visual user-driven approach for hybrid recommendation. In Proceedings of the 19th International Conference on Intelligent User Interfaces, pp. 235-240. ACM, 2014
[221] Diakopoulos, Nicholas. Accountability in algorithmic decision making. Communications of the ACM 59, no. 2 (2016): 56-62
[222] Neyland, Daniel. Bearing accountable witness to the ethical algorithmic system. Science, Technology & Human Values 41, no. 1 (2016): 50-76

transparency regarding board members and the focus of their work[223].

197. AI and machine learning are constantly evolving areas of research and practice, and consequently so are the discussions around transparency and accountability. Therefore we do not claim that the above approaches to algorithmic transparency are a complete list. As stated at the start of this section, no single approach to algorithmic transparency would work for every big data organisation. Rather, a combination of complementary approaches, both technical and organisational, should be adopted to suit the design and purpose of a particular big data application. This view is reflected in the findings of research on algorithmic transparency in the USA[224] and Switzerland[225].

[223] Shead, Sam. DeepMind is staying silent on who sits on Google's AI ethics board. Business Insider UK, 5 December 2016. http://uk.businessinsider.com/deepmind-is-remaining-silent-on-who-sits-on-googles-ai-ethics-board-2016-12 Accessed 19 December 2016

[224] Burrell, Jenna. How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data & Society 3, no. 1 (2016)

[225] Saurwein, Florian; Just, Natascha and Michael Latzer. Governance of algorithms: options and limitations. info 17, no. 6 (2015)

Chapter 4 – Discussion

198. In chapters 2 and 3 we have shown that the use of big data analytics has several implications for data protection and privacy rights, but that those implications are not insurmountable barriers to the legal and ethical application of such analytic techniques. Various tools and approaches are available to help with compliance. Rather than restricting the use of big data analytics, these tools can encourage innovation and support delivery of the benefits that flow from big data.

199. However, we recognise the emerging view that the data protection principles, as embodied in UK and EU law, are no longer adequate to deal with the big data world. The World Economic Forum characterised the "traditional data protection approach" as one where "the individual was involved in consenting to data use at the time of collection. The organisation that collected the data then used it for a specified use, based on user consent, and then deleted the data when it was no longer needed for the specified purpose."[226] It is this model that critics of data protection have in mind.

200. This so-called 'notice and consent' model has been criticised on the grounds that users lack the time, willingness or ability to read lengthy privacy notices; therefore, even if they give consent on this basis it is effectively meaningless[227]. It is also argued that there are practical difficulties in giving privacy notices in situations where data is collected by observing or recording the actions of individuals, rather than individuals consciously providing it. We have discussed these issues in the sections on privacy notices and consent above.

201. The notice and consent model is a fundamental facet of the data protection principle of transparency. Criticism of this model is mirrored by wider criticism of the role of transparency in the evolving world of big data analytics. Some suggest, for instance, that the concept of transparency is inadequate when it comes to the complex and opaque nature of algorithms[228] and that it can lead to "gaming of the decisionmaking process."[229]

---

[226] World Economic Forum Unlocking the value of personal data; from collection to usage. WEF, February 2013 http://www3.weforum.org/docs/WEF_IT_UnlockingValuePersonalData_CollectionUsage_Report_2013.pdf Accessed 16 June 2016

[227] USA. Executive Office of the President. President's Council of Advisors on Science and Technology. Big data and privacy. A technological perspective. White House, May 2014. https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf Accessed 16 June 2016

[228] Ananny, Mike and Crawford, Kate. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media and

202. In addition to such arguments about the limitations of transparency, there is a view that the problems of big data analytics, and any potential harms, arise not from how the data is *collected* but from how it is *used*. For example, some commentators have questioned what harm is caused in any case to an individual simply by the *collection* of their data, for example through mass surveillance by governments[230]. An increased focus on the *use* of data has led to the championing of accountability as an answer to big data issues, as opposed to transparency[231]. Rather than focusing on providing people with the 'hows' and 'whys' of the processing of their personal data, accountability concentrates on monitoring its use through mechanisms such as scrutinising the technical design of algorithms[232], auditability[233] and software-defined regulation[234].

203. The emerging importance of accountability is reflected in the GDPR, which includes it explicitly as a new data protection principle[235] and is, in part, being introduced to address the implications of the processing of personal data in a big data world. New provisions regarding data protection by design and default[236], data protection impact assessments[237] and certification[238] all emphasise the growing role accountability has to play both within organisations but also externally. It should be noted, for instance, that under the GDPR

Society, 13 December 2016. http://journals.sagepub.com/doi/pdf/10.1177/1461444816676645 Accessed 21 December 2016

[229] Kroll, Joshua et al. Accountable Algorithms. University of Pennsylvania Law Review, March 2016 Vol 165

[230] For example in chapter 8 of: van der Sloot, Bart; Broeders, Dennis and Schrijvers, Erik. Exploring the boundaries of big data. Netherlands Scientific Council for Government Policy/ Amsterdam University Press, April 2016. http://www.wrr.nl/fileadmin/en/publicaties/PDF-Verkenningen/Verkenning_32_Exploring_the_Boundaries_of_Big_Data.pdf Accessed 16 June 2016

[231] Beresford, Tom. Algorithmic transparency is not the solution you're looking for – algorithmic accountability is. Gamification of Work, 2 November 2016. http://gamificationofwork.com/2016/11/algorithmic-transparency-not-solution-youre-looking-algorithmic-accountability/ Accessed 21 December 2016.

[232] Bomhof, Freek. In Order to Trust Big Data, Transparency Is Not Enough. DataFloq, 23 October 2016. https://datafloq.com/read/transparency-in-big-data-is-not-enough/138 Accessed 21 December 2016

[233] Diakopoulos, Nicholas and Friedler, Sorelle. How to Hold Algorithms Accountable. MIT Technology Review, 17 November 2016

[234] Taneja, Hemant. The need for algorithmic accountability. TechCrunch, 8 September 2016. https://techcrunch.com/2016/09/08/the-need-for-algorithmic-accountability/ Accessed 21 December 2016

[235] GDPR Article 5(2)

[236] GDPR Article 25

[237] GDPR Article 35

[238] GDPR Articles 42-43

data protection regulators will have a role in giving prior authorisation to certain forms of 'high risk' processing[239]. The prominence of accountability is further exemplified by the large volume of discussion about its features and use among academics[240], practitioners[241] and journalists[242].

204. We freely acknowledge the increased weight being placed on accountability in a big data context, but we do not see it as the death of transparency as a data protection principle. To return to the example cited above about the mass collection of personal data: even if such processing were held to account regarding the potential issues arising from the *use* of the data – for instance, the increased severity and likelihood of data breaches, internal misuse by rogue employees and undesirable secondary use[243] – the lack of a transparent process in the *collection* of the personal data would still cause other problems. Quite simply, if people have not been informed about the processing of their personal data, they are unlikely to be able to exercise their rights, even if they would regard the processing as unfair, because they would not be aware of it.

205. As we have shown in chapter 3, particularly in the privacy notices section, achieving transparency is not impossible in a big data world. But the methods by which it is achieved are altering, with a shift towards a more 'layered' approach to transparency. This approach is exemplified in the layering of privacy notices to individuals (as and when the purposes for collecting and using their personal data emerge), and also in the layering of information about the inner workings of big data analytics, with a greater level of detail and access given to regulators, auditors and accredited certification bodies.

206. As we have shown at points throughout this paper, big data analytics and data protection should not be viewed in simple binary terms; the same also applies to the principles of transparency and

---

[239] GDPR Article 36

[240] Butin, Denis; Chicote, Marcos and Le Metayer, Daniel. Strong Accountability: Beyond Vague Promises. In Reloading Data Protection: Multidisciplinary Insights and Contemporary Challenges. Springer, pp.343-369, 2014

[241] Bellamy, Bojana and Heyder, Markus. Protecting Privacy in a World of Big Data: The Role of Enhanced Accountability. SCL – The IT Law Community, 18 May 2016. http://www.scl.org/site.aspx?i=ed47678 Accessed 22 December 2016

[242] Burgess, Matt. Holding AI to account: will algorithms ever be free from bias if they're created by humans? Wired, 6 December 2016. http://www.wired.co.uk/article/creating-transparent-ai-algorithms-machine-learning Accessed 22 December 2016

[243] Brookman, Justin and Hans, GS. Why collection matters. Surveillance as a de facto privacy harm. In: Big data and privacy. Making ends meet pp 11-14. Future of Privacy Forum and Stanford Law School, January 2014. https://fpf.org/2014/01/29/essays-on-big-data-and-privacy/ Accessed 16 June 2016

accountability. There has been somewhat of a paradigm shift regarding the emerging importance of accountability, but this is not a wholesale replacement for transparency. In fact, the Centre for Information Policy Leadership lists transparency as part of one of the essential elements of accountability[244]. In our view, a combination of both approaches will help to ensure the protection of privacy rights while delivering the benefits of big data.

207. This view is supported by other regulators in the EU and other jurisdictions. The USA does not have overarching data protection legislation in the same way as the EU, but in a speech on big data the chair of the Federal Trade Commission said that "focusing on consumer choice at the time of collection is critical, but use restrictions have their place too."[245] The Australian Information Commissioner has also stressed the continuing importance of notice and consent in a big data context, as well as the need for organisations to have a robust and accountable privacy governance framework in place[246].

[244] Centre for Information Policy Leadership. Data Protection Accountability: The Essential Elements A Document for Discussion. Hunton and Williams LLP, October 2009 https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/data_protection_accountability-the_essential_elements__discussion_document_october_2009_.pdf Accessed 16 February 2017

[245] Ramirez, Edith. The privacy challenges of big data: a view from the lifeguard's chair. Federal Trade Commission, August 2013. https://www.ftc.gov/sites/default/files/documents/public_statements/privacy-challenges-big-data-view-lifeguard%E2%80%99s-chair/130819bigdataaspen.pdf Accessed 16 June 2016

[246] Pilgrim, Timothy. Big data and privacy: a regulator's perspective. Office of the Australian Information Commissioner, June 2015. https://www.oaic.gov.au/media-and-speeches/speeches/big-data-and-privacy-a-regulators-perspective Accessed 16 June 2016

Chapter 5 – Conclusion

208. Big data, AI and machine learning are becoming widespread in the public and private sectors. They may increasingly be seen as 'business as usual', but the key characteristics of big data analytics still represent a step change in the processing of personal data.

209. The analysis of big data using techniques made possible by AI creates implications for data protection, and it can be more challenging to apply the data protection principles when using personal data in a big data context. These implications arise not only from the volume of the data but from the ways in which it is generated, the propensity to find new uses for it, the complexity of the processing and the possibility of unexpected consequences for individuals. In this paper we have tried to map data protection principles against the features of big data analytics and highlight the areas of potential difficulty.

210. However, we have also discussed several tools and approaches, including anonymisation, PIAs and privacy by design, that can help organisations to ensure their processing complies with data protection legislation and minimises the impact on privacy. We welcome the trend towards organisations developing their own ethical principles and building relationships of trust with the public, because putting this into practice will assist compliance with data protection requirements. Recent moves towards setting up 'councils of ethics', within organisations and nationally, are a positive development that should also support this.

211. We recognise the many benefits that can flow from the use of big data for individuals, public services, business and society in general. But these benefits will only truly be felt when privacy rights and data protection are embedded in the methods by which they are achieved. So it is crucial for organisations to be clear about the potential benefits of what they are doing and the steps they have taken to address privacy risks.

212. Yet data protection is not simply a legal requirement for big data analytics; it is also prudent and advantageous for other reasons. Further to the creativity and innovation it encourages, we also argue that getting data protection right helps to ensure data quality – which is becoming ever more critical for businesses and public sector organisations in a big data world. Therefore, for organisations that rely on big data, data protection is important not just in the legal or compliance department but also for people working in data analytics, marketing and research. They need to be aware of the data

protection implications of big data analytics and of tools such as PIAs. This also suggests that data protection should be part of the higher education curriculum for people going into these roles.

213. We are aware of the view that, given some of the challenges of applying data protection principles to big data analytics, a different legal or regulatory approach is required. However, we do not accept the idea that data protection, as currently embodied in legislation, does not work in a big data context. We maintain that big data is not a game played by different rules. We acknowledge the increasing importance of accountability in addressing some of these challenges, but we do not see it as a replacement for the more traditional principle of transparency. Transparency still has a significant role to play and we argue it can still be achieved, even in a complex world of AI and machine learning.

214. Throughout this paper we have referred to relevant provisions of the GDPR. The GDPR is intended partly to address some of the questions raised by big data analytics. It aims to strengthen privacy rights in this context and refers specifically to issues such as profiling, and tools such as data protection impact assessments and data protection by design and default.

215. The ICO has a role in helping big data organisations to meet new and existing data protection obligations. Although this paper is intended primarily as a discussion document, we have added a practical dimension, particularly in exploring compliance tools in chapter 3 and conducting PIAs in Annex 1. We have also pulled out six key recommendations from these discussions, which we present in chapter 6.

216. However, this paper is not the end of our work on data protection and big data analytics. We have several current and planned activities that are linked with our work in this area, particularly regarding the GDPR and its provisions on matters such as 'profiling' and 'risk', both of which are especially relevant for big data, AI and machine learning. Outputs from this and other work will be published (or linked to) on our website www.ico.org.uk and promoted in our monthly e-newsletter.

217. We will continue our work in this area to help and encourage organisations to meet their obligations. But we also have a role in responding to breaches of data protection legislation with proportionate regulatory action, which can include issuing enforcement notices and monetary penalty notices. This is no different in the world of big data, AI and machine learning, and we

will continue to use our powers where necessary and in line with our Regulatory Action Policy[247].

[247] Information Commissioner's Office. Data Protection Regulatory Action Policy. ICO, August 2013. https://ico.org.uk/media/1853/data-protection-regulatory-action-policy.pdf Accessed 18 January 2017

Chapter 6 – Key recommendations

218. Based on discussions regarding compliance tools in chapter 3 of this paper, we have pulled out six key recommendations that we feel will help organisations to achieve and go beyond data protection compliance in a big data world:

Organisations should…
1. …carefully consider whether the big data analytics to be undertaken actually requires the processing of personal data. Often, this will not be the case; in such circumstances organisations should use appropriate techniques to **anonymise** the personal data in their dataset(s) before analysis… …read more in the anonymisation section.
2. …be transparent about their processing of personal data by using a combination of innovative approaches in order to provide meaningful **privacy notices** at appropriate stages throughout a big data project. This may include the use of icons, just-in-time notifications and layered privacy notices… …read more in the privacy notices section.
3. …embed a **privacy impact assessment** framework into their big data processing activities to help identify privacy risks and assess the necessity and proportionality of a given project. The privacy impact assessment should involve input from all relevant parties including data analysts, compliance officers, board members and the public… …read more in the privacy impact assessment section and Annex 1.
4. …adopt a **privacy by design** approach in the development and application of their big data analytics. This should include implementing technical and organisational measures to address matters including data security, data minimisation and data segregation…

… read more in the privacy by design section.

5. …develop **ethical principles** to help reinforce key data protection principles. Employees in smaller organisations should use these principles as a reference point when working on big data projects. Larger organisations should create ethics boards to help scrutinise projects and assess complex issues arising from big data analytics…

   …read more in the ethical approaches section.

6. …implement innovative techniques to develop **auditable machine learning algorithms**. Internal and external audits should be undertaken with a view to explaining the rationale behind algorithmic decisions and checking for bias, discrimination and errors…

   …read more in the algorithmic transparency section.

Annex 1 – Privacy impact assessments for big data analytics

Introduction

Feedback on version 1 of this paper, and subsequent discussions with industry sectors, identified an interest in the development of some specific guidance on conducting privacy impact assessments (PIAs) in a big data context. PIAs are particularly important in this area because of the capabilities of big data analytics and the potential data protection implications that can arise, as identified and discussed in this paper.

Furthermore, although PIAs are not required under the DPA, they will be required under the GDPR in situations where processing is likely to result in a high risk to the rights and freedoms of individuals, in particular when using new technologies[248]. Specifically, the GDPR states that a PIA – referred to as a "data protection impact assessment" (DPIA) – is required in the case of:

> *"a systematic and extensive evaluation of personal aspects relating to natural persons which is based on automated processing, including profiling, and on which decisions are based that produce legal effects concerning the natural person or similarly significantly affect the natural person"*[249]

Therefore it's highly likely that, under the GDPR, a DPIA will be legally required for most big data applications involving the processing of personal data. So we share the view that some guidance on PIAs/DPIAs for big data analytics would be useful, and we seek to provide it here.

The GDPR sets out a structure for DPIAs[250] which, broadly speaking, maps on to the PIA framework used in our 'Conducting privacy impact assessments code of practice'[251] (PIA COP); this is shown in the table below. So the guidance provided here uses our existing PIA COP framework as a basis for a discussion of the issues at play. This is followed by a checklist of the key points for conducting a PIA/DPIA for big data analytics.

[248] GDPR Article 35(1)
[249] GDPR Article 35(3)(a)
[250] GDPR Article 35(7)
[251] Information Commissioner's Office. Conducting privacy impact assessments code of practice. ICO, February 2014.

Step 1	PIA COP	Identify the need for a PIA

Step 2	PIA COP	Describe the information flows
	GDPR	A systematic description of the envisaged processing operations and the purposes of the processing, including, where applicable, the legitimate interest pursued by the controller

Step 3	PIA COP	Identify the privacy and related risks
	GDPR	An assessment of the necessity and proportionality of the processing operations in relation to the purposes
	GDPR	An assessment of the risks to the rights and freedoms of data subjects

Step 4	PIA COP	Identify and evaluate privacy solutions
	GDPR	The measures envisaged to address the risks, including safeguards, security measures and mechanisms to ensure the protection of personal data and to demonstrate compliance with this Regulation taking into account the rights and legitimate interests of data subjects and other persons concerned

Step 5	PIA COP	Sign off and record the PIA outcomes

Step 6	PIA COP	Integrate the PIA outcomes back into the project plan

Step 1

PIA COP – Identify the need for a PIA

In our discussions with organisations about conducting PIAs in a big data context, an argument was made that an ethical assessment will probably already have been done by a big data team, so there would be little point in a data protection officer (DPO) 'waving a PIA at them'. This raises two key points about identifying the need for a PIA.

The first point is that, while a form of assessment (such as a general risk assessment or ethical assessment) may already have been done, when the GDPR comes into force it will be legally required, for certain big data activities, to do a DPIA that covers several specific areas (as detailed in the table above). Organisations will therefore need to ensure their existing assessment methodology addresses these areas, or amend their processes accordingly.

The second point is that identifying the need for a PIA should not rest solely with a DPO or compliance department. A DPO should be consulted throughout the PIA process (the GDPR actually requires it for DPIAs[252]) but it is very important that big data analysts are themselves able to recognise the need for a PIA at the outset.

As this paper has highlighted, while several features make big data analytics unique, it is still subject to the same data protection principles as any other processing operation. Therefore, in terms of identifying the need for a PIA, the screening questions detailed in our PIA COP remain relevant and appropriate to use by those involved in big data analytics. In particular, the following three questions are specifically relevant to big data and the DPIA requirements set out in the GDPR:

> *Are you using information about individuals for a purpose it is not currently used for, or in a way it is not currently used?*
>
> *Does the project involve you using new technology that may be perceived as being privacy intrusive?*
>
> *Will the project result in you making decisions or taking action against individuals in ways that can have a significant impact on them?*

[252] GDPR Article 35(2)

A key issue that arose from our discussions with organisations was reluctance to start the PIA process too soon. This was because there is often a lack of clarity about the direction that a big data project will take during its early stages (discussed in more detail in step 2 below). If this is the case, we would encourage big data analysts to err on the side of caution and start the PIA process as soon as possible if they can reasonably foresee that the analysis may lead to further work that would result in one of the above screening questions being answered 'yes'.

For example, an insurance company may be planning to run some unsupervised machine learning algorithms on a dataset in the hope of finding interesting correlations in the data. At the outset the insurer does not know what the potential correlations might show. But it knows that one possible outcome is an additional piece of work to adjust premiums based on the correlations. So, even though it cannot be sure this will be the case, the insurer begins the PIA process anyway. This is to ensure it is already thinking about privacy risks as the project begins to develop.

Note that there are other innovative ways to help big data analysts recognise the need for a PIA. For instance, during our discussions with organisations in the telecoms sector, one company mentioned it uses a matrix of different types of data so that an operative can easily identify which types are high risk before starting a project. Another company in the same sector highlighted the importance of having 'data champions' in departments. While not necessarily being a DPO, a data champion would know about data protection, so could help to properly identify areas of concern.

Checklist

☐ We have a DPO available for consultation on PIAs.

☐ Our big data analysts use appropriate screening questions to help identify the need for a PIA.

☐ If the direction of a big data project seems unclear, we err on the side of caution and begin the PIA process anyway.

Step 2

PIA COP – Describe the information flows

DPIA under the GDPR – A systematic description of the envisaged processing operations and the purposes of the processing, including, where applicable, the legitimate interest pursued by the controller

Discussions with organisations highlighted this step as difficult to complete in the context of conducting PIAs for big data analytics. The consensus was that describing information flows is often much harder because the discovery phase of big data analytics (thinking with data) involves finding unexpected correlations as opposed to the testing of a particular set of hypotheses. Additionally, companies in insurance and telecoms highlighted the difficulty of mapping information flows when using the Agile project management methodology.

It is clear that this step can be challenging for big data analytics, but under the GDPR it will be an explicit part of a DPIA[253]. Furthermore, if the 'legitimate interests' condition is being relied on for the processing of personal data, the GDPR requires it to be described as a part of this step. This requirement links with the new accountability principle in the GDPR which, among other things, obliges organisations to maintain internal records of their processing activities[254].

Therefore, if it's a realistic outcome of a big data project that decisions will significantly affect individuals, every effort needs to be made to observe the requirements of this step by describing the relevant information flows, the purposes for the processing and, where necessary, the organisation's legitimate interests.

Although our discussions with organisations revealed a common theme of difficulties with this step, several companies in the telecoms sector emphasised the need for clarity in the aims of data processing and the importance of having an end product in mind. This view is reflected in a paper by the Information Accountability Foundation, which refers to big data analytics beginning with a "sense of purpose" as opposed to a hypothesis[255]. We encourage organisations undertaking big data analytics

[253] GDPR Article 35(7)(a)
[254] GDPR Article 30(1)
[255] Abrams, Martin et al. A Unified Ethical Frame for Big Data Analysis. Information Accountability Foundation, 7 October 2014. http://informationaccountability.org/wp-

to think carefully about their sense of purpose for a given project, even if it may change somewhat as the project develops. This will help illuminate the potential information flows that could arise as a big data project progresses. It also complements the advice in our PIA COP about Agile project management and the description of information flows:

> "Describe the information flows as part of a user story which you can refer to while implementing the project. As the project progresses, record how each stage has changed how you use personal information."[256]

For big data projects where there are genuinely no aims or objectives at all at the outset, a potential solution may be to take the processing outside the data protection sphere by using only anonymised datasets during the discovery phase. Should correlations of any interest be discovered, the organisation would then be able to identify the aims of any further processing before starting any analysis of the original dataset containing personal data. At this point, the organisation should therefore be able to describe the envisaged information flows, the purposes for processing and, where necessary, its legitimate interests.

Checklist

☐ Where possible, we clearly describe the predicted information flows for our big data project.

If the purposes of the processing are uncertain:

☐ we use only anonymised data, or

☐ we describe the information flows as the project progresses.

content/uploads/IAF-Unified-Ethical-Frame-v1-08-October-2014.pdf Accessed 17 November 2016

[256] Information Commissioner's Office. Conducting privacy impact assessments code of practice. ICO, February 2014.

Step 3

PIA COP - Identify the privacy and related risks

DPIA under the GDPR – An assessment of the necessity and proportionality of the processing operations in relation to the purposes

DPIA under the GDPR – An assessment of the risks to the rights and freedoms of data subjects

In our discussions with organisations about this step, similar issues to those identified in step 2 were also highlighted; namely that the discovery phase of big data analytics can make it particularly difficult to identify privacy risks because, at this stage, it's not clear what the analysis might reveal.

While our PIA COP and the GDPR set out specific frameworks, a PIA should not be seen as a rigid process that restrains the progress of a particular big data project. Rather, PIAs should be treated as scalable and 'living' procedures that develop alongside a project's evolution. Therefore, as with step 2, the identification of risks can take place as the project moves forward and the potential risks become clearer.

Based on our research for this paper and discussions with organisations, we have developed the following questions that may help to identify and record the relevant risks to individuals and organisations in a big data context. This list is not meant to be complete and the questions are relatively high level and broad. Organisations should develop their own questions based on the specifics of the big data analytics they are undertaking.

- Have individuals been made aware of the use of their personal data?
- Could our analysis involve sensitive personal data – for example, in the analysis of social-media posts?
- Is the dataset representative and accurate?
- What are our retention policies for the data?
- Are the datasets held across multiple and disparate systems?

- Do the systems have appropriate inbuilt security measures?
- Does our proposed analysis involve cloud processing?
- Will a third-party organisation do the analytics for us?
- Could anonymised data be re-identified?
- Will we be able to explain the reasons behind any decisions we make that result from the big data analytics?

As regards the wording of this step for DPIAs in the GDPR, assessing "the risks to the rights and freedoms of data subjects" is largely covered by what we discussed above. However, in addition, "an assessment of the necessity and proportionality of processing operations" will also be an explicit part of the DPIA process.

For organisations involved in big data analytics, assessing necessity will mean considering whether the proposed type of analytics is the only method of achieving the project's purposes or whether another less privacy-intrusive method could be used. For instance, an organisation may need to consider why a more traditional research project (using a sample of a total population) would not be sufficient to achieve the project's objectives.

Additionally, assessing proportionality will involve considering whether the proposed analytics are justified in the circumstances. To put it another way, are the project's purposes important enough to compensate for the potentially privacy-intrusive methods to be used? For example, if a big project's objective is to target an offer to a particular group of people, does the value of the offer to that group justify the profiling of people during the application phase of the analytics?

Consultation, both internal and external, is key for a successful PIA and should take place throughout the process. We highlight it here because of the value of seeking the views of individuals in identifying privacy risks. In our discussions with organisations there seemed to be some uncertainty about the need to consult customers about big data projects. But we would encourage such consultation and remind organisations of the potential commercial benefits of increased trust and competitive advantage through transparency[257]. The Council of Europe's Consultative

[257] Del Vecchio, Steve, Thompson, Chris and Galindo, George. Trust but verify: From transparency to competitive advantage. PricewaterhouseCoopers, View Issue 13. http://www.pwc.com/us/en/view/issue-13/trust-but-verify.html Accessed 17 January 2017

Committee of Convention 108's guidelines on big data recommend the involvement of individuals and groups as part of risk assessments if the use of big data may affect fundamental rights[258]. Furthermore, under the GDPR, consultation with data subjects will be required for DPIAs in circumstances where it would be "appropriate"[259]. The GDPR does not define such circumstances, but the requirement is likely to apply in situations where data subjects will be significantly affected by the outcomes of the big data analytics.

Checklist

☐ We ask ourselves questions about the proposed big data analysis to identify and record the associated privacy risks.

☐ As the project develops we regularly return to these questions and develop new questions to identify and record any new risks.

☐ We assess whether the proposed big data analytics is the only method by which the project could be conducted.

☐ We assess whether the proposed big data analytics is justified in relation to its potential benefits.

☐ We consult internally and externally throughout the big data project.

[258] Consultative Committee of the Convention for the protection of individuals with regard to automatic processing of personal data. Guidelines on the protection of individuals with regard to the processing of personal data in a world of big data. Council of Europe, 23 January 2017. https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000016806ebe7a Accessed 6 February 2017

[259] GDPR Article 35(9)

Step 4

PIA COP – Identify and evaluate privacy solutions

DPIA under the GDPR – The measures envisaged to address the risks, including safeguards, security measures and mechanisms to ensure the protection of personal data and to demonstrate compliance with this Regulation taking into account the rights and legitimate interests of data subjects and other persons concerned

Once the questions developed as part of step 3 have helped identify all the relevant privacy risks associated with a big data project, the next step is to consider how these risks will be addressed before the analytics begin. Using the example questions above, relevant to identifying privacy risks in a big data context, we have listed some potential solutions below. These solutions are merely examples; organisations should identify and record their own list of solutions appropriate to the specifics of the big data analytics they are undertaking.

- Have individuals been made aware of the use of their personal data?
    - Yes, we provided a privacy notice at the point of collection and obtained consent to use the personal data for analysis to identify and provide relevant offers and discounts to individuals.
- Could our analysis involve sensitive personal data (for example, in the analysis of social-media posts)?
    - No. We have developed an algorithm to identify and omit all instances of sensitive personal data from the dataset before any analysis, eg references to race, ethnicity, religion and health.
- Is the dataset representative and accurate?
    - If a dataset is unlikely to be representative of the total population, we do not use the analysis results for the purposes of profiling or significant decision making.
    - We regularly check samples of the dataset with individuals to make sure it is accurate and up to date.
- What are our retention policies for the data?

- We maintain appropriate retention schedules for the datasets we use for big data analysis. These are regularly reviewed and enforced by our records management team.

- Are the datasets held across multiple and disparate systems? Do the systems have appropriate inbuilt security measures?
  - Yes. We employ information security experts to implement appropriate security measures including encryption and access controls.

- Does the proposed analysis involve cloud processing? Will a third-party organisation do the analytics for us?
  - Yes. We will make an extensive assessment of cloud providers and data analysis organisations to select those that can provide the most secure environment for the data processing. We will put contractual agreements in place regarding the security and use of the data.

- Could anonymised data be re-identified?
  - We follow the guidance in the ICO's Anonymisation code of practice to reduce the likelihood of anonymised data being re-identified.

- Will we be able to explain the reasons behind any decisions we make that result from the big data analytics?
  - Yes, we audit our machine learning algorithms to check for bias and decision-making rationale.

We recognise that several unique features of big data analytics can make it difficult to identify practicable privacy solutions that would appropriately address the risks in question. In our discussions with organisations, two particular areas of concern emerged.

First, several organisations talked about using consent as a mitigation measure. But there was uncertainty as to whether such consent could truly be 'informed' in the context of big data when an organisation may not know exactly what they will do with the data at the point of obtaining of consent. As we said in the consent section in chapter 2, rather than treating consent as a one-off transaction at the start of a relationship between an organisation and an individual, a graduated consent model could be used instead to obtain consent from an individual for new uses of

their personal data as part of an ongoing relationship with them. Thus, as the purposes of a particular big data project are defined, an organisation could then re-approach individuals for informed consent, or as part of business-as-usual activities that involve contact with its customers.

The second area of concern was about transparency and the difficulties of ensuring people understand, and therefore expect, what is happening with their personal data, given the complexities of big data analytics. Again, referring to the main paper, and in particular the privacy notices section in chapter 3, an innovative and layered approach to providing clear, concise and intelligible information about big data analytics may involve using just-in-time notifications, icons, videos, and other visual representations. These can help to explain complex concepts in an easy way. Additionally, individuals' expectations about the use of their personal data can be linked to their trust in an organisation. Trust can be fostered in several ways, but transparency is an important component[260]. An organisation should therefore be completely honest with individuals about the use of their personal data. This may even mean explaining at the outset of a relationship with an individual that the exact purposes of any data analysis may not yet be defined, but that more information will be provided when the purposes become apparent, in line with the graduated consent model.

It is useful to reiterate here that, as a fluid process, a PIA should not limit the identification of privacy solutions to a specific phase of a big data project. As a project progresses and objectives shift, new privacy risks will emerge. Organisations will need to be able to continue considering how they will address these emerging risks.

Checklist

☐ We identify and record appropriate measures to address the privacy risks previously identified.

☐ As the big data analysis progresses and new risks are identified, we continue to identify and record measures to address these risks.

☐ If the direction of a big data project is unclear, we use novel methods of obtaining consent and providing privacy notices.

[260] Morey, Timothy; Forbath, Theodore and Schoop, Allison. Customer Data: Designing for Transparency and Trust. Harvard Business Review, May 2015. https://hbr.org/2015/05/customer-data-designing-for-transparency-and-trust Accessed 16 January 2016

Step 5

PIA COP – Sign off and record the PIA outcomes

The fifth step of the PIA involves recording the process and signing off the measures identified to address the privacy risks. This is not an explicit part of the DPIA framework in the GDPR, but it links up with the new accountability principle that requires organisations to maintain internal records of their processing activities[261].

In line with our view that information security should be considered a boardroom issue[262], we recommend that sign-off for a big data PIA is sought from board level or an equivalent senior level for smaller organisations. This view was reflected in our discussions with organisations in the technology sector. They said engagement with privacy issues at board level varies but saw the need for buy-in from this level to properly address privacy risks.

In our PIA COP we state that the ICO does not take a role in approving or signing off PIAs. However, for DPIAs under the GDPR, organisations will in some circumstances need to consult the ICO before they process personal data. For instance, such consultation will be required if a proposed big data project will involve high-risk processing but the organisation undertaking the project has been unable to identify a way of mitigating the risk as part of their DPIA[263]. If this is the case, the ICO will provide written advice to the organisation within 8 weeks (or 14 weeks if the matter is particularly complex) of receiving a request for consultation. If necessary, we may also use our powers to prohibit the proposed processing operations.

Finally, as in our PIA COP, we would still encourage organisations to make their PIA reports publicly available (but with business-sensitive information redacted). This will help to increase the transparency of big data processing operations that will contribute to data protection compliance and help to build customers' trust.

[261] GDPR Article 30(1)
[262] Information Commissioner's Office. TalkTalk gets record £400,000 fine for failing to prevent October 2015 attack. ICO, 5 October 2016. https://ico.org.uk/about-the-ico/news-and-events/news-blogs-and-speeches/2016/10/talktalk-gets-record-400-000-fine-for-failing-to-prevent-october-2015-attack/ Accessed 17 January 2017
[263] GDPR Article 36, Recital 94

Checklist

☐ We obtain board-level sign-off for the measures identified to address the privacy risks of the proposed big data analytics.

☐ We keep a record of the sign-off and the whole PIA process.

☐ If we have identified high risks but not the measures to mitigate them, we consult the ICO before starting any data processing.

☐ We produce and publish a PIA report.

Step 6

PIA COP – Integrate the PIA outcomes back into the project plan

It is very important that the sixth and final step of the PIA is not forgotten. This is when the privacy solutions identified in step 4 and signed off in step 5 are actually folded back into the big data project. Organisations' compliance functions have a role here. But it is vital that the analysts actually undertaking the big data project understand the solutions, why they are necessary and how they can be implemented. This view was echoed in our discussions with insurance companies, when it was suggested that a PIA should be owned by the business as opposed to the compliance department.

This may be the last step in the PIA process, but organisations should not see it as a point from which they no longer need to consider privacy risks. Regular reviews should ensure that the privacy solutions implemented are working as expected. Furthermore, as discussed, the aims, objectives and applications of big data operations may be subject to change throughout a project's lifecycle. Regular reviews will help to pinpoint such changes and check whether the outcomes of the PIA still apply. If they don't, the earlier steps of the PIA can be revisited or a new PIA can be undertaken. Then any new privacy risks can be addressed.

Checklist

☐ We ensure that the agreed privacy solutions are folded back into the big data project.

☐ We regularly review our big data processing operations to check whether the privacy solutions are working as expected.