Title

Big Data and the Reform of the European Data Protection Framework: an Overview of Potential Concerns Associated with Proposals for Risk Management-based Approaches to the Concept of Personal Data.

Henry Pearce, University of Hertfordshire, Lecturer in Law

School of Law, Criminology and Political Science

University of Hertfordshire,

Hatfield,

Hertfordshire,

AL10 9EU

Email: h.pearce@herts.ac.uk

Keywords:

Big data, data protection, personal data, risk management, regulation

Abstract:

This article considers the emergence of the phenomenon of big data and how it poses considerable theoretical and practical difficulties for the European Data Protection framework's key enabling concept: the notion of personal data. The article starts by outlining the fact that there is an emerging body of opinion which suggests some of these problems might best be addressed by way of a shift to a risk management-based model, or models, of data protection. This, it is argued, is suggestive of the emergence of a possible fusion between the disciplines of data protection law and risk management. The article contends, however, that there are a number of severe complications associated with the adoption of risk management-based regulatory strategies, both generally and in the immediate context, which have to date, despite being widely recognised in the risk research literature, have not been meaningfully explored, and in some cases not considered at all, in the legal and regulatory literature pertaining to data protection. Consequently, these are issues in need of address.

Whilst not intending to counsel against the adoption of risk management-based regulatory strategies in the big data context, the aim of this article is begin bridging the metaphorical "gap" between legal, regulatory, and risk research and management discourses, to stoke much-needed debate in this topical area, and highlight the fact that debates about risk management-based reform of the European data protection framework should not in any way be thought of as being concluded. To this end, the article presents several areas which are in need of further consideration, and where there will likely be possibilities for future inter-disciplinary research.

Over the course of the last few years much has been written regarding the emergence of big data and how it poses considerable challenges to the effectual operation of the European data protection framework. In particular, it has been suggested that, as traditionally envisaged, the notion of personal data, the central enabling concept at the heart of European data protection law, is not fit for purpose in the emerging big data environment. A popular response from academics, regulatory bodies, and other observers, has been to suggest that the most appropriate way to respond to the challenges faced by the concept would be to move to a model, or models, of data protection regulation that could broadly be described as risk management-based.

The notion of risk, and the analysis and management thereof, are no strangers to one another. There is, for example, a wealth of risk research literature dedicated to the consideration of the identification, quantification and analysis of risk. However, a review of the relevant literature pertaining to data protection law and policy reveals that hitherto little analysis has been devoted to a number of important issues that would have to be considered were such a regulatory shift to be initiated.[1] In other words, despite a variety of observers advocating for the adoption of risk management-based models of data protection in Europe, a number of key fundamental issues well-traversed in the risk research literature have been overlooked.

This article, whilst not intending to counsel against the adoption of risk management-based approaches to data protection law and policy, highlights a number of significant issues and complications that are highly pertinent in the context of risk analysis and management but have to date been underexplored in the data protection literature and, by beginning to bridge the gap between legal and risk research discourses aims to engender much needed debate in the area.

The structure of the article is as follows. First the rationale, emergence and development of data protection law as a distinct field of legal practice at the European level is outlined. Second, the phenomenon of big data is defined, and it is examined how its emergence has caused difficulties for data protection law's key enabling concept, the notion of personal data itself. The final sections of the article are then dedicated to a consideration of suggestions that a move to risk management-based models of data protection would provide an effective and desirable way to respond to and subvert the some of these problems. In so doing the article ultimately highlights several issues that will need to be more rigorously considered, and where future research will be required, if risk management-based models of data protection law and policy are to have any realistic prospect of success.

1. The development of data protection law in Europe and the concept of "personal data"

In the mid-to-late twentieth century a range of concerns over computerised processing of information, which was becoming particularly prominent in the private sector, spawned widespread calls for fresh legislative and regulatory interventions. Particularly, these calls focused on the need to protect individuals from potential harms and abuses, put them in a position whereby they could more readily

[1] The literature considered specifically by this article is comprised of works authored by both academic and scholarly observers, as well as those produced by institutions such as the UK Information Commissioner's Office and the World Economic Forum. Notable works cited include: Schwartz, P. and Solove, D. (2012) "The PII Problem: Privacy and a New Concept of Personally Identifiable Information", *New York University Law Review*; Cavoukian, A. and El Emam, K. (2011) "Dispelling the Myths Surrounding De-identification: Anonymization Remains a Strong Tool for Protecting Privacy", *Information and Privacy Commissioner, Ontario, Canada*; Kuan Hon, W. Millard, C. and Walden, I. "What is Regulated as Personal Data in Clouds?" in *Cloud Computing Law*, ed. by Millard, C. (2013) Oxford: Oxford University Press; Aldhouse, F. (2014) "Anonymisation of personal data: A missed opportunity for the European Commission", *Computer Law & Security Review*; ICO (2012) "Anonymisation: managing data protection risk code of practice"; World Economic Forum (2012) *Rethinking Personal Data: Strengthening Trust*, available at: http://www.weforum.org/reports/rethinking-personal-data-strengthening-trust.

control how information about them was used, and to ensure that the free movement of data for economic purposes was not unduly restricted.

In response, the Council of Europe began drafting bespoke codes of practice in response to the growing perception that existing legal instruments, notably the European Convention on Human Rights, were not suitably geared towards the achievement of these objectives.[2] This period of drafting eventually culminated in the adoption of the Convention on Data Protection which contained basic principles for the processing of personal data in automated data files.[3] These principles imposed various obligations on both organisations and individuals handling large quantities of data, and outlined rights for individuals as well as arrangements pertaining to institutional oversight, enforcement, and international cooperation.

However, due to concerns over inconsistencies in the way in which the Convention had been incorporated into the national laws of Member States, and the detrimental effects this was thought to be having on inter-community trade,[4] proposals were later brought forward by the European Commission for a Directive 'On the Protection of Individuals with Regards to the Processing of Personal Data and on the free Movement of Such Data', later adopted as the Data Protection Directive in 1995.[5] The underlying rationale of the Directive, as with the Convention before it, was to fulfil the dual objectives of facilitating the free flow of data in the internal market whilst, concurrently, protecting individuals from harms that may stem from the computerised, or non-computerised, processing of data.[6]

In order to meet these objectives, the Directive, which at the time of writing remains the EU's primary legislative instrument concerning the regulation of data-handling practices, defines personal data as '*any information relating to an identified or identifiable natural person'*,[7] and subjects the processing[8] thereof to a range of substantive rules. Article 6 of the Directive, for instance, contains a series of provisions relating to the quality of personal data; notably that personal data must be accurate, kept up to date, collected for a specified legitimate purpose, not be excessive for the purposes for which they were collected, and processed fairly and lawfully in a way that is not incompatible with the original specified legitimate purpose.[9] Article 7 then provides an exhaustive list of circumstances under which data can be processed fairly and lawfully.[10]

Unfortunately, like the Convention before it, the Directive's existence has been troubled and wrought with complication. Though the Directive's objective was to ensure an equivalent level of data protection

---

[2] Bygrave, L. (2014) *Data Privacy Law: An International Perspective*, Oxford: Oxford University Press, pp.12-24, and Hustinx, P. "The Reform of EU Data Protection" in *Emerging Challenges in Privacy Law* ed. by Witzleb, N. *et al.* (2014) Cambridge: Cambridge University Press. Pp.62-71.

[3] Officially known as the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, January 1981, CETS no.108.

[4] Generally speaking, the lack of consistency between the data protection regimes of different EU Member States was identified as having made it difficult and onerous to transfer data across borders in a way that was legally compliant. In turn, this was thought to be inimical to the development of economically and socially advantageous data processing activities. For an overview of these issues, see: Lloyd, I. (2017) *Information Technology Law*: Oxford: Oxford University Press, pg.40.

[5] Directive 95/46/EC. [hereinafter, the Data Protection Directive]

[6] Perhaps ironically, however, some observers have suggested that the evolution of data protection law at the European level has acted as major obstacle to free international data flows, rather than a facilitator. See: Kong, L. (2010) "Data Protection and Transborder Data Flow in the European and Global Context", *The European Journal of International Law* 21(2).

[7] Article 2 (a) Data Protection Directive.

[8] The Directive defines processing broadly, as '…any operation or set of operations which is performed upon personal data, whether or not by automatic means, such as collection, recording, organization, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction'. See: *Ibid*., Article 2 (b)

[9] *Ibid.,* Article 6

[10] *Ibid*., Article 7 (a)

within the EU, a considerable divergence in the way its provisions have been applied across member states has been observable throughout its lifespan.[11] As a consequence, a not insignificant number of data controllers, particularly those operating online, have been left in a position of having to potentially deal with 28 different national data protection laws, creating a fragmented legal environment rife with uncertainty and unequal protection for individuals.[12] Moreover, advancements in information technology, notably the emergence of new data mining and analytical technologies, have given rise to fresh concerns in respect of the Directive's efficacy.

These anxieties led to the European Commission presenting a package of proposals designed to update and modernise the European data protection framework in January 2012. Significantly, the package contained a proposal for a General Data Protection Regulation[13] to replace the apparently outdated Directive. This proposal was subject to intense discussion before trilogue discussions between the European Commission, Parliament, and Council were concluded in December 2015, resulting in an agreed final text which will apply in all EU Member States from May 2018.[14]

This drive to revise the data protection framework should have represented an excellent opportunity to re-assess the central tenets of data protection law, and undertake a challenging exercise in how to develop a new structure by which modern data-handling practices could be regulated. However, despite the apparent intention of European lawmakers to modernise the data protection framework, the text of the General Data Protection Regulation still retains, and places a great deal of emphasis on, many of the core concepts that were integral to the Directive before it, notably the notion of personal data itself. Using an approach similar to the Directive before it, the Regulation defines personal data as "any information relating to an identified or identifiable natural person",[15] and makes it clear that the processing thereof is subject to a number of substantive rules and conditions.[16]

From this brief overview of key EU legislation in the data protection field we are immediately able to identify the data protection framework's key enabling concept: personal data itself. The Directive and Regulation alike are primarily concerned with the regulation of the processing of personal data, and the key substantive rules and provisions of each are exclusively engaged in situations where data processing operations involve data that are "personal". Conversely, the same substantive rules and provisions are not engaged in situations where no personal data are involved. In so doing, the data protection framework appears to operate under the assumption that there exists a hard dichotomy between data which are personal and those which are not.

The continued reliance on the concept of personal data in this key central role in the data protection framework seemingly reflects the belief of European lawmakers that the concept of personal data remains a fundamentally useful construct around which to build a regulatory framework for contemporary data-handling activities. The continued faith being placed in this concept, however, should give us pause. Not only has the notion of personal data proved troublesome to date, but there is

[11] For instance, in the Analysis and impact study on the implementation of Directive EC 95/46 in it was revealed that, amongst other divergences, the national data protection regimes of Member States contained inconsistent and contrasting definitions of a number of key terms, notably "data controller" and "consent". *Analysis and Impact Study on the implementation of Directive EC 95/46 in Member States.* Available at: http://ec.europa.eu/justice/policies/privacy/docs/lawreport/consultation/technical-annex_en.pdf

[12] Reding, V. (2012) "The European data protection framework for the twenty-first century", *International Data Privacy Law*.

[13] Officially known as Regulation (EU) 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L119/1

[14] Available at: http://static.ow.ly/docs/Regulation_consolidated_text_EN_47uW.pdf

[15] Article 4(a) General Data Protection Regulation.

[16] See: Arts 4-11 General Data Protection Regulation.

reason to suspect that, as traditionally envisaged, it may be entirely unfit for purpose in an age of big data.

2. Big Data and its implications for the concept of personal data

Generally speaking, big data can be considered a loosely defined term, which is broadly used to describe datasets that are so large and complex they have become awkward to work with using standard statistical software, or data which are too large to be stored, managed, or analysed in a single organisation.[17] Essentially, the existence of such datasets is made possible by the unprecedented, and exponentially increasing, amount of data produced and put into circulation in the world today.[18]

As noted by boyd[19] and Crawford, however, big data is in many ways a poor term. In their words, there is little doubt that the quantities of data now available in the world are often huge, but that is not the defining feature of this new data ecosystem.[20] Big data is less about data that are big, and more about increased capacities to search, aggregate, and cross-reference large datasets.[21] A better, or at least more nuanced way of defining big data, therefore, might be to consider it to be a cultural, technological, and scholarly phenomenon that rests on the interplay of:

> '*1. Technology: maximizing computation power and algorithmic accuracy to gather, analyse, link and compare large datasets.*
>
> *2. Analysis: drawing on large data sets to identify patterns in order to make economic, social, technical and legal claims.*
>
> *3. Mythology: the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy.*'[22]

As alluded to by boyd and Crawford, working with datasets of such extraordinary size and scale allows for those in possession of sophisticated analytical tools to identify patterns, and make inferences and predictions that would not have previously been possible when working with datasets of a smaller size. Significantly, this may often entail data that have been collected for a specific purpose being repurposed to serve an entirely different end.

In other words, big data are large volumes of inter-connected data that can be stored, processed, and shared in ways that allow us to develop a richer, deeper analysis and picture of what those data represent.[23] The types of scenarios which would fall within these categories can range drastically, from data derived from vast international science projects the Large Hadron Collider of the European

---

[17] Snijders, C. et al. (2012) "Big Data": Big Gaps of Knowledge in the Field of Internet Science", *International Journal of Internet Science*, pg.1.

[18] For instance, *The Economist* reported in its 2013 Outlook that the quantity of global digital data expanded from 130 exabytes in 2005, to 1,227 in 2010, and will have increased to 7,910 by 2015 – to highlight the enormous quantity of this amount of data in lay terms, the 1,227 exabytes of data, if stored on DVDs, would require a fleet of 16 million Boeing 747 aircraft in order to transport it globally. The Economist (last accessed August 2014) "Welcome to the yotta world", http://www.economist.com/node/2153792

[19] In accordance with her wishes, this article refers to danah boyd using lower case letters only.

[20] Some of the data encompassed by big data, for instance trending Twitter posts, will not be as large as earlier datasets that we not considered big data, such as national censuses.

[21] boyd, d. and Crawford, K. (2012) "Critical Questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon", *Information, Communication and Society*

[22] *Ibid*., pg.663.

[23] Williamson, A. (2014) "Big Data and the implications for government", *Legal Information Management*.

Organization for Nuclear Research, to the wealth of data collated by online companies such as Facebook and Google.[24]

Big data is, therefore, gathered from a vast, and constantly increasing, array of sources, ranging from data gathered online through the use of services like social networking sites and search engines, credit and debit card activity, scientific and medical experiments, the use of financial services, and increasingly, the use of sensor networks, surveillance cameras, and the rollout of pervasive computing technologies, often referred to as the Internet of Things.[25] Moreover, not only have analytical capacities soared, but they have also become far more inexpensive and widely distributed. Modern mobile devices such as telephones and tablets possess more computing power than the desktop computers of a decade ago, and it is now possible to link data and devices virtually so that huge computational tasks can be undertaken affordably and conveniently.[26]

Whilst big data's latent value is seemingly huge, and its uses are likely to lead to a significant number of benefits, both social and economic in nature, it has also been noted that certain big data analytics operations may also be capable of causing harmful and undesirable consequences for individuals, and thus they require regulation. In particular, automated-algorithmic profiling, a key constituent part of many big data analytics operations, has been identified as having the potential to lead to discriminatory practices and, as a result, the diminution of individual autonomy.[27] An in-depth examination of these issues is beyond the scope of this article. What is significant to this article, however, is the fact that many big data analytics operations, regardless of their consequences, will involve the processing of personal data and so will fall under the jurisdiction of the European data protection framework. As alluded to above, however, there are a number of difficulties that necessarily rise when attempting to apply the concept of personal data itself in the big data environment.

As outlined above, the concept of 'personal data' is of fundamental importance to the European data protection framework. This is reflected in the texts of both the Data Protection Directive and the imminent General Data Protection Regulation, the substantive provisions on which apply exclusively to data which are personal, and not to those which aren't. This is made clear in the Directive's recitals where it is specified that the principles of data protection apply exclusively to data which are 'personal', and are categorically not applicable to data which have been rendered anonymous in such a way that an individual cannot be identified or considered to be identifiable by all means 'likely reasonably to be used' by any data controller or other person who sought to make such an identification. [28] This provision

---

[24] Fishleigh, J. (2014) "A non-technical journey into the world of Big Data: an introduction", *Legal Information Management*

[25] Rubinstein, I. "Big Data: A Pretty Good Privacy Solution" Future of Privacy Forum and the Stanford Center for Internet & Society's "Big Data and Privacy: Making Ends Meet" Workshop 2013; Kuner, C. et al. (2012) "The challenge of 'big data' for data protection", *International Data Privacy Law* 2(2), pg.47.

[26] *Ibid*

[27] See: Hildebrandt, M. (2008) "Profiling and the Rule of Law", *IDIS*, pp.55-70; Schiller, B. (2014) "First Degree Prince Discrimination Using Big Data", *Brandeis University*; Cumbley, R. and Church, P. (2013) "Is "Big Data" creepy?", *Computer Law & Security Review* 29(5); Magnani, L. "Abducing personal data, destroying privacy: diagnosing profiles through artificial mediators" in *Privacy, Due Process and the Computational Turn: The philosophy of law meets the philosophy of technology,* ed. by Hildebrandt, M. and de Vries, J. (2013) Oxon, Routledge; Hildebrandt, M. (2008) "Profiling and the Rule of Law", *Identity in the Information Society*, pp.55-70; Bollier, D. (2010) "The Promise and Peril of Big Data", *The Aspen Institute*; Tene, O. (2010) "Privacy: The new generations", *International Data Privacy Law*. It should also be noted that big data analytics may also pose risks to those whose personal data are not swallowed up by these sorts of activities. There remain billions of people in the world today who remain on the margins of such activities and their inherent implications, simply because they do not routinely engage in activities that advanced analytics are designed to capture. It has been suggested that such individuals, despite their non-participation, may still find themselves subject to new forms of inequality and subordination. On this issue, see: Lerman, J. (2013) "Big Data and its Exclusions", *Stanford Law Review Online*.

[28] Recital 26 Data Protection Directive

is again mostly retained in the recitals of the General Data Protection Regulation, which specify that the principles of data protection should only apply to information concerning an identifiable or identified natural person, and that when attempting to determine whether a person is identifiable account should be taken of all the means reasonably likely to be used to identify them.[29]

As alluded to in the article's previous section, however, the concept of personal data has endured a troubled existence. Reaching an understanding as to precisely which types of information fall within its definition, for instance, has proved problematic for courts,[30] and much academic and regulatory ink has been devoted to its elucidation.[31] However, the emergence of big data and the possibilities inherent in its associated analytical operations has fundamentally cast considerable doubt on the continued suitability of the law making a distinction between data which are personal data and those which are not, and treating these categories of data differently.

The original rationale for this distinction appears to have been an underlying belief held by the drafters of early data protection legislation that data which are non-personal cannot be related to an individual, and therefore the processing of such data will not entail any threats to any specific person and so should not be subject to the full remit of data protection law's substantive rules. Concurrently, anonymisation techniques – processes by which data can be manipulated to make it more difficult, or effectively impossible, to link individual persons to whom they relate – have traditionally been widely used by data controllers to render data they have at their disposal anonymous, allowing them to share, analyse, disclose or sell those data without engaging data protection rules

Advances in re-identification methods made possible by the emergence of big data, however, have meant that the suggestion that it is impossible for individuals to be identified, and suffer resultant harms, when identifying information is removed from a dataset has effectively lost all scientific basis. Seminally, in 2008 Arvind Narayanan and Vitaly Shmatikov of the University of Texas demonstrated that by applying their bespoke de-anonymisation methodology to the Netflix Prize dataset,[32] which contained anonymised film ratings of 500,000 Netflix subscribers – the world's largest online film rental service – it would be possible for an adversary who possessed minimal information about an individual subscriber to easily identify said subscriber's record in the dataset. By using the Internet Movie Database[33] as a source of background knowledge, Narayanan and Shmatikov were able to successfully identify a number of the Netflix records of known users, and, as a result, uncovered their apparent political preferences and other potentially sensitive information.[34]

More recently, in a 2010 study, it was demonstrated that information about the membership of groups on social networking sites (i.e. information regarding Facebook groups, and similar, to which an individual user belongs) will often be sufficient to uniquely identify an individual, or at the very least identify a category of people in which the individual can be pinpointed. Building on the abovementioned work of Narayanan and Schmatikov, which showed that statistical methods can be used to de-

[29] Recital 26 General Data Protection Regulation

[30] See, for instance: *Durant v Financial Services Authority* [2003] EWCA Civ 1746

[31] See, for instance: Millard, C. and Kuan Hon, (2012) "Defining 'Personal Data' in e-Social Science", *Information, Communication and Society;* Welfare, D. (2012) "Clarifying the scope of personal data", *Privacy & Data Protection.* See also: Article 29 Data Protection Working Party, Opinion 4/2007 on the Concept of Personal Data, WP 136; United Kingdom Information Commissioner's Office (2007) "Determining what is personal data".

[32] The Netflix Prize was an open competition held by Netflix, where entrants were provided an anonymized dataset of over 100,000,000 film ratings, and invited to improve on Netflix's film recommendation algorithm, without any other information about Netflix's users being provided. See: Netflix (accessed June 2014) "The Netflix Prize Rules", http://www.netflixprize.com//rules

[33] The Internet Movie Database is an online database containing information relating to films, and other media. www.imdb.com

[34] Narayanan, A. and Shmatikov, V. (2008) "Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset)", *University of Texas at Austin.*

anonymize data through the cross-correlation of datasets, Wondracek et al showed that the identification of a particular individual could be made using only information from a single social networking site and combining it with the browsing history of a user, without the need to correlate it with other auxiliary information.[35] Even more recently, additional studies have shown that, due to human mobility traces being highly unique, even supposedly anonymised location data can be linked to a particular individual with relative ease if correlated with a piece of outside information.[36]

The inference to be drawn from these findings is that, quite simply, the use of anonymisation techniques and the removal of all obvious personal identifiers from datasets will, in many instances, not prevent individuals from being identified and, as a result, will not equate to the potential harms inherent in contemporary data processing activities being solved by technology rather than law or other regulatory interventions. In simple terms, in a world of big data, anonymisation techniques will often fail to render data truly anonymous, as there are quite simply too many data 'out there' online, in publicly available repositories, which can be used to link specific identities to records in datasets that have supposedly been anonymised through the use of conventional techniques.

The challenges posed to the law by the limits of anonymisation techniques were first brought to the attention of the legal world in 2009 by Paul Ohm in his influential article, 'Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization'.[37] Here Ohm outlined how, as outlined above, the European data protection framework implicitly embraces the assumption that anonymisation provides a 'silver bullet' against harms stemming from the processing of data, and suggested that, given the ease with which individuals can now be identified from data which have supposedly been anonymised, and since theoretically all data can now be related to an identifiable individual, the processing of non-personal data should theoretically be subject to the same conditions and legal safeguards as the processing of data which are personal.

In Ohm's eyes, therefore, the hard dichotomy prevalent in data protection legislation surrounding the treatment of data which are personal and those which are not should, therefore, be abandoned since this is a distinction which has become severely and irreversibly blurred.[38] Though Ohm's line of argument is not universally accepted,[39] it seems now that a consensus has emerged between scholars and policymakers that enough work has been done to dispel the once widely-held idea that anonymisation is a panacea to the risks associated with contemporary data-handling practices.

When introducing the Data Protection Directive in 1995, by defining 'personal data' as any information that could be related to an individual, either directly or indirectly, EU lawmakers seemingly envisioned items like documents with an identification number, or other small pieces of information that could be used to link data to the identity of a specific individual. Today, in a world of big data, this definition evidently encompasses far more information than could ever have been imagined at the point of the Directive's inception approximately twenty years ago.[40]

Nevertheless, despite these complications, as noted above, the impending General Data Protection Regulation retains the concept of personal data as its central tenet, and once again subjects the

---

[35] Wondracek, G. et al. (2010) "A Practical Attack to De-Anonymize Social Network Users", *IEEE Symposium on Security and Privacy*. pg.237.
[36] See: de Montjoye, Y. (2013) "Unique in the Crowd: The privacy bounds of human mobility", *Scientific Reports*.
[37] Ohm, P. (2009) "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization", *UCLA Law Review*, pg.1707.
[38] *Ibid*
[39] Some notable observers continue to put faith in anonymisation as being a vital tool in the quest to protect individuals against harms and abuses associated with the processing of data. See, for example: Cavoukian, A. and El Emam, K. (n 1).
[40] Tucker, P. (2013) "Has Big Data Made Anonymity Impossible?", *MIT Technology Review*, Kuan Hon, W. Millard, C. and Walden, I. (n 1).

processing thereof to its substantive rules. The wording of the Regulation makes it clear that personal data is to remain a binary concept, and presupposes the existence of a conclusive divide between data which are personal and those which are not. This is clearly problematic as there are now notable reasons to doubt that in its traditionally envisaged form the notion of personal data remains a suitable construct upon which to base a wide-ranging legislative framework for the regulation of big data analytics. Accordingly, the questions raised by Ohm regarding the continuing suitability of laws framed around a dichotomy over data which are personal, and those which are not, remain pertinent and in need of address.

3. A risk management-based approach to the concept of personal data?

The previous sections of the article have outlined how, despite big data posing apparently severe problems to the operation of data protection law's key concept: personal data, European lawmakers apparently remain strongly committed to its continued use in both current and prospective data protection regulatory regimes. Evidently, however, personal data is a concept that can no longer function as traditionally envisaged. Accordingly, it seems beyond doubt that fresh approaches to its understanding and application will be required if it is to have greater meaning and practical value big data environment.[41] Whilst a number of varied proposals have emerged in respect of how the notion of personal data might be refined or adapted to render it more fit for purpose in this context,[42] the suggestion that the abovementioned difficulties might best be addressed by way of a shift to risk management-based approach to the concept is becoming increasingly prominent.

Generally speaking, risk management-based approaches to regulation can be described as regulatory strategies that involve the targeting of enforcement and resources on the basis of assessments of the risks that a particular regulated activity poses to the regulator's objectives.[43] The key components of these assessments will be the evaluations of the risks of noncompliance and calculations pertaining to the impact that said noncompliance may have on the regulatory body's ability to achieve its objectives. In its idealised form, therefore, risk management-based regulation offers an evidence-based means of targeting the use of resources and of prioritising attention to the highest risks in accordance with a transparent, systematic, and defensible framework.[44]

Over the course of at least the last ten years there has been an observable rise in regulators adopting risk management-based frameworks of supervision worldwide, particularly with respect to regulatory issues regarding the environment, food safety, occupational health, and financial services.[45] As alluded to above, however, in accordance with the rise of big data we are now seeing the emergence of proposals for risk management-based regulatory strategies to be adopted as a means of repairing the areas of European data protection law that have come under strain. In particular, calls for the adoption of a risk management-based approach to the concept of personal data have become prominent.

For instance, in response to the difficulties associated with the concept of personal data which, as highlighted above, were outlined by Ohm in 2009, Cavoukian and El Emam argue that despite anonymisation's shortcomings, it may in some instances remain an effective mode of concealing

---

[41] On this issue generally, see, Pearce, H. (2016) "A systems approach to data protection law and policy in a world of big data?", *Computer and Telecommunications Law Review* 22(4).
[42] See, for instance: Pearce, H. (2015) "Online Data Transactions, Consent, and Big Data: Technological Solutions to Technological Problems?", *Computer and Telecommunications Law Review* 6, pp.149-153.
[43] Black, J. (2008) "Risk-based regulation: choices, practices and lessons learnt" in *Risk and Regulatory Policy: Improving the Governance of Risk*, OECD: Paris, pp.185-224.
[44] Black, J. and Baldwin, R. (2010) "Really Responsive Risk-Based Regulation", *Law & Policy* 32(2)
[45] Black, J. (n 43). It is also worth noting that over the last two decades there has been a huge quantitative growth in references to risk in the academic literature, with the number of conferences, courses, centres and journals focusing on, or making use of, the word risk expanding rapidly since the early 1990s. Durodie, B. (2005) "The concept of risk" *Nuffield Trust Global Programme on Heath, Foreign Policy and Security*.

individuals' identities. In their words, Ohm's earlier work on anonymisation merely highlighted the fact that one cannot look to such techniques in the hope of securing absolute guarantees of anonymity.[46] Precisely when the use of such techniques will be effective at rendering data truly anonymous will depend on the consideration of a variety of factors. In sum, in the emerging big data environment, whether information and data can be used to identify, or re-identify, an individual will depend on technology and practices that entail the linking of supposedly anonymised data with information and data that already identify an individual. Moreover, as additional pieces of data that identify, or are capable of identifying, individuals become available (e.g. as individuals make accessible more and more of their details in public locations, such as social networking sites) it will become ever more simple to link them to anonymised data, because there will likely be more elements in common that can be used to make a match.

Identifiability, therefore, will depend on context. Accordingly, it is unrealistic to rely on the hollow premise that there exists a sharp divide between data which are intrinsically personal and those which are not.[47] For instance, an individual Google search for 'poodles' may not be enough to identify an individual, but a collection of queries when cross analysed, or a highly specific search, might be. At some point the identification of an individual will be possible from most types of data, but when will depend entirely on circumstance.[48]

Rather than discarding personal data as a regulatory construct, recognising that data protection law does not require an absolute guarantee of non-identifiability for data to fall outside the definition of personal data, Cavoukian and El Emam argue in favour of a reshuffle of the existing legislative framework to a model of personal data based on the idea of a risk of harm continuum rather than a hard dichotomy between personal and non-personal data. Under this heading data would only constitute personal data, and thus the processing thereof would only become subject to regulation, if there was a significant risk of identification and/or harm. A shift to such a model would allow for the concept of personal data to be retained as an important regulatory construct, but would circumvent some of the problems associated with its traditional binary nomenclature that have come about due to the emergence of big data.

Support for a shift to a risk management-based model of personal data can be found not only in the work of Cavoukian and El Emam, but in a number of other sources. For instance, in its 2012 report, entitled *Rethinking Personal Data: Strengthening Trust*,[49] the World Economic Forum proposed that rather than continuing to frame data protection laws around a binary black and white dichotomy, whereby data are either categorically personal or non-personal, it would be preferable to shift to a new paradigm of data protection completely, based on the idea of context and risk. Rather than focusing on data itself, data protection laws constructed on this premise would take a contextual and risk-based approach to the different uses of data and how linked the usage is to a particular individual before determining whether the processing of such data should be subject to legal rules.[50]

The anonymisation code of practice, published by the UK Information Commissioner's Office[51] in November 2012 advocates a seemingly similar approach. Noting that the substantive provisions of data protection law do not require an absolute guarantee of non-identifiability to avoid being invoked, the Information Commissioner advises that data should only be considered personal when the risk of identification is more than reasonably likely. Though the code notes that making determinations over

---

[46] Cavoukian, A. and El Emam, K. (n 1). See also: Knight, A. Saxby, S. and Pearce, H. "Piercing the Anonymity Veil: Re-identification Risk and the UK Transparency Agenda" in *Information Ethics and Security: Future of International World Time*, ed. by Kierkegaard, S. (2014) IAITL, pp.1-33.
[47] Schwartz, P. and Solove, D. (n 1). pp.1847-1848.
[48] *Ibid*
[49] World Economic Forum (n 1).
[50] *Ibid*
[51] The Information Commissioner's Office is the UK's independent regulatory body which deals with matters regarding privacy and data protection. [hereinafter the ICO]

whether data ought to be considered personal, or sufficiently anonymised so to render them non-personal, will in many instances be far from straightforward, it sets forth a number of good practice recommendations for data handlers. To assess the risk-of re-identification, the ICO recommends that organizations evaluate whether any other person could identify an individual from the anonymised data, either by itself or in combination with other available information, at present or in the future. When considering 'other information', the Code recommends organisations adopt a 'motivated intruder' test, which would entail asking whether an individual, or individuals, could be re-identified by someone who is "*reasonably competent, has access to resources…and would employ investigative techniques such as making enquiries of people who may have additional knowledge of the identity of the data subject*."[52]

Numerous leading scholars have also added their voice the claim that a risk and context-based approach to personal data should be formally recognised and reflected in legal reform. For instance, Kuan Hon et al, whilst noting that current data protection law already requires data controllers to undertake a consideration of specific circumstances when assessing which data are personal and which are not, suggest assessing the risks of identification and harm posed by a particular processing operation would not seem a great deal more burdensome than determining whether that information is 'personal'.[53] They note that the definition of personal data is currently the single trigger for applying all of the law's requirements but, on this definition, given the abovementioned scientific advancements in re-identification techniques, almost all data could qualify as such, highlighting the limits of the personal data concept in its current framing. A more suitable and practical approach, they suggest, would be to consider data to be personal data only when there is a realistic risk of an individual being identified or harmed from those data being processed. They go on to suggest a two-stage test: a technologically neutral, accountability-based approach to address the types of concerns the personal data concept is intended to address. First, appropriate technical and organisational measures should be taken to minimise identification risk. Only then, if the resulting risk of identification remains sufficiently high, should data be considered personal data and trigger data protection obligations. Second, the risk of harm and its likely severity should then be assessed and appropriate measures taken.[54]

Aldhouse agrees, and suggests that the law should not speak of any firm divide between identifiability and non-identifiability. Furthermore, in his view, the law should be framed in terms which dictate that the application of its principles must be driven by considerations such as the nature of the data in question, the purposes for which it is to be used, the possible motivation of an adversary, the means available to such a person, and the extent of the harm the data subject would experience as a result of identification.[55]

From this overview of some important responses to the challenges posed by big data to the notion of personal data, it is evident there is a growing appetite from both regulatory bodies and other observers for a risk management-based approach to the concept. The key driver of these suggestions appears to be the underlying belief that the substantive rules of the European data protection framework should only be engaged when there is a clear risk of individuals being identified, or harmed, as a consequence of certain data being processed. As a corollary, the proposals for reform considered above seemingly suggest that said substantive rules should not be engaged when the risk of individuals being identified or harmed as a result of certain data being processed is negligible. Effectively, therefore, the adoption and enactment of these proposals would effectively render the enforcement of data protection law an exercise in risk management. Accordingly, they can be grouped together and defined as risk management-based approaches to data protection law and policy. Given the way in which the adoption of risk management-based models of data protection would apparently nullify and dissolve some of the

[52] ICO (n 1). pp.22-23.
[53] Kuan Hon, W. Millard, C. and Walden, I. (n 1). pg.187.
[54] *Ibid*
[55] Aldhouse, F. (n 1). pg.418.

problems currently observable in the data protection framework, they seemingly have considerable potential.

There, is, however, a problem. From the proposals for a shift to risk management-based model of personal data considered above it is possible to infer a possible presumption on behalf of the respective proponents that the construction and administration of risk management-based models of regulation in this vein would be relatively straightforward. Risk management is, however, a discipline in its own right, and boasts a wealth of academic literature and research geared towards the understanding of difficulties inherent in the identification, evaluation, and minimisation of risks arising from complex situations.

The remainder of the article is dedicated to a consideration of some severe difficulties inherent in the construction and application of a risk management-based models of personal data, and argues that the proponents thereof have too-readily eschewed, or have otherwise failed to consider, a number of complications and challenges that have been well documented in the risk research literature. In so doing, it attempts to bridge the gap between legal and risk research disciplines, and identify where further debate and research is needed in this topical area.

4. Risk management-based approaches to the concept of personal data: a risky regulatory approach?

The previous section of this article highlighted the fact that there is evidently a growing body of opinion which advocates for the adoption of a risk management-based approach to the concept of personal data as a means of responding to big data's emergent challenges. This is perhaps reflective of the general global trend of regulators adopting risk management-based strategies to manage their resources and reputations that has been observable for at least the best part of a decade.[56]

As alluded to above, however, there are several challenges inherent in the adoption of risk management-based regulatory strategies that, despite being well-traversed in the risk research literature, have not, to date, been rigorously addressed by those advocating a regulatory shift in the manner outlined above. These challenges can generally be separated into two categories: challenges of design, and challenges of enforcement and application.

4.1. Challenges of design

The first challenges encountered by regulators and lawmakers when attempting to devise any risk management-based model of personal data will occur in the design stage. Primarily, arguably the most significant challenge associated with the design of risk management-based regulatory regimes in general terms is the fact that, as noted above, such frameworks necessarily require an assessment, or assessments, of any possible harms that may emanate from the activity, or activities, that are the target of regulation, and the likelihood of them occurring.

Risk management-based approaches to regulation presume, therefore, that regulators will be well positioned to comprehensively discern the level of risk inherent in the activity or object that is to be regulated. In some scenarios this will be straightforward as there may, for example, be a high number of incidents from which data on the risk of any potential harm arising from a particular situation can be assessed. In other scenarios, however, regulators will be confronted with situations and events from which reliable probabilistic calculations cannot easily be drawn, or with conditions of uncertainty in

[56] Black, J. and Baldwin, R. (n 44). pg.182.

which the risk in question is inherently insusceptible to probabilistic assessment.[57] This is a point that has been well-noted in the risk research literature.[58]

Significantly, however, it is likely that many data processing activities which take place in the big data environment are likely to fall within the latter category. In other words, determining the level of risk of harm inherent in a particular big data analytics operation will in numerous instances be extremely complicated. Firstly, given the inherently uncertain nature of big data, in many circumstances it will be difficult, if not impossible to indicate what the potential harms associated with a particular processing operation, and the severity thereof, might be.[59] Secondly, even if the nature of such harms were calculable, determining the likelihood of them *actually* occurring could in itself also often be tremendously challenging.

Despite expressly advocating in favour of a shift to a risk management-based model of data protection, this is a point that is acknowledged by the abovementioned ICO anonymization code of practice:

> '*…the risk of re-identification through data linkage is essentially unpredictable as it can never be assessed with certainty what data is already available or what data may be released in the future.*'[60]

This dearth of knowledge or understanding will have obvious, but potentially severe, implications for those wishing to undertake an analysis as to whether a particular data processing activity carries with it, or does not carry, a risk of harm and the likelihood of that harm occurring.

For instance, when attempting to adopt the ICO's motivated intruder test in order to determine whether data should be considered "personal", as outlined above, questions over what 'other information' might be 'out there' capable of identifying an individual, as well as how many 'motivated intruders' exist – and what their desires, capabilities and background knowledge are – both now and in the future, will in many instances be answerable only on the basis of conjecture. Particularly, by asking data controllers to consider what information might be 'out there' at a later point in time, the ICO effectively asks that that data controllers predict the future and, as is widely accepted in general discourse, the future is inherently unpredictable; the course of technological evolution, and the effect it has had on data protection law and policy to date, should be sufficient to illustrate this point.

In any event, despite acknowledging that the level of risk associated with data processing in the big data environment will often be unknowable, the ICO nonetheless invites data controllers to undertake such assessments themselves. As has been suggested elsewhere, however, such is the complexity of making such calculations, in reality the only people who will have a chance of making an accurate assessment over the true risks associated with big data analytics operations are expert statisticians.[61] This in itself is troublesome. Basing a regulatory framework on the basis of risks that can likely be determined, much less understood, only by those with considerable expertise would surely be ill-advised.[62] It is one thing

---

[57] *Ibid*., pg.184.

[58] See, for instance: Aven, T and Renn, O. (2009) "On risk defined as an event where the outcome is uncertain", *Journal of Risk Research* 12(1), pp.1-11.

[59] Narayanan, A. Huey, J. and Felten, E. "A Precautionary Approach to Big Data Privacy" in *Data Protection on the Move* ed. by Gutwirth, S. et al. (2016) New York: Springer.

[60] ICO (n 1). pg. 18.

[61] Alexin, Z. (2014) "Does fair anonymization exist?", *International Review of Law, Computers & Technology* 28(1)

[62] This point was famously recognised by the American legal philosopher, Lon Fuller, who, in his seminal work, *The Morality of Law*, published in 1964, argued that the rules of any 'good' system of law must be understandable to those to whom they apply. Fuller, L (1969) *The Morality of Law*, New Haven, Connecticut: Yale University Press. It might also be argued that the existence of a system of law which had non-understandable rules would also be contra to the principle of legality and the rule of law itself, both of which

for the terms of such a law to be understandable by experts, and quite another for it to be understood by all the individual data controllers to whom such a law would apply. As has already been widely noted in the literature, one of the primary faults with the existing European data protection framework is its scale and complexity.[63] Shifting to a model of regulation which would invite data controllers to determine which data are likely to cause harm or lead to the identification of a person on the basis of a risk that is essentially unknowable to anyone without expertise in the fields of statistics and computer science would surely only exacerbate these existing problems.

Whilst these issues will surely prove to be extremely problematic, however, they may not necessarily prove to be fundamental. The very nature of risk management, after all, is to ensure that uncertainties, such as those described above, do not unduly impede societal or organisational goals. In other words, risk analysis and risk management are theoretically designed to deal with this exact kind of situation. In addition to the abovementioned design issues, however, what is perhaps more, or at least equally problematic in the context of proposals for a shift to a risk management-based approach to personal data is that its proponents appear, for the most part, to have overlooked the fact that there is more than one prominent methodology that can be used to identify and analyse risks, each with their own strengths and weaknesses. These different approaches and methodologies, as is noted in the risk research literature, have been extensively empirically tested in practice.[64] Acknowledgement of, much less any consideration of which, if any, of the prominent models of risk analysis and management would be suitable for use in the context of a risk management-based approach to personal data is something that is notably absent from the relevant data protection literature. This is troubling, as surely if risk management-based approaches to data protection are to have any success the determination of which methodology for assessing risk is appropriate in this context is something that needs to be considered in depth from their very outset.

In any event, difficult as they may be, deliberations regarding the identification and quantification of risks in the big data context may only be the tip of the proverbial iceberg so far as designing risk management-based approach to personal data may be concerned. Another highly salient question that would also have to be confronted by regulators as a part of the design process of such an initiative, but to date has been paid scant attention, would be how best to determine the regulatory appetite for risk; that is to say, even if the level of risk associated with a particular big data analytics operation *could* be accurately assessed, the above difficulties in doing so notwithstanding, what type of risks ought to be considered tolerable, and to what extent? In addition to this there are broader questions to be asked in respect of *who* should be making these determinations. Should, for instance, the acceptable level of risk be determined by politicians and lawmakers, regulators, or individuals themselves? These questions will likely prove difficult to answer, and should not be hastily eschewed or dismissed.

As noted by Black and Baldwin, for instance, that it is normally for regulators and policymakers themselves to set an acceptable level of risk, and that in practice a regulator's risk tolerance will often ultimately be driven by a variety of different considerations:

---

emphasise the importance of laws being clear and ascertainable. See: Le Sueur, A. Sunkin, M. and Murkens, J. (2013) *Public Law*, Oxford: Oxford University Press, pp.78-112.

[63] For instance, as a central feature of his article, *How to Make Bad Law: Lessons from Cyberspace*, Reed considers how the EU data protection regime provides a clear example of the way in which excessive detail has made European data protection law difficult to understand. It consists not simply of the Data Protection Directive and its national implementing law, but also of a large mass of reports, recommendations, and guidance notes from regulators and advisory bodies. Reed, C. (2010) "How to Make Bad Law: Lessons from Cyberspace", *Modern Law Review* 73(6), pg.914.

[64] See, for instance: Cagliano, A. et al. (2015) "Choosing project risk management techniques. A theoretical framework", *Journal of Risk Research* 18(2), pp232-248; Klinke, A. and Renn, O. (2002) "A New Approach to Risk Evaluation and Management: Risk-Based, Precaution-Based, and Discourse-Based Strategies", *Risk Analysis* 22(6), pp.1035-1209.

> *"All regulators face political risk, the risk that what they consider to be an acceptable level of risk will be higher than that tolerated by politicians, the media, and the public, and that the uncertainties they face will be unrecognized and/or not tolerated."*[65]

Allowing policymakers and regulators to determine the acceptable level of risk in respect of any risk management-based approach to the concept of personal data would, however, likely prove troublesome. It might be asked, for instance, by what metric should regulators assess the acceptability of a particular risk inherent in a particular big data analytics operation when, in reality, notions of acceptability will be viewed differently from one person to another, as well as being assessed quite differently by supposed experts, such as data analysts and privacy regulators?

Perhaps because of questions like this it has been suggested by leading risk researchers that whilst the technical expertise of regulators may be of use in determining the average probability of an adverse effect linked to an object or activity, it is public perception that should invariably govern the selection of criteria on which the acceptability or tolerability of a risk are to be judged.[66] Even if this logic was to be universally accepted, however, due to the divergence of views between different cross-sections of the European populace this would be far from the end of the potential complications.

As noted by Brownsword, for instance, there are at least three major ethical constituencies with a significant presence within contemporary Europe. Namely, these are Utilitarian, rights-based, and so-called 'dignitarian' schools of thought, each of which has a differing attitude to the concept of risk.[67] For the Utilitarian constituency, risk speaks to a concern about safety, that is to say, safety relative to any conditions that are material to the maximisation of pleasure and the avoidance of pain. Utilitarians will want to know what the cost of a risk is; and that the anticipated net result of running said risk is greater to that of not doing so.[68] For members of the rights-based constituency, the notion of risk speaks to a concern over possible infringements of fundamental rights and values. In the same way that Utilitarians may want to consider what price (qua loss of utility) is being paid as a result of running a particular risk, rights-based theorists will want to know whether the running of said risk comes at an acceptable price relative to competing or conflicting rights.[69] Conversely, those of the Dignitarian constituency regard the compromising of human dignity as a self-standing reason for restraint, the notion of risk does not enter into such an ethic in quite the same way. For Dignitarians, to propose that we should exercise precaution against the risk that undesirable consequences may materialise if we fail to do so is to miss the point. If a certain action entails the possibility of compromising human dignity, then regulation should de facto prohibit it, irrespective of any other considerations.[70]

Applying this brief sketch of competing ethical outlooks to the notion of adopting a risk management-based approach to the concept of personal data we might deduce the following. A Utilitarian theorist might hold the view that substantive data protection rules should only be engaged when the risk that a particular data processing operation might result in significant negative or harmful consequences was so high it outweighed any potential benefits that the processing of those data would otherwise accrue. A rights-based theorist, alternatively, may hold the view that a substantive data protection rules should only be engaged when the risk was such that the processing of data to be undertaken came at an unacceptable price relative to any competing rights held by affected individuals. A Dignitarian theorist would outright object to any processing of data capable of compromising human dignity in any

---

[65] Black, J. and Baldwin, R. (n 44). pg.184.
[66] Renn, O. (1998) "Three decades of risk research: accomplishments and new challenges", *Journal of Risk Research* 1(1), pp.49-71.
[67] Brownsword, R. (2009) "Nanoethics: Old wine, new bottles?", *Journal of Consumer Policy* 32(4), pp.355-379.
[68] *Ibid*
[69] *Ibid*
[70] *Ibid*

circumstances. As noted previously, certain data processing activities undertaken as a part of big data analytics may generate an explicit loss of freedom for people make decisions regarding their own lives.[71] As free will is integral to most understandings of human dignity, the mere presence of a possibility that a particular data processing operation might lead to such an eventuality may cause Dignitarian ethicists to object to certain data processing activities in their entirety; the idea of a risk management-based approach to personal data would be moot. This is a point alluded to by Aldhouse, who when advocating a shift to a risk management-based approach to personal data, recognises the opposition such a move would surely evoke from those who hold dear the German doctrine of informational self-determination.[72] Quite clearly, therefore, reaching a consensus as to an acceptable level of risk in the immediate context would be extremely difficult. Another important relevant issue is the fact that even within these three distinct ethical constituencies there may yet be further divisions. Empirical research has indicated, for instance, that attitudes to risk tend to differ drastically between persons of different age groups, genders, and personality types, as well as being heavily affected by factors such as peer pressure and social status.[73]

Accordingly, from this hypothetical scenario we might deduce that even if a broad consensus between policymakers and regulators could be reached as to an acceptable level of risk, the abovementioned difficulties in respect of identifying and quantifying risks in the big data context notwithstanding, it is conceivable that a not insignificant number of individuals within Europe would consider whatever the conclusion might be to be illegitimate or otherwise unsatisfactory.[74] This could prove to be severely problematic for any risk management-based approach to data protection because, as has been widely noted in the literature, a perceived lack of legitimacy is one of the primary reasons for why, historically, numerous high-profile regulatory regimes have ended in failure.[75]

Yet again, however, even if the design-related challenges considered thus far could be successfully negotiated, there would still likely be further complications. When devising risk any management-based approach to data protection, for instance, it will be important for regulators to bear in mind the fact that, generally speaking, the actors they purport to regulate – in this case, primarily individuals and organisations handling and processing large quantities of personal data – can interact quite differently across different regulatory tasks, and that cultures and attitudes to regulation may differ within different limbs of the actors themselves.

In respect of data protection regulators and firms undertaking big data analytics, for instance, some parts of a particular firm (e.g. individual departments) might be more amenable to regulation than others. Furthermore, even if such a firm's response to regulation was not internally fragmented, said firm may

---

[71] See above at n.27.

[72] Aldhouse, F. (n 1) pg.418.

[73] A study undertaken in Switzerland in 2005 involving 388 randomly selected individuals strongly suggested that an individual's general trust and general confidence, as well as their age and gender, have an impact of their perception of risks, particularly those associated with emerging technologies. See: Siegrist, M. et al. (2005) "Perception of risk: the influence of general trust, and general confidence", *Journal of Risk Research* 8(2), pp.145-156. See also: Lahno, A. and Serra-Garcia, M. (2015) "Peer effects in risk taking: Envy or conformity?", *Journal of Risk and Uncertainty* 50(1), pp.73-95; Rohde, I and Rohde, K. (2015) "Managing social risks – tradeoffs between risks and inequalities", *Journal of Risk and Uncertainty* 51(2), pp.103-124; Nicholson, N. et al. (2005) "Personality and domain-specific risk taking", *Journal of Risk Research* 8(2), pp.157-176.

[74] This is despite the fact that recent research has suggested that individuals often agree that greater protection of personal data is a desirable goal. See: ICO (2015) "Data protection rights: what the public want and what the public want from Data Protection Authorities", available at: https://ico.org.uk/media/about-the-ico/documents/1431717/data-protection-rights-what-the-public-want-and-what-the-public-want-from-data-protection-authorities.pdf and DataIQ (last accessed July 2016) "Consumers give firms false personal data due to a lack of trust", available at: http://www.dataiq.co.uk/news/201505/consumers-admit-lying-about-personal-data-due-lack-trust

[75] Brownsword, R. and Goodwin, M. (2012) *Law and the technologies of the twentieth century*. Cambridge: Cambridge University Press, pg.61.

prove to be highly resistant and uncooperative in relation to the work undertaken by the regulator in respect of investigating possible risks, but it may be very compliant once its behaviour is placed at issue (i.e. in the event an errant practice was unearthed).[76] Moreover, the use of risk management-based analyses to guide regulatory operations may, as a general rule, prove far more helpful for some tasks as opposed to others. In the present context of big data analytics it should not be assumed, for instance, that the use of a particular mode of risk management-based analysis as a means of detecting risk in a particular type of big data analytics operation would necessarily identify the existence of risks inherent in other operations of a similar, but different, nature.[77]

Accordingly, whilst the use of risk analyses in the data protection field may prove a useful basis for detecting high-risk actors (e.g. a large data analytics firm), it may not necessarily be especially useful in terms of identifying which individual aspects of the operations of such actors (i.e. particular analytics operations) are deserving of regulatory attention. In any event, the identification of a high level of risk (again, assuming it would be possible to make such a determination) may provide some indication that a regulatory intervention may have to be made into a certain data processing activity undertaken by a particular big data analytics firm, but would not in itself necessarily provide any indication as to how any prospective regulatory response ought to be shaped as a means of reducing that level of risk. This is well highlighted by a fictional example provided by Baldwin and Cave:

> '*Two firms with similarly high risk scores may…be respectively well-intentioned, and ill-informed, or ill-intentioned and ill-informed. The former may respond well to an educative program while the latter is unlikely to. The former does not need to be met with a punitive threat; the latter may have to be.*'[78]

Designing and developing risk analyses which are sufficiently fine-grained to accommodate all such nuances, rather than settling for a rudimentary one-size-fits-all mode of evaluation, may prove far from straightforward, but is another issue that has received scant attention in the data protection literature advocating for the adoption of risk management-based approaches to personal data.

4.2. Challenges of enforcement and application

In addition to the abovementioned design challenges associated with setting an acceptable level of risk, and devising evaluations by which such risks can be assessed, there may also be further complications that would later arise in respect of putting risk management-based regulatory frameworks into practice, as well as their subsequent enforcement.

Firstly, given the fact that, generally speaking, risk management-based models of regulation are necessarily geared towards events that may happen in the future, if a predicted risk does not materialise it can be difficult, if not impossible, to demonstrate that its failure to manifest was in any way resultant of the regulator's actions.[79] In effect, this means that the assessment of the quality and performance any risk management-based approach to the concept of personal data, once constructed and put into operation, would inevitably be extremely difficult to undertake. This is problematic, as being able to identify the successes and shortcomings of any regulatory regime is clearly paramount in terms of assessing its overall quality.[80] The construction of any method or metric against which such a determination could be made or undertaken is something on which the existing data protection literature does not examine in any meaningful way. In so doing, this precisely highlights another area where

[76] Black, J. and Baldwin, R. (n 44). pg.189.
[77] *Ibid*
[78] Baldwin, R. and Cave. M. (1999) *Understanding Regulation: Theory, Strategy. And Practice.* Oxford: Oxford University Press, Chapter 8.
[79] Black, J. and Baldwin, R. (n 44). pg.200.
[80] On regulatory effectiveness generally, see Brownsword, R. and Goodwin, M. (n 75).

further research and more in-depth thinking about the precise practical implications of adopting a risk management-based approach to personal data is required.

Perhaps more troublingly, however, regulating on the basis of risks which would in many cases likely be largely unknowable, such as those inherent in the big data environment, could inadvertently create perverse incentives for organisations handling large quantities of personal data, or those which use personal data for sophisticated analytical purposes, to err on the side of caution when attempting to determine whether they are bound by any regulatory obligations. It has been suggested in the risk management literature, for instance, that risk management, when done well, should be inherently precautionary in respect of perceived risks and harms.[81] In situations where notable and obvious risks can be identified, for instance, the taking of reasonable or sensible precautions is thought to be highly prudent. However, it has also been noted that the taking of reasonable or justified precautions becomes extremely challenging in situations of considerable complexity or uncertainty, as they demand an understanding of the potential false positives as well as false negatives that can be derived from any relevant evidence relating to the risk that is to be managed.[82]

The danger inherent in situations of this type is that the undertaking of such analyses can lead to overly precautionary approaches to the management of potential risks which, in turn, could lead to wider negative consequences. A risk management-based approach to personal data, such as those outlined above, may be particularly likely to encounter such troubles. For instance, given that the risk of an individual being identified or harmed as a result of specific data being processed as part of an individual data analytics operations would in many cases be extremely difficult to ascertain, rather than risking invoking, or breaching, the substantive tenets of data protection law, it is conceivable that data controllers would judge it safer to avoid undertaking certain analytical operations for fear of acting unlawfully and being penalised. Accordingly, this could effectively usher in a regulatory culture which goes beyond what might be described as sensible precautions and, in fact, could be said to embody the much maligned, and altogether more heavy-handed precautionary principle.

The precautionary principle suggests that decision-making bodies should take precautions to guard against certain harms, even in the event there is no clear evidence of such harms occurring, notwithstanding the costs of such action. In other words, the precautionary principle holds that if an action or policy has a suspected risk of causing harm, in the absence of consensus that the action or policy is *not* harmful the burden of proof that the action is not harmful will fall on those taking the action. Quite simply, it requires proof that a certain activity will *not*, or is very unlikely to, cause harm before it should be permitted.[83] The principle has long been a staple of environmental law,[84] and it now seems there may be a chance of it entering the data protection field should a shift to a risk management-based approach to the concept of personal data be initiated, particularly if the level of risk deemed acceptable was particularly difficult to satisfy.

For instance, the standard of an acceptable risk proposed by those supportive of a shift to a risk-driven models of data protection appears to be that data should only be processed free of legal constraints if the risk of harm resulting from that processing is negligible, or at least less than substantial. Whilst this is not quite the 'proof of harmlessness' standard demanded by the precautionary principle, it is arguably not far removed, and would place a considerable demand on data controllers to prove that the data at their disposal either did not, or would be extremely unlikely to, give rise to any risk of harm. When considering the difficulties in making this determination, it does not seem implausible to suggest that,

---

[81] Hrudley, S. and Leiss, W. (2003) "Risk management and precaution: insights on the cautious use of evidence", *Environmental Health Perspectives* 111(13), pp.1577-1581.
[82] *Ibid*
[83] Cross, F. (1996) "Paradoxical Perils of the Precautionary Principle", *Washington and Lee Law Review* 59(3); Manson, A. "Formulating the Precautionary Principle", *Environmental Ethics* 24(3), pp.263-274.
[84] See: Da Cruz, J. (2004) "The Precautionary Principle in EC Law", *European Public Law*.

given the high levels of uncertainty associated with big data, under such an arrangement data controllers would conceivably often end up finding themselves in positions where it would be more attractive to err on the side of caution and avoid undertaking a certain data analytics operation, rather than act, engage the substantive tenets of data protection law, and risk regulatory sanction. Such an approach would be heavily, rather than moderately, precautionary.

As alluded to above, this would not necessarily be a development to be welcomed. The precautionary principle is a fiercely debated construct, and it is often propounded, for instance, that as a regulatory tool it is logically and theoretically suspect, and thus, that it is doubtful that initiatives which have the potential to embody it should ever be afforded a central role in any significant regulatory framework.[85] From a purely theoretical standpoint, and in even the best case scenario, it has been noted that in the face of an emergent risk the presumption that maintaining the status quo, instead of pursuing a concerted course of action, is worthy of priority over conscious change is without empirical foundation and, against the background of the unstoppable evolution of life, even illogical. We often, for example, have as little notion of how undesirable a future where a certain action or behaviour is prohibited would be as we do about how undesirable a future where the same behaviour is allowed.[86] As noted at numerous times throughout this article, this is particularly likely to be true in respect of big data analytics operations and their possible consequences.

It is also worth noting that the cost of overly liberal use of precautionary-type regulatory approaches has historically, in several notable instances, come in the shape of lost commercial and economic advancement could have answered real societal needs.[87] Furthermore, it can be argued when applied fully and logically, the precautionary principle effectively cannibalises itself. If, for instance, the principle was applied rigidly to the idea of precautionary regulation itself, according to the principle's burden of proof approach, advocates of precautionary regulation would be required to demonstrate to a certainty, or at least a near certainty, the absence of counterproductive effects that would stem from said regulation itself. The practical consequences of precautionary regulation, however, are so uncertain that its advocates typically could not meet this burden, and the precautionary principle would prohibit its own use.[88] To many, observers in the risk research field these are all assertions that are serious enough to justify some strong reservations about the precautionary principle, or indeed risk management-based regulatory practices that could potentially give rise to it.[89]

It is not just in general terms, however, that the precautionary principle is thought to be suspect. Writing in the context of modern information technologies, and particularly the way in which such technologies and the law are in a permanent state of mismatch, Thierer notes that many calls for fresh legislative and regulatory action in response to problems raised by contemporary data processing technologies are premised on precautionary principle logic, and rest on the assumption that since many of the relevant recent technological advances could pose some theoretical danger or risk, public bodies should prevent people from using them unless it can be proved that they will not cause any harm. In particular, he offers the following caution:

---

[85] On the difficulties surrounding conceptual and practical aspects of the precautionary principle, see: Allhoff, F. (2009) "Risk, Precaution, and Emerging Technologies", *Studies in Ethics, Law, and Technology* 3(2); Renn, O. (2008) "Precaution and analysis: two sides of the same coin?", *EMBO Reports* 8, pp.303-304; Stirling, R. (1999) "On Science and Precaution In the Management of Technological Risk", *ESTO*; Sunstein, C. (2005) *Laws of Fear: Beyond the Precautionary Principle*, Cambridge: Cambridge University Press;

[86] Somsen, H. "Precautionary Regulation of Reproductive Technologies", in *Regulating Technologies: Legal Futures, Regulatory Frames and Technological Fixes*, ed. by Brownsword, R. and Yeung, K. (2008) Oxford: Hart. pg.229.

[87] *Ibid*

[88] Cross, F. (n 83)

[89] See, for example: Petersen, M. (2007) "The precautionary principle should not be used as a basis for decision-making: Talking Point on the precautionary principle", *EMBO Reports* 8, pp.305-308.

> *'…The views set forth by some of these scholars represent a rather succinct articulation of precautionary principle thinking as applied to modern data collection practices. They are essentially claiming that – because there are various privacy risks associated with data collection and aggregation – policymakers should consider pre-emptive and potentially restrictive approaches to the initial collection and aggregation of data. The problem with that logic should be obvious, however, and as identified by the late political scientist and risk analysis expert Aaron Wildavsky, "if you can do nothing without knowing how it will turn out, you cannot do anything at all". Best case scenarios will never develop if we are gripped with fear by the worst case scenarios and try to pre-emptively plan for them with policy interventions.*'[90]

It is not difficult to draw comparisons between Thierer's warning and the risk management-based approach to the concept of personal data considered earlier in the article. As outlined in the opening sections of the article, the analysis of big data is likely to yield considerable societal and economic benefits.[91] It is not difficult to see, for instance, how adopting a precautionary approach to the regulation of these activities could potentially prevent many of these purported benefits from being realised. Given the huge difficulty in determining levels of risk, it arguably would make little sense for these substantial benefits to be foregone due to a data controller being unable to prove that the processing of certain data would *not*, or at least be very unlikely to, lead give rise to some risk of harm.

However, a shift to risk management-based approach to the concept of personal data would undeniably have the potential to cause such outcomes and, in so doing, put the law on a collision course with the possibilities for innovation and economic and societal development inherent in big data analytics. It might potentially be argued, therefore, that in even a best-case scenario pursuing such an initiative would be inimical to the advancement of knowledge and societal development. According to this line of thinking, the potential risks associated with shifting to a risk management-based model of data protection are so severe, there should be considerable reservations about pursuing with such a course of action. Troublingly, however, as with the other challenges inherent in the creation and deployment of risk management-based approaches to personal data considered above, the majority of the data protection literature proposing the initiation of such a move pays scant attention to both the possibility of the emergence of a system of data protection in this vein, and the indeed the risk research literature which highlights and warns against the problems that this would necessarily entail.

5. Conclusion

As considered above at the article's outset, the emergence of big data has brought forth a number of emergent problems which challenge the smooth operation of the European data protection framework. Whilst a process of change is underway in the area, it has become clear that the European data protection framework's key concept of personal data is to be retained in future regulatory endeavours pertaining to the processing of data. This is despite the fact that there are reasons to believe that, as traditionally envisaged, it may no longer be fully fit for purpose in the emerging big data environment. Evidently, change will be needed if it is to be fully functional in the years to come.

In light of the pressing need for legal reform and the identification of new ways forward, this article has highlighted the apparent general trend of a rise in appetite for an approach to the concept of personal data which can be described as risk management-based. A popular response of academics, regulators, and other observers alike, to big data's emergent policy challenges, for instance, has been to propose a shift to a regulatory position whereby data protection law's substantive rules are only be engaged in

---

[90] Thierer, A. (2014) "Privacy Law's Precautionary Problem", *Maine Law Review* 66(2)

[91] It is noted that in spite of the criticisms of the precautionary principle outlined in this article, some observers have expressly advocated for the adoption of precautionary approaches to big data's regulation. See: Narayanan, A. Huey, J. and Felten, E. (n 59). On the defence of the principle more generally, see: Sandin, P. et al. (2002) "Five charges against the precautionary principle", *Journal of Risk Research* 5(4).

instances where the processing of data poses significant risks of harm, or other negative consequences, to individual persons.

In so doing, these proposals essentially bring together, or at least bring into closer alignment, the fields of law, regulation, and risk research. However, as things stand, all is not well with this arrangement. Despite the proposals bringing these disciplines closer together, many of the observers counselling for the adoption of risk management-based approach to the concept of personal data have apparently overlooked several important issues widely traversed in the risk research literature in any meaningful way.

Particularly, for instance, a survey of the risk research literature reveals that, generally speaking, there are numerous complex challenges associated with the identification, evaluation, and minimisation of risks that arise in conjunction with attempts to deal with complex situations. The big data policy challenges facing the European data protection framework are undoubtedly complex in nature, but yet, in spite of this, scant attention appears to have been paid to these challenges in the legal literature pertaining to data protection law and its reform.

Despite the appetite for risk management-based models of data protection becoming more pronounced, therefore, there are still an array of severely challenging questions to be asked in respect of how risks in the big data context should be quantified, how they should be evaluated, what level of risk ought to be considered acceptable or tolerable, and what could be done to ensure that risk-management based approaches to data protection law and policy do not lead to overly precautionary outcomes.

As outlined in the article's introduction, it was not the intention of the author to counsel against the adoption of risk management-based approaches to the concept of personal data, nor to risk management-based approaches to data protection regulation more generally, but to instigate much needed debate and discussion in respect of the abovementioned perplexing nascent issues. Whilst risk management-based regulatory strategies may have thus far emerged as popular, and potentially promising, responses to some of big data's emergent challenges, it is clear there is still a way to go before debate in this area of data protection reform can be brought to a close. Accordingly, what the present situation now represents is an excellent opportunity for scholars and policymakers to close the gap between legal and risk-related scholarship, and for future inter-disciplinary research. It is hoped that this article may act as a starting point and catalyst towards the realisation of this potential.