When data protection by design and data subject rights clash

Michael Veale*, Reuben Binns** and Jef Ausloos***

Key Points

- Data protection by design (DPbD), a holistic approach to embedding principles in technical and organizational measures undertaken by data controllers, building on the notion of Privacy by Design, is now a qualified duty in the GDPR.
- Practitioners have seen DPbD less holistically, instead framing it through the confidentiality-focussed lens of privacy enhancing technologies (PETs).
- We show that some confidentiality-focussed DPbD strategies used by large data controllers leave data reidentifiable by capable adversaries while heavily limiting controllers' ability to provide data subject rights, such as access, erasure and objection, to manage this risk.
- Informed by case studies of Apple's Siri voice assistant and Transport for London's Wi-Fi analytics, we suggest three main ways to make deployed DPbD more accountable and data subject–centric: building parallel systems to fulfil rights, including dealing with volunteered data; making inevitable trade-offs more explicit and transparent through Data Protection Impact Assessments; and through ex ante and ex post information rights (Articles 13–15), which we argue may require the provision of information concerning DPbD trade-offs.
- Despite steep technical hurdles, we call both for researchers in PETs to develop rigorous techniques to balance privacy-as-control with privacy-as-confidentiality, and for DPAs to consider tailoring guidance and future frameworks to better oversee the trade-offs being made by primarily well-intentioned data controllers employing DPbD.

Introduction

Data protection law has historically faced significant enforcement challenges. Data protection authorities (DPAs) have classically been underfunded and outgunned, possessing limited ability to scrutinize the on-the-ground practices of data controllers and restricted capacity to meaningfully act when transgressions are suspected. In response to these governance challenges, concerned communities have advocated a range of technological approaches that allow effective but non-invasive use of data, or 'DIY' protections which data subjects can adopt unilaterally.[1]

These approaches, often called 'privacy-enhancing technologies' (PETs), are commonly discussed in regulatory circles within the context of 'privacy by design' (PbD). PbD emphasizes that issues of privacy should be considered from the start and throughout the design process through creative social and technical means. Most point to its intellectual home in a report undertaken by the Dutch Data Protection Authority and TNO, with support of the then Information and Privacy Commissioner for Ontario, Tom Wright,[2] although its heritage can be traced further back to the considerations given to 'technical and organizational measures' in the

* Department of Science, Technology, Engineering and Public Policy (STEaPP), University College London, London, UK

** Department of Computer Science, University of Oxford, Oxford, UK

*** Centre for IT & IP Law (CiTiP), KU Leuven, Leuven, Belgium. This article was funded by Engineering and Physical Sciences Research Council, UK (Grant numbers: EP/M507970/1 (to M.V.) EP/J017728/1, EP/N023013/1, EP/N02334X/1 (to R.B.).

1 Claudia Diaz, Omer Tene and Seda Gürses, 'Hero or Villain: The Data Controller in Privacy Law and Technologies' (2013) 74 Ohio St LJ 923, 925.

2 Henk van Rossum and others, *Privacy-Enhancing Technologies: The Path to Anonymity* (Registratiekamer, Den Haag 1995).

data protection directive (DPD)[3] and in the national and regional laws that preceded it.[4] The term PbD entered use around 2000, with the Workshop on Freedom and Privacy by Design at the Computers, Freedom and Privacy 2000 conference in Toronto,[5] and a variety of papers made use of the term around that time.[6] As laid out by the Information and Privacy Commissioner for Ontario from 1998–2014, Ann Cavoukian, PbD is not simply a set of organizational and technical measures to prevent information disclosure, but maps more broadly onto a wider idea of privacy as represented by the Fair Information Practices (FIPs) and even extends beyond them, aiming at a 'significant "raising" of the bar in the area of privacy protection'.[7]

While recommendations of PbD by regulators have significant history,[8] the concept has only recently made it onto the statute books in Europe as part of the General Data Protection Regulation (GDPR).[9] In doing so, it underwent a shrewd transformation into 'data protection by design' (DPbD). This metamorphosis, which some scholars have commented on as wise,[10] makes it clear that the aim is to ensure privacy as enshrined in data protection rights and principles, rather than the flexible, multi-layered and hard to pin down concept of privacy in general.[11] While the European Commission has historically referred to the two concepts synonymously,[12] the focus on DPbD alone provides scope for further clarity. Lee Bygrave summarizes that DPbD requirements, as now enshrined in Article 25 of the GDPR (and also in Article 20 of the Law Enforcement DP Directive),[13] impose a 'qualified duty on controllers to put in place technical and organisational measures that are designed to implement data protection principles effectively and to integrate necessary safeguards into the processing of personal data so that such processing will meet the Regulation's requirements and otherwise ensure protection of data subjects' rights'.[14] The relevant article, Article 25 of the GDPR, reads:

1. Taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organisational measures, such as pseudonymization, which are designed to implement data-protection principles, such as data minimization, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects.
2. The controller shall implement appropriate technical and organizational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed. That obligation applies to the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility. In particular, such measures shall ensure that by default personal data are not made accessible without the individual's intervention to an indefinite number of natural persons.
3. An approved certification mechanism pursuant to Article 42 may be used as an element to demonstrate compliance with the requirements set out in paragraphs 1 and 2 of this article.

3 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 on the free movement of such data, OJ 1995 L 281/31 (Data Protection Directive, hereafter 'DPD').

4 See generally, Gloria González Fuster, *The Emergence of Personal Data Protection as a Fundamental Right of the EU* (Springer, Dordrecht 2014).

5 'Computers, Freedom and Privacy 2000: Full Program' <http://www.cfp2000.org/program/full-program.html#tuesday> accessed 19 November 2017.

6 See eg Julie E Cohen, 'Examined Lives: Informational Privacy and the Subject as Object' (2000) 52 Stanford Law Review 1373; Marc Langheinrich, 'Privacy by Design—Principles of Privacy-Aware Ubiquitous Systems' in Gregory D Abowd, Barry Brumitt and Steven Shafer (eds), *Ubicomp 2001: Ubiquitous Computing*, vol 2201 (Springer, Dordrecht 2001).

7 Ann Cavoukian, *Privacy by Design: The 7 Foundational Principles* (Information and Privacy Commissioner of Ontario, Toronto, Canada 2010) 1.

8 van Rossum and others (n 2).

9 Regulation (EU) 2016/679 of the European Parliament and of the Council 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), OJ 2016 L 119/1 (hereafter 'GDPR').

10 See Mireille Hildebrandt and Laura Tielemans, 'Data Protection by Design and Technology Neutral Law' (2013) 29 Computer Law & Security Review 509, 517.

11 Kieron O'Hara, 'The Seven Veils of Privacy' (2016) 20 IEEE Internet Computing 86.

12 Commission, 'A Digital Agenda for Europe' (Communication) COM (2010) 0245 final.

13 EU Law Enforcement Directive: Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA, OJ 2016 L 119/89 (hereafter the 'Law Enforcement DP Directive').

14 Lee A Bygrave, 'Data Protection by Design and by Default: Deciphering the EU's Legislative Requirements' (2017) 1 Oslo Law Review 105, 114.

What are these principles that technical and organizational measures should take aim at? They are found primarily in Article 5(1): lawful, fair and transparent processing; purpose limitation; data minimization; accuracy; storage limitation; and integrity and confidentiality. Article 5(2) introduces a further, additional overarching principle to the GDPR, 'accountability', laying the burden of proof on the controller to prove compliance with the six principles in Article 5(1).

Yet, in contrast to these wide-ranging principles, within which reside the rights and obligations the legislation details, the PETs literature takes relatively single-minded aim at information disclosure. It focusses in particular on guarantees rooted in either information theory or the computational "hardness" of the resultant re-identification or disclosure problem.[15] Despite attempts in related literature on complementary approaches to coin terms such as 'transparency-enhancing technologies' and 'profile transparency by design', the PET paradigm has dominated the 'by design' discussion in data protection contexts.[16] Unlike the data protection paradigm, which has increasingly shifted to placing accountability obligations upon data controllers in an effort to make them trusted custodians of personal data, the PET paradigm departs from a 'diametrically opposed perception', not of the data controller as a trusted third party, but as an adversary.[17] In a similar vein, recent taxonomies of privacy-enhancing technologies claiming to be 'comprehensive' consider privacy primarily in terms of disclosure risks present at different levels, rather than in terms of the multi-faceted nature of privacy espoused by Cavoukian and the European Commission.[18]

This notion of privacy-as-confidentiality sits at least apart from, and potentially at tension with, the notion of privacy-as-control as espoused by the FIPs and the GDPR.[19] As the Article 29 Working Party notes, PbD incorporates rights such as erasure, noting that 'functionality should be included facilitating the data subjects' right to revoke consent, with subsequent data erasure in all servers involved (including proxies and mirroring).' They note that in addition to data confidentiality, 'controllability', 'transparency', 'data minimization', and 'user friendly systems' should be considered under the PbD umbrella.[20]

Despite these clarifications by regulators, the renaming of the term to emphasize its focus, and the commentary in the literature on the wide array of protection goals that privacy engineering should have,[21] 'privacy by design' in practice is often a narrower affair. Where data are of high dimensionality (where they have many distinct variables), many PbD approaches aimed at the 'unlinkability' of data[22] will inevitably fail to prevent information disclosure where faced with a capable adversary. This does not mean that PETs cannot be used to minimize or reduce risk in this way, but we argue that this minimization comes at a cost. That cost can be, as we demonstrate with case studies, the effective ability to wield data protection rights—the 'intervenability' promoted by privacy-as-control—over such data.[23] The important rights of access and portability (Articles 15, 20), erasure (Articles 17), and the right to object to processing (Article 21) suffer in particular as a result.

There is a danger that data controllers implement privacy design strategies[24] that leave them with data that is difficult for them to re-identify, but far from trivial for an adversary to, given that adversaries likely have a high tolerance for inaccuracy and access to many additional, possibly illegal, databases to triangulate individuals with. The situation is worsened by the fact that a data controller may have relatively little technical re-identification capacity, while also having a very low tolerance for inaccuracy when it comes to their provision of core data protection rights, such as access or erasure. Indeed, to erroneously provide a data subject sensitive personal data of another in response to a subject access request would usually be in breach of the same law the controller would be seeking to comply with.

15 See generally, Casey Devet and Ian Goldberg, 'The Best of Both Worlds: Combining Information-Theoretic and Computational PIR for Communication Efficiency', *Privacy Enhancing Technologies* (Springer, Cham 2014).

16 See eg Hildebrandt, Mireille, 'Profile Transparency by Design?: Re-Enabling Double Contingency', in *Privacy, Due Process and the Computational Turn: The Philosophy of Law Meets the Philosophy of Technology* (Routledge, London 2013) 221–46 ; and Milena Janic, Jan Pieter Wijbenga and Thijs Veuge, 'Transparency Enhancing Tools (TETs): an Overview' (2013) *Third Workshop on Socio-Technical Aspects in Security and Trust (STAST), New Orleans, LA, USA, 29 June 2013.*

17 Diaz and others (n 1).

18 Johannes Heurix and others, 'A Taxonomy for Privacy Enhancing Technologies' (2015) 53 Computers & Security 1. Note that not all conceptions of privacy engineering share these assumptions: cf Marit Hansen, Meiko Jensen, and Martin Rost, 'Protection Goals for Privacy Engineering' (2015) *IEEE Security and Privacy Workshops.*

19 Seda Gürses, 'Can You Engineer Privacy?' (2014) 57 Communications of the ACM 20.

20 Art 29 Working Party, Working Party on Police and Justice, 'The Future of Privacy: Joint Contribution to the Consultation of the European Commission on the Legal Framework for the Fundamental Right to Protection of Personal Data' (WP 168, 1 December 2009) 14.

21 Hansen, Jensen and Rost (n 18).

22 Andreas Pfitzmann and Marit Köhntopp, 'Anonymity, Unobservability, and Pseudonymity—A Proposal for Terminology', in *Designing Privacy Enhancing Technologies* (Springer, Dordrecht 2001).

23 On the varied goals of privacy engineering, see Hansen, Jensen and Rost (n 18).

24 Jaap-Henk Hoepman, 'Privacy Design Strategies', *ICT Systems Security and Privacy Protection* (Springer, Dordrecht 2014).

These controllers, some of which we will illustrate below in case studies, have bound their hands in a very particular way. Their actions have reduced their own data protection obligations and shifted a risk onto the data subject, who has been stripped of her ability to manage the risk herself. When the data subject concerned loses trust in a previously trusted controller, there is nothing she can do but wait for a breach and hope that her record is unable to be effectively triangulated.

We do not intend to suggest that this is a deliberate tactic by the data controllers in our case studies (even though it might be an effective one). However, it does not need to be deliberate to be problematic. Trade-offs are a natural part of all complex decision-making, and the need to make them clearly rather than implicitly is a core component of good decision-making in value-laden contexts.[25] Where there are very few organizational or technical measures supporting data protection deployed, DPbD is likely to benefit everyone. But where basic safeguards are already in place, satisfying everyone and their varying privacy preferences[26] may become more difficult, as 'privacy' is no longer a case of Pareto-improvement (under which it can masquerade as a unified concept), but requires choosing a certain approach (eg confidentiality) to the detriment of others (eg control). Thinking in terms of data protection rights and obligations as we do in this article can make this challenge clearer: achieving one makes it more difficult, or even impossible, to achieve others. Not engaging with these trade-offs does not make them disappear, it simply means they have been determined in an arbitrary fashion. In this article, we do however present some vignettes which indicate that certain controllers do pursue an interpretation of these provisions, deliberately or not, which is unfavourable to the effective exercise of data subject rights.

Deliberate or not, these implicit trade-offs are not even contemplated by pre-emptive provisions in data protection law, such as data protection impact assessments (DPIAs). We believe that there are indeed grounds in the GDPR to support more consideration and transparency regarding the way these trade-offs are determined and communicates—and it is important we identify and use them in this way—but it requires new readings of many of the relevant obligations which this article aims to provide. Firstly, however, we turn to real-world case studies to explore this concern in context.

Case studies of rights lost in the balance

WiFi analytics on the London Underground

Between 21 November and 19 December 2016, Transport for London (TfL), the public transit agency for the UK's capital, ran an in-house trial using the WiFi networks installed at 54 of the stations they manage. They collected more than 500 million connection requests from devices passively transmitting their MAC addresses, with the aim of improving (i) customer information for journey planning and congestion; (ii) management of events and disruption; (iii) timetable planning and station upgrades; (iv) retail unit and advertising positioning.[27]

Transport for London, unlike many undertaking WiFi analytics,[28] were aware of legal obligations in this area, data protection in particular. TfL undertook a data protection impact assessment (DPIA) and met with the UK's DPA, the Information Commissioner's Office (ICO).[29] They cite the ICO's WiFi Analytics Guidance[30] in their use of salting MAC addresses to make re-identification on the basis of device hardware data highly challenging for an attacker. In consultation with the ICO, users were informed using a 'layered approach', which included a press release picked up by the media, a news story on 21 November in the Metro (a free morning newspaper widely distributed and read on London transport), a linked website (tfl.gov.uk/privacy) adapted throughout the trial on the basis of feedback with users, 300 large posters on platforms and at station entrances, through social media and through briefings packs issued to station staff and stakeholder organizations.[31]

25 The ubiquity of trade-offs and the importance of making them explicitly is a core component of public policy education. See eg Eugene Bardach, *A Practical Guide for Policy Analysis: The Eightfold Path to More Effective Problem Solving* (CQ Press, Washington DC 2011).

26 A Westin, *Privacy on & off the internet: What consumers want* (Privacy & American Business, 2001); Mark S Ackerman, Lorrie Faith Cranor, and Joseph Reagle, 'Privacy in e-Commerce: Examining user Scenarios and Privacy Preferences', *Proceedings of the 1st ACM Conference on Electronic Commerce, EC'99, New York, NY, USA* (ACM, New York 1999).

27 Transport for London, 'Insights from Wi-Fi Data: Proposed Pilot' <https://perma.cc/6FZX-VHHK>; Transport for London, *Review of the TfL WiFi Pilot* (Transport for London 2017) 6–8 <https://perma.cc/97DG-KU35> accessed 24 October 2017.

28 See eg College bescherming persoonsgegevens, *Wifi-Tracking van Mobiele Apparaten in En Rond Winkels Door Bluetrace (Rapport z2014-00944)* (Autoriteit Persoongegevens 2015) <https://perma.cc/2JVA-9HYR> accessed 20 September 2017.

29 Transport for London, *Review of the TfL WiFi Pilot* (n 27) 22.

30 Information Commissioner's Office, *Wi-Fi Location Analytics* (ICO 2016).

31 See Transport for London, *Review of the TfL WiFi Pilot* (n 27) 22; Transport for London, *TfL WiFi Analytics Briefing Pack* (Transport for London 2016) <https://perma.cc/7PHN-WBGH> accessed 24 October 2017. Note that some NGOs felt that the posters displayed in and around stations were insufficiently clear about how to opt-out. See Ed Johnson-

As location data is high dimensional, it is highly likely to be unique and easy to re-identify. A now classic study showed that only four spatiotemporal points are needed to single out the vast majority of individuals in a dataset, even where records are rendered significantly coarser (something that often heavily diminishes the data's utility).[32] Unsurprisingly, TfL are therefore uncomfortable with releasing the dataset, refusing it on privacy grounds when requested under Freedom of Information law. They note, correctly, that:

> Although the MAC address data has been pseudonymised [...] given the possibility that the pseudonymised data could, if it was matched against other data sets, in certain circumstances enable the identification of an individual, it is personal data. The likelihood of this identification of an individual occurring would be increased by a disclosure of the data into the public domain, which would increase the range of other data sets against which it could be matched.[33]

Some concerns have been raised over the nature of the 'salt' added to the MAC address or other identifier to generate the string to be hashed.[34] While the ICO recommends that a salt be changed after 'a short period of time',[35] and the Article 29 Working Party recommends that a unique device identifier should only be stored 'for a maximum period of 24 hours for operational purposes',[36] it appears that TfL used a constant salt, generated by once typing letters at random on the keyboard with averted eyes.[37] Such an approach creates two risks. Firstly, anyone who knew or discovered this salt could reverse engineer the process. Secondly, and arguably more probably, a constant salt links devices across days, making attacks not aimed at cryptography but based on external sources of data, such as knowing where someone was at four particular times in a week, more feasible.

One approach would seek to make extra efforts to de-identify the held data. The main way to make data more difficult to re-identify would be to give records more frequently-changing, difficult-to-reverse hashed identifiers. But this would likely be unacceptable for some controllers, as it makes the purpose of the analysis they seek to undertake difficult to fulfil, and so data subjects might suspect that data controllers would wish to transform the data in this way. For example, it would preclude the use of analysis to understand longitudinal patterns in data, restricting them only to what can be learned in a snapshot of time. This is far from the logic of the A/B testing style trials favoured in both industry and policy circles right now.[38]

Yet, another approach sits on the side of the data subject, rather than the controller. More specifically, it sits with capabilities and behaviours of the hardware used. Much of data protection law aims to build trust in data controllers as responsible stewards of sensitive information. Yet, proponents of personal PETs take what some may consider as a contrasting, comparatively dismal view of the world—a gloomy planet where every other actor is a potential adversary that wants to do harm to them with their data—and as such seek to adopt technical practices in order to minimize the information that any third party can learn about them. These practices are increasingly popular with some software and hardware producers. Apple's portable devices include MAC address randomization, which seeks to foil third parties working to build a longitudinal record of a particular device's network scanning activity. Some Android devices utilize this, although many manufacturers, such as Samsung, do not support or practice it.[39] This has a similar, although not identical,[40] effect to regularly changing the salt, and serves to make persistent tracking harder.

Williams, 'TfL Needs to Give Passengers the Full Picture on WiFi Collection Scheme' (*Open Rights Group*, 25 November 2016) <https://perma.cc/8YEA-BV8D> accessed 24 October 2017.

32 Yves-Alexandre de Montjoye and others, 'Unique in the Crowd: The Privacy Bounds of Human Mobility' (2013) 3 Scientific Reports 1376.

33 Natasha Lomas, 'How "anonymous" Wifi Data Can Still Be a Privacy Risk' (*TechCrunch*, 7 October 2017) <https://perma.cc/Y63T-MAC8> accessed 24 October 2017.

34 A hash function is a one-way transformation of data. For example, the md5 hash of 'iheartdataprotection' is '374d67ace049664f8837250bab7010ed'. A salt is a string added to data before it is hashed. For example, to add the salt '1' would result in 'iheartdataprotection1', which has a different md5 hash ('d6790618285a4f41c79aba2eb9bced3e'). There should be reversible mathematical link between those two outputs; the only way to reverse engineer is through 'brute force'. Yet, as someone could (and people do) make extremely large tables of all possible MAC addresses and their resultant hashes, salts are crucial to avoid reversal of the hash process.

35 Information Commissioner's Office (n 30) 6.

36 Article 29 Working Party, 'Opinion 13/2011 on Geolocation Services on Smart Mobile Devices' (WP 185, 16 May 2011), 19.

37 Lukasz Olejnik, 'Privacy of London Tube Wifi Tracking' (*Security, Privacy & Tech Inquiries [blog]*, 11 September 2017) <https://blog.lukaszolejnik.com/privacy-of-london-tube-wifi-tracking/> accessed 24 October 2017; Lomas (n 33).

38 While snapshot analytics might help an organisation like TfL better understand overcrowding and crowd management, for example, it would not, for example, allow them to easily understand something such as whether individuals that often run down escalators that subsequently stop by certain posters telling them not to indeed change their behaviour in the future. Whether analytics tracking individuals over time *should* be allowed is not a topic we weigh in on here, only noting that this is the type of analytics prevalent in online industries today.

39 Jeremy Martin and others, 'A Study of MAC Address Randomization in Mobile Devices and When It Fails' (2017) 2017 Proceedings on Privacy Enhancing Technologies 802.

40 In particular, MAC randomization does not prevent attackers recovering the several MAC addresses from unsalted or poorly salted hashes through brute force.

Yet, even with these approaches enabled, researchers are consistently finding ways, both statistical and based on technical implementation or other features of smartphones, to link individuals across contexts.[41] As TfL recognize, despite protections placed at either the controller side or the device side, such data is not safe from re-identification attacks.

Given this risk of reidentification, particularly from adversaries were data to leak, does a data subject not have a right to understand the data that is being collected about them, and utilize their rights, such as the right to object to processing, or the right to erase data relating to them? It is not difficult to imagine a situation where a previously trusted data controller now loses trust, either to be a well-intentioned custodian of data, or to be capable of keeping it confidential with high certainty.[42] While a data subject may well wish to do so, these protections, whilst not fully mitigating any risk, do effectively remove the ability of data controllers to provide the full range of data protection rights usually afforded to data subjects. Indeed, TfL note:

> The salt is not known by any individual and was destroyed on the day the data collection ended. Therefore, we consider the data to be anonymous and are unable to identify any specific device. As we cannot process known MAC addresses in the same manner as we did in the pilot, we are unable to complete any Subject Access Request for the data we collected.[43]

Were TfL to attempt this, they would find that in the cases of some hardware, the difficulty would be compounded by the device MAC randomization practices described above. In particular, while devices can be identified with acceptable levels of accuracy for an attacker,[44] the levels of identification achieved would be insufficient for providing guaranteed and comprehensive erasure, or accurate access (including avoiding divulging information about others).[45] This reduces the protection afforded by law to the security provisions in the GDPR, as well as the trust in the controller to adhere to the principle of purpose limitation, giving the data subject little-to-no control over the data observed about them after the fact.

As mentioned, beyond subject access requests, another provision in the GDPR relates to the right to object to processing.[46] Where the legitimate or public interest grounds are relied upon, data subjects should be 'entitled to object to the processing of any personal data' unless the controller can 'demonstrate that its compelling legitimate interest overrides the interests or the fundamental rights and freedoms of the data subjects'.[47] This manifests as an 'opt-out' provision, recommended in relation to 'big data' analytics grounded in legitimate interests by both the ICO and the European Data Protection Supervisor (EDPS).[48] Opt-outs from Wi-Fi analytics in particular feature in the ICO's guidance on the matter,[49] although whether they are mandatory under European law is unclear.[50] This may yet change in the proposed updated ePrivacy regulation (at the time of writing entering trialogue negotiations), which has been amended to require opt-outs when WiFi analytics have been used.[51] The Dutch Data Protection authority stopped short of mandating WiFi tracking firm Bluetrace to be required to offer opt-outs, instead settling for the company to undertake research

41 Martin and others (n 39); Mathy Vanhoef and others, 'Why MAC Address Randomization is not Enough: an Analysis of Wi-Fi Network Discovery Mechanisms', *Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security (ASIACCS 2016)* (ACM 2016).

42 It could be argued that as TfL already have potentially re-identifiable gate-to-gate data on users travel behaviour collected by their Oyster smart ticketing system, such erasure would often not do much to reduce data on their behaviour from data controllers, unless individuals relied on higher-cost disposable paper tickets. However, WiFi analytics would provide re-identification capacity above this, and could even reveal additional information, such as adverts looked at, or which individuals were travelling in proximity to each other.

43 Transport for London, *Review of the TfL WiFi Pilot* (n 27) 22.

44 One study found identification success ranged from around 20–50% in the presence of MAC randomization, becoming more difficult with more individuals present, and increased time of tracking demanded. See Vanhoef and others (n 41).

45 On the risks of subject access requests creating privacy breaches, see Andrew Cormack, 'Is the Subject Access Right Now Too Great a Threat to Privacy' (2016) 2 European Data Protection Law Review 15. For further discussion on subject access rights and re-identification see below in section **'Case studies of rights lost in the balance'**. Acquiring additional information (arts 11; 12(2) and Recital 57, GDPR)'.

46 Art 21(1), GDPR.

47 Recital 69, GDPR. A similar argument can be made in relation to the public interest ground for processing.

48 Information Commissioner's Office, Big Data, Artificial Intelligence, Machine Learning and Data Protection (ICO 2017) para 69; European Data Protection Supervisor, Meeting the Challenges of Big Data, Opinion 7/2015 (EDPS 2015).

49 Information Commissioner's Office (n 30).

50 BFE Bosch and NANM van Eijk, 'Wifi-Tracking in de Winkel(straat): Inbreuk Op de Privacy?' (2016) 19 Privacy & Informatie 245.

51 Committee on Civil Liberties, Justice and Home Affairs, Report on the proposal for a regulation of the European Parliament and of the Council concerning the respect for private life and the protection of personal data in electronic communications and repealing Directive 2002/58/EC (Regulation on Privacy and Electronic Communications), European Parliament, 2017. See also the original proposal, containing weaker provisions around wireless analytics: Proposal for a Regulation of the European Parliament and of the Council concerning the respect for private life and the protection of personal data in electronic communications and repealing Directive 2002/58/EC (Regulation on Privacy and Electronic Communications), COM(2017) 10 final, 10 January 2017.

into their technical feasibility, and pointing them towards opt-out registers being developed by Dutch civil society organizations.[52] To the authors' knowledge, the company never did so, having instead opted to cease WiFi analytics entirely, its business model being incompatible with the requirements of the regulator.[53]

In addition, the 'Mobile Location Analytics Code of Conduct' proposed by the Future of Privacy Forum (FPF) has opting-out as one of its principles, noting that the option should be available on the website of an operator.[54] Indeed, FPF themselves run an opt-out service which partners with some organisations selling Wi-Fi tracking technologies to provide a global opt-out list (https://optout.smart-places.org/). FPF note on their website, however, that

> Owners of iOS 8 devices that wish to opt-out of Mobile Location Analytics can still do so by visiting the Smart Store Privacy Opt Out Page. However, since this opt out works by recognizing the MAC address of an opted-out device, in the case of iOS 8 devices, any such opt out will be reset when the device's MAC address changes.[55]

This highlights another rights issue—that the Privacy by Design approach taken in the development of Apple devices, among others[56] prevents effective opting out without necessarily providing effective privacy. The ambient environment, much of which is rightfully untrusted, as anybody could silently set up a device capturing MAC addresses, leads hardware providers to make a value choice for data subjects. Whether opting out is possible given MAC randomization is a research question in and of itself. Legally, enforceable Do Not Track signals may be required—something which raise many issue in and of themselves that we do not seek to unpack in this article, suffice to say that they would require unprecedented coordination between the manufacturers of wireless tracking systems and those of mobile devices.[57]

Apple's 'Siri' voice assistant

Voice assistants are commonplace in a range of devices. Typically, these systems, including Microsoft's Cortana, Google's Assistant and Apple's Siri, work by recording and compressing audio data, processing it for transcription on the company's servers, and returning the transcript to the phone, where a local speech synthesis system may 'reply' to the user. The use of this approach has allowed unprecedented accuracy in speech recognition, as well as avoiding energy, resource and space–intensive processing on the terminal device. Many people use these technologies to activate device functionalities, or to dictate messages or documents.

Firms may provide the recording data to data subjects. Google, for example, provide a tool where voice and audio data can be searched and managed.[58] These can be seen as meeting their access obligations under European data protection law, although unlike many implementations of access rights, there does not appear to be a difference in these tools inside or outside the USA.

Other firms, notably Apple, despite providing a near-identical service to their competitors in this regard, do not provide these data to data subjects automatically, nor do they provide such data upon explicit request under the Irish Data Protection Acts.[59] In correspondence with one of the authors, they cite privacy-by-design as the reason for this. Apple's notion of Privacy by Design in relation to voice assistant data seems to hinge on three aspects.

Firstly, Apple claims that voice identifier data is divorced from the usual identifiers that Apple users are familiar with. While Google users log-in with their account details, under which all their voice data are then listed, Apple generate device-specific identifiers that are separate from these identities.

> When Siri is turned on, the device creates random identifiers for use with the voice recognition and Siri servers. These identifiers are used only within Siri and are utilized to improve the service. If Siri is subsequently turned off, the device will generate a new random identifier to be used if Siri is turned back on.[60]

Nevertheless, these are persistent identifiers. It appears that if the user never disables Siri in the device's settings, as we might expect few users to do rather than simply

52 College bescherming persoonsgegevens (n 28) 20. On opt-out registers in relation to the Internet of Things, see generally, Lilian Edwards, 'Privacy, Security and Data Protection in Smart Cities' (2016) 2 European Data Protection Law Review 28, 55.

53 Autoriteit Persoonsgegevens, 'Bluetrace Beëindigt Overtredingen Wifi-Tracking Na Optreden AP' (*Autoriteit Persoonsgegevens,* 20 April 2017) <https://autoriteitpersoonsgegevens.nl/nl/nieuws/bluetrace-be%C3%ABindigt-overtredingen-wifi-tracking-na-optreden-ap> accessed 24 October 2017.

54 Future of Privacy Forum, *Mobile Location Analytics Code of Conduct* (Future of Privacy Forum 2013) <https://perma.cc/LC4B-FHY5> accessed 11 November 2017.

55 Future of Privacy Forum, 'About Mobile Location Analytics Technology' (*Smart Places*) <https://smart-places.org/mobile-location-analytics-opt-out/about-mobile-location-analytics-technology/> accessed 24 October 2017.

56 Martin and others (n 39).

57 The need for such collaboration was emphasised by the Article 29 Working Party (n 36), 18.

58 Google, 'Manage Google Voice & Audio Activity' (*Google Search Help*) <https://perma.cc/BEJ3-PM3G> accessed 24 October 2017.

59 The lead author of this article submitted a subject access request to Apple Distribution International, Ireland, which was denied. The grounds for the denial are referred to in this article.

60 Apple, Inc, *iOS Security: iOS 10* (2017) 49 <https://perma.cc/8EQE-TFW5> accessed 24 October 2017.

opting not to use it, the identifier persists throughout the lifetime of the device. Apple claimed in correspondence that they do not have a technical means to access the Siri identifier on the device, nor to search the data by identifier, as they have chosen not to build one.[61]

Secondly, Apple claim that data usually have their linked identifiers scrubbed, and are eventually deleted after certain times have elapsed:

> User voice recordings are saved for a six-month period so that the recognition system can utilize them to better understand the user's voice. After six months, another copy is saved, without its identifier, for use by Apple in improving and developing Siri for up to two years. A small sub-set of recordings, transcripts and associated data without identifiers may continue to be used by Apple for ongoing improvement and quality assurance of Siri beyond two years. Additionally, some recordings that reference music, sports teams and players, and businesses or points of interest are similarly saved for purposes of improving Siri.[62]

Thirdly, Apple claim that while Siri is able to recognize your name, it does this by sending such details from your phone each time Siri is used, until such a time where it has not been used for ten minutes, upon which it is deleted from the remote server.[63]

Issues with this conception of privacy by design

Upon first glance, the above may seem like privacy-promoting design features. Yet, there are significant conceptual flaws with each, as well as the entire system, that means while Apple currently find it difficult to access this data, re-identification would be possible, if not relatively trivial in some cases.

Firstly, refusing to build a database retrieval tool is no basis on which to refuse data subject rights. Retrieval is generally a standard feature of database systems. Indeed, it is arguably their very purpose. In most cases, data controllers would have to proactively modify their systems in order to remove such functionality from standard database software.[64]

Secondly, refusing to make the device identifier accessible to the data subject through the design of the software while still enabling it to be transmitted regularly to the data controller serves little practical purpose other than obstructing the data subject's ability to verify it is indeed them requesting the data.[65] Indeed, this seems to be doing more to stand in the way of data protection rights than provide privacy by design. Recital 30 is quite clear that such identifiers would be considered associated with a natural person, noting that 'online identifiers provided by their devices', including those provided by RFID (which, being imperceptible, are similarly inaccessible to the average data subject), may either directly enable profiling or identification, or may do so indirectly, such as in combination 'with other information received by the servers'.

Thirdly, while Apple note that they do not permanently save the name you provide on the server, they do save many kinds of information of similar or even greater use in re-identification alongside your identifier. Indeed, Apple note that because it is onerous to send details such as relationships with family members, reminders, and playlists to the server each time a Siri session is started (and would likely introduce unwanted lag and/or data use), they send those initially, and store them there. Even if we were to accept that a device specific identifier was not personal data (despite the rulings surrounding MAC addresses and even dynamic IP addresses), a list of their contacts and their relations to you is relatively trivial even for non-experts to use to re-identify individuals by using easily accessible data sources, like social media. It seems similarly likely that simple re-identification attacks could be formulated against things such as reminders, particularly as they often mention the names of organizations or individuals.

Fourthly, a significant body of research has demonstrated that individuals can be re-identified and clustered by voiceprints alone, which have such re-identification potential that they are being used and proposed for biometric authentication.[66] Apple themselves even possess several patents in this area from their own in-house research activities.[67] Even based on text transcripts without the voice data, researchers have

61 '[W]e have not built any tool that allows us to retrieve this data'; email from Apple Distribution International to author (3 August 2017).

62 Apple, Inc (n 60) 50.

63 Ibid 49.

64 Incidentally Apple, bizarrely, argued in correspondence with one of the authors on the basis of a complaint that data protection rights were not being upheld, that Siri data was not stored in a 'filing system', citing art 2 of the GDPR on material scope. The exemption for data which do not "form part of a filing system" is explicitly intended to apply only to data not processed by automated means. It would be unlikely that this line of argument would find much traction with regulators or in courts.

65 There can be useful reasons for obscuring data from both the user and the controller at a *hardware* level—*secure enclaves*, such as those that enable fingerprint scanning locally without making the verification data directly accessible to the rest of the system, work in this way.

66 See eg Najim Dehak and others, 'Front-End Factor Analysis for Speaker Verification' (2011) 19 IEEE Transactions on Audio, Speech, and Language Processing 788. For opposing work on systems attempting to dodge re-identification, cf Federico Alegre and others, 'Evasion and Obfuscation in Speaker Recognition Surveillance and Forensics', *2nd International Workshop on Biometrics and Forensics (IWBF), 27–28 March 2014, Valetta, Malta* (IEEE 2014).

67 Jerome R Bellegarda and Kim EA Silverman, 'Fast, Language-Independent Method for User Authentication by Voice (Patent US8645137 B2)' (2014) <https://www.google.com/patents/US9218809> accessed 27 October 2017; Adam J Cheyer, 'Device Access Using Voice Authentication (Patent US9262612 B2)' (2016) <https://www.google.com/patents/US9262612>

demonstrated attacks that can re-identify or cluster individuals stylometrically, based on the words and grammar they use.[68]

Compounding this, it is not just how things are said, but what is being said. Sensitive data can be said and held in textual form. How to redact terms that might disclose sensitive data is an active field of research.[69] This is very challenging even when the forms of text are relatively standardized, such as in medical documents[70]—standardization not present in messages or other spoken interactions. As a recent review notes, '[g]eneral-purpose privacy solutions for plain text are scarce and they only focus on the protection of sensitive terms, which are assumed to be manually identified beforehand'.[71] These systems have not been developed with conversation transcripts in mind; it is unclear that there are effective privacy mechanisms in place here that would defend against re-identification. Furthermore, sensitive data is likely to be recorded, including special categories of data under the GDPR, such as political opinions. Without guarantees that private and re-identifiable parts of a conversation have been redacted, which seem technically difficult, if not currently impossible, to provide, little assurance can be given.

Breaches of data which identify a user but do not contain contact information

Many of the most high-profile data breaches concern data which is conventionally personally identifying, such as full names, home or email addresses, or telephone numbers. In such cases, contacting the individuals affected is relatively straightforward since the controller holds relevant details which would enable a communication channel. However such data types needn't be involved for significant negative effects to occur. As discussed in section 'Case studies of rights lost in the balance', there might be unique identifiers from an end-user's device such as a MAC address, IMEI number, and device-generated advertising IDs. Other examples of data sources that would allow a data subject to be identified but not easily communicated with include high-dimensional data, such as web browsing history or lists of available plugins.[72] Web tracking is a common practice, often using browser fingerprints rather than an explicit, provided identifier. Data such as these have, unsurprisingly, been shown to contain sensitive insights. Facebook has in the past used such approaches to profile individuals by 'ethnic affinity', for example.[73] These high-dimensional data are also often highly identifying. One 2010 study found that 83.6% of users visiting a website had a unique device fingerprint, with an additional 5.3% sharing their fingerprint with just one other record. A breach involving web browsing data could enable an attacker to single out an individual's entire browsing history based on supplementary data about just a few pages they visited, an attack demonstrated by researchers on German members of parliament, judges and other public figures.[74] Other devices and modalities than the web are similarly vulnerable. For example, researchers have been able to identify individuals from their gait using just the gyroscopic sensors on a phone for over ten years.[75] The resulting societal situation has meant we are seeing growing instances of high dimensional datasets covering many data subjects' activities that are capable of being mined for their sensitive information, but lacking straightforward nominal identifiers or contact information.

What are the consequences of this? Imagine a breach of such data. What would be required to occur? A notification to the data protection authority under Article 33(1) GDPR would be required. But what about a notification to the individuals whose data has been accessed? That falls under Article 34, which states that '[w]hen the personal data breach is likely to result in a high risk to the rights and freedoms of natural persons, the controller shall communicate the personal data breach to the data subject without undue delay'. This is the case unless there were appropriate technical and organizational

accessed 27 October 2017; Allen P Haughay, 'User Profiling for Voice Input Processing (Patent US9633660 B2)' (2017) <https://www.google.com/patents/US9633660> accessed 27 October 2017.

68 Sadia Afroz and others, 'Doppelgänger Finder: Taking Stylometry to the Underground', *2014 IEEE Symposium on Security and Privacy* (IEEE 2014).

69 David Sánchez and Montserrat Batet, 'Toward Sensitive Document Release with Privacy Guarantees' (2017) 59 Engineering Applications of Artificial Intelligence 23.

70 See eg Stephane M Meystre and others, 'Automatic de-Identification of Textual Documents in the Electronic Health Record: A Review of Recent Research' (2010) 10 BMC Medical Research Methodology 70.

71 Sánchez and Batet (n 69) 24.

72 Seungyeop Han, Jaeyeon Jung and David Wetherall, 'A Study of Third-Party Tracking by Mobile Apps in the Wild' (University of Washington, Tech. Rep. UW-CSE-12-03-01 2012) <https://perma.cc/5L38-VLQN> accessed 15 November 2017.

73 Julia Angwin and Terry Parris Jr, 'Facebook Lets Advertisers Exclude Users by Race—ProPublica' (*ProPublica*, 28 October 2016) <https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race> accessed 31 October 2017. Note, it is unclear whether Facebook advertising operates in this way within the European Union.

74 Mark Ward 'It is easy to expose users' secret web habits, say researchers' (BBC News, 31 July 2017), <http://www.bbc.co.uk/news/technology-40770393> accessed 15 November 2017.

75 Jani Mantyjarvi and others, 'Identifying Users of Portable Devices from Gait Pattern with Accelerometers', *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005 (ICASSP '05)* (IEEE 2005).

protection measures that 'render the personal data unintelligible to any person who is not authorised to access it'; subsequent measures to mitigate the risk had been taken; or it would involve 'disproportionate effort', in which case a 'public communication or similar measure' could be substituted.

Firstly, in some cases an argument could be made that the personal data were unintelligible, and therefore they do not trigger individual notification requirements under Article 34. This might indeed be the case where data were suitably encrypted with state-of-the-art technologies. It is important, however, to draw a distinction between unintelligibility owing to encryption, and merely not being able to associate a record with a particular individual without supplementary data points. What may seem unintelligible, and therefore non-identifiable from the data controller's perspective, might in fact be identifiable with access to minimal additional data, as demonstrated in the aforementioned attack on the browsing habits of German public figures. As attackers will likely hold different and usually illicitly acquired, personal datasets not available to the breached data controller, this distinction becomes highly important to consider.

Secondly comes the thornier topic of the 'disproportionate effort' that may be required to communicate to individuals. Data such as cookies, browser fingerprints, or device-specific identifiers are not traditional means of identifying someone for the purposes of communicating with them. However, they are frequently assigned and collected by behavioural advertising networks precisely to 'communicate' with individuals through the specific medium of in-browser or in-app advertisements. In a lighter vein, some well-known internet pranks involve buying eerily targeted adverts aimed at individual friends, intending to unsettle them.[76] Such examples demonstrate that even without personal contact details for traditional communication channels, such organizations may have ways to contact their data subjects despite lacking traditional contact details.

Thus, one option for such organizations to communicate breaches to affected individuals in a manner that could be deemed proportionate might even be to purchase advertising space, using the same data to target them once more, in order to tell them their data had been breached. Particularly when it comes to shadowy data brokers that the majority of data subjects are unaware of the identity of, let alone details of their practices or how to contact them, a 'public communication' would be unlikely to be effective, not least because these companies have purposefully never developed channels or a capacity to communicate with data subjects.

Another option would be to facilitate communication via those service providers who do have the capacity to link non-traditional identifiers with traditional communication channels, and who could facilitate the communication of a data breach from an organization to the affected individual. Selecting an appropriate intermediary would depend on the context, but for instance, a device manufacturer/operating system provider like Apple or Google can readily link a device ID to an email address, and thus allow an ad network who identifies users by device IDs to communicate a breach to the affected individuals by email. Another example might be a cellphone network service provider who can easily link SIM numbers to IMEI numbers and thus facilitate breach notifications via SMS or phone call.

While the viability of various breach notification measures is highly context-dependent, these examples demonstrate another way in which privacy-as-confidentiality is in tension with data protection principles—in this case, transparency regarding breach notifications. Technical choices which promote privacy-as-unlinkability (eg not being able to associate high-dimensional browsing data to a relevant contact address), could end up denying data subjects the right to know about breaches which have a high risk of affecting them.

Putting the data protection in DPbD

If DPbD risks taking away rights, as in the cases illustrated above, how might we rectify or ameliorate this? Here, we propose some approaches that might help do this, assess their possibilities and pitfalls, and place them in legal context.

Parallel systems to fulfil DP rights

One set of options would be to maintain parallel systems with the explicit purpose of upholding these rights. Here, we outline two main types of these systems in legal and technical context: systems designed to retain data to provide access and better enable erasure and objection, and systems designed to process additional data, which may be provided by the data subject, to make re-identification possible.

76 Brian Swichkow, 'How I Pranked My Roommate with Eerily Targeted Facebook Ads' (*ghostinfluence [blog]*, 6 September 2014) <https://perma.cc/9FGR-JVWQ> accessed 1 November 2017.

Obligations to retain data

In *College van burgemeester en wethouders van Rotterdam v MEE Rijkeboer* (hereafter *Rijkeboer*),[77] the CJEU was referred a question relating to a case where Mr Rijkeboer, a Dutch citizen, asked the Mayor and Executive Board of Rotterdam to provide him with details of the third parties to which any information relating to him held by the municipality had been communicated to. In Mr Rijkeboer's case, the data controller had replied positively but partially, providing only information relating to the previous year, as Dutch law and practice provided that the data from the year preceding had been wiped. The question to the court was whether, in the absence of a timeframe provided within the access rights of Article 12 of the DPD, Member States could impose deletion of such data—which was in a sense metadata about the data held on Mr Rijkeboer[78]—after a certain period of time—meaning that such access rights could not refer to data outside this time period.

The Mayor and Executive Board of Rotterdam, the United Kingdom, the Czech, Spanish, and Dutch governments submitted that the right of access 'exists only in the present and not in the past', while the Greek government and the European Commission submitted that it applies not only to the present but also extends into the past.[79] The court ruled that such a right must necessarily relate to the past to ensure the practical effect of access, erasure and rectification provisions,[80] that the exact time limitation was up to further Member State rule-making, but that a period of one year alone does 'not constitute a fair balance of the interest and obligation at issue' unless it can be shown that anything longer would lead to an 'excessive burden' on the controller.[81] Indeed, any time limit upon this metadata should constitute 'a fair balance between, on the one hand, the interest of the data subject in protecting his privacy, in particular by way of his rights to object and to bring legal proceedings and, on the other, the burden which the obligation to store that information represents for the controller'.[82]

This ruling is pertinent to the current discussion, as it directly places the question of data subjects' rights against what the court described as the 'burden' that data storage places on the controller—a burden which consists increasingly of securing this data against adversaries, rather than just the simple cost of storage media. The Court acknowledged that this was a trade-off that the DPD did not contemplate explicitly; the same can be said of the GDPR.

Another relevant point from this ruling is the distinction made by Court between two types of data in light of the right of access.[83] Firstly, that the 'basic data', used for the functionality of local service provision, was being stored for a longer period than the data regarding the transfers (which of course may be sensitive to the data subject), was noted to be a source of the unfair balance that had been struck by the Rotterdam Mayor and Executive Board.[84] Put differently, one could say the controller adopted a different retention policy for 'content data' (ie the actual personal data such as individuals' names) as opposed to 'metadata' (eg information relating to how the personal data was used and its source). This has an interesting, although not exact, parallel to some alleged PETs. In these technologies, we can also distinguish between different types of data; the full, potentially identifiable data collected, and the transformed data which is now more difficult to link to data subjects. The former is erased after a certain timeframe,[85] often at the time that it is transformed into the latter for retention. Is this erasure a "fair balance"?

Distinguishing different types of information for the purposes of data management is a common industry practice. Yet, the ways in which data are classified within organizations do not always have neat analogues the legal framework. Take the Siri identifiers. Siri identifiers clearly single out a data subject, as they are persistent identifiers that link to a device typically only used by one person. One of the major purposes of this system is to deliver a personalized voice assistant to a data subject. As a result, Article 11(1) GDPR, which relieves data controllers from having to 'maintain, acquire or process additional information in order to identify the data

77 College van burgemeester en wethouders van Rotterdam v M.E.E. Rijkeboer, case C-553/07, 7 May 2009. See also case commentaries by Cécile De Terwangne, 'L'étendue Dans Le Temps Du Droit D'accès Aux Informations Sur Les Destinataires de Données à Caractère Personnel: Note Sous C.J.U.E, 22 Décembre 2010' [2011] Revue du Droit des Technologies de l'Information 65.; and G Overkleeft-Verburg, 'EU sHof van Justitie 7 mei 2009, zaaknr. C-553/07', (2009) Jurisprudentie Bestuurecht 159.

78 Note that the *Rijkeboer* judgement does not refer to this as metadata, we do so here for explanatory purposes.

79 *Rijkeboer* (n 77), paras 37–39.

80 Ibid para 54.

81 Ibid para 66.

82 Ibid para 64.

83 Ibid para 42 et seq.

84 Ibid.

85 In the case of Siri, Apple *further* de-identifies this data after a six month period, noting that 'User voice recordings are saved for a six-month period so that the recognition system can utilize them to better understand the user's voice. After six months, another copy is saved, without its identifier, for use by Apple in improving and developing Siri for up to two years.' See Apple, Inc (n 60) 50.

subject for the sole purpose of complying' with the Regulation (notably accommodating data subject rights), does not apply. And even if it were, Article 11(2) still enables data subjects to have their data subject rights accommodated upon providing additional information that does allow the controller to (re-) identify them. However, the Regulation does not seem to contemplate that technological developments have allowed identification of a data subject for the purpose of service delivery and data processing, but not for the purposes of data access.

In this case, how might a data controller use a parallel system to augment the identification process being undertaken for service delivery to also allow data access? Apple IDs could be stored alongside Siri identifiers in a separate database. While Recital 64 of the GDPR does note that a controller 'should not retain personal data for the sole purpose of being able to react to potential requests', here the controller already holds both sets of personal data, and only needs to establish a link between them. Asking controllers to purposively make it difficult to consistently find a data subject across many datasets held seems problematic in light of the practical challenges of GDPR implementation. A mechanism could also be implemented on the device to obtain the identifiers used. The core question that relates to these approaches is of security. A centralized list in the first case presents a significantly heightened re-identification risk were attackers to gain access to this data. Both options jeopardize what one might suspect to be among Apple's deeper aims—to claim their hands are tied when faced with law enforcement or intelligence services requests, as they have done publicly before.[86]

While these approaches rely on burying data in a haystack, it may also be possible for Apple to provide this data to users in a form only they can access, using encryption techniques. In this case, the data controller would not be retaining the data in a form they could access, but instead providing portability from the outset. Indeed, users might find it useful to have a repository of speech data and transcripts in order to quickly train any new system, were they to change providers. If data protection by design means access and portability by design, there are feasible design solutions that could form part of the strategy from the outset. This may also allow a user the effective right of erasure; the data they hold could be automatically compared with the de-identified database, and the matches removed.

Just as *Rijkeboer* made the data controller have to rethink their data retention process, it seems feasible that future rulings could also take aim further upstream, at the "fair balances" being struck in the design process.

Acquiring additional information (Articles 11; 12(2) and Recital 57, GDPR)

Re-identifying data to an acceptable percentage of certainty to exercise data protection rights may be difficult in practice for any data controller practicing certain types of DPbD, regardless of intention. The GDPR recognizes this in Articles 11 and 12(2), which exempt the data controller from having to accommodate data subject rights if it can demonstrate it is not in a position to identify the data subject. Article 11(2) however, grants data subjects the ability to provide additional data to enable such (re-)identification, something not every data subject might be inclined to do.[87] The final call though, seems to be in the hands of the controller. Pursuant to Article 12(2) *in fine*, the data controller still has an opportunity to demonstrate not being in a position to (re-)identify, even after being provided with additional information by the data subject.

Having said all that, it would still require a considerable burden of proof to adequately demonstrate reidentification is not possible, even despite additional information being provided by the data subject. This does not only appear from the GDPR's general emphasis on accountability and weightier focus on data controllers' responsibilities, but also manifests itself through Recital 57. This Recital notes that while the data controller 'should not be obliged to acquire additional information in order to identify the data subject' to comply with the regulation, they 'should not refuse to take additional information provided by the data subject in order to support the exercise or his or her rights'.[88] Taking a step back, it is of course important to emphasize that an (alleged) inability to fully accommodate data subject rights cannot be exploited to evade data protection law altogether, and that all other provisions (notably those in Article 5 and 6) still apply in full.[89]

An unanswered question remains—such acquired or volunteered additional information still requires a re-identification process that while very possible, may not be straightforward to the data controller to undertake. Indeed, data controllers may not have expertise in this

86 Karl Stephan, 'Apple Versus the Feds: How a Smartphone Stymied the FBI' (2017) 6 IEEE Consumer Electronics Magazine 103.

87 Indeed, as data are increasingly processed by cloud compute services, individuals may not store or retain the copies themselves needed to identify them, particularly when this data is not used directly by data subjects.

88 Recital 57, GDPR.

89 Art 29 Working Party, 'Opinion 1/2010 on the Concepts of 'Controller' xand 'Processor.' (WP 169, 16 February 2010).

space, particularly when it does not form part of their core processing activities. While Recital 57 shines little light on this, Recital 26 provides some guidance, suggesting that factors such as cost, time, available and emerging technology should be taken into consideration.[90]

This provides an interesting avenue for a policy intervention—a possibly controversial one—that could support data subjects' rights. While theoretical attacks for re-identification are often possible, and would likely undermine the privacy-by-design approaches taken above, there is a valid argument about whether these technologies are 'available' in the context of Recital 26. While they might be available to attackers, and traded, like stolen data, on shady online markets, this creates an imbalance between the deployable technologies available to data controllers and those available to their adversaries. Is there an obligation on data controllers to develop (or to procure from security companies) state-of-the-art re-identification tools in order to make data subject rights possible?

There is a parallel here with other examples, albeit not all in the EU, in which certain individuals are owed redress by an organization who lacks the means of identifying them. After a U.S. financial lender, Ally Financial, was found to have racially discriminatory car loan pricing, they were ordered to used census data to estimate which of their borrowers were Black, Hispanic, or Asian in order to (imperfectly) identify the rightful recipients of compensation.[91] Similar efforts might be beneficial in the case of data breaches, where publicly available information could be mined in order to identify a means of contacting affected individuals who are otherwise only known to the controller by their browsing history, device fingerprint, or other data.

Data protection law, in an attempt to be technologically neutral, is silent on imposing specific innovation requirements on data controllers—which is probably a good thing, as mandating technological advancement through legislating it seems like a misguided idea.[92] But were governments, academia, or civil society to develop and make re-identification tools for high-dimensional data publicly available, with a codebase compatible with many types of commercial systems, it would be hard to deny these technologies were 'available' in the sense of Recital 26. Additionally, the possibility for certification bodies outlined in the GDPR may provide a further avenue for keeping up-to-speed on the state of the art technologies in this space.[93] Yet, this comes with its own security risks. Not only are these tools then available to attackers, but they may even be installed and calibrated on the systems that data is being illicitly obtained from, leaving adversaries a little like 'a kid in a candy store'.

When these re-identification mechanisms are already designed however, and out in the published research literature, 'putting a lid on them' would appear to be a poor policy approach. Even where the codebase is scrappy and unreliable, these are precisely the types of tools that 'script kiddie' adversaries are used to working with. Imagining that making these tools more useful and deployable would only serve to help attackers is likely to underestimate adversaries' existing capacity to use and generate knowledge to valorize stolen personal data, as well as to understate the benefit of such tools for giving data subjects more control over the data they are entitled to legal rights over. Indeed, making them more usable may not vastly increase the capabilities of attackers that were always willing to string together unreliable code, and may primarily serve to empower data subjects to manage risks relating to them.

Making trade-offs with Data Protection Impact Assessments

Given the value-laden nature of these trade-offs, it is important that they are made in an explicit way, with care and with rigour. As it stands, the GDPR, being extremely vague about what DPbD means, does not acknowledge either in recitals or the enacting terms the existence of trade-offs within design approaches. When these trade-offs, as we have shown, involve the fundamental rights of data subjects, this is unacceptable.[94]

Data Protection Impact Assessments[95] (DPIAs) are positioned as a potentially apt point in the compliance process to consider the trade-offs present when employing DPbD strategies. DPIAs are the main form of

90 Recital 26, GDPR: 'To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments.'

91 Annamaria Andriotis and Rachel Louise Ensign, 'U.S. Uses Race Test to Decide Who to Pay in Ally Auto-Loan Pact' *Wall Street Journal* (30 October 2015) <https://www.wsj.com/articles/u-s-uses-race-test-to-decide-who-to-pay-in-ally-auto-loan-pact-1446111002> accessed 12 November 2017.

92 This is not to say that the state should not have a role in steering innovation, or strategically funding particular areas—indeed, it often has—but that innovation policy is more complex than imposing a statutory requirement. See generally, Mariana Mazzucato, *The Entrepreneurial State: Debunking Public Vs. Private Sector Myths* (Anthem Press, London 2015).

93 Art 42, GDPR.

94 The very ability for an individual to have access to their personal data forms an explicit part of the fundamental right to the protection of personal data: art 8, Charter of Fundamental Rights of the European Union, OJ 2010 C 83/389 (hereafter the 'Charter').

95 Art 35, GDPR.

preemptive analysis and documentation requirement in the GDPR, taking particular aim at high-risk processing.[96] The GDPR explicitly, albeit in the recitals, notes that the 'risk to the rights and freedoms of natural persons, of varying likelihood and severity, may result from personal data processing which could lead to physical, material or non-material damage, in particular: [...] where data subjects might be deprived of their rights and freedoms or prevented from exercising control over their personal data'.[97] Accordingly, the Article 29 Working Party identifies, as one of the criteria leading to high risks to data subjects, situations where 'the processing in itself "prevents data subjects from exercising a right or using a service or a contract"'.[98]

Yet in the same guidance, it is worth noting that DPbD gets only a fleeting mention as another preemptive approach comparable to DPIA; the Working Party is silent on including DPbD itself as a topic within DPIAs.[99] It is furthermore easy to see how DPbD measures such as those discussed could be seen as one of several 'measures envisaged to address the risks, including safeguards, security measures and mechanisms'.[100] As processing undertaken in response to otherwise risky processing with the intention of decreasing that risk, they might escape the scrutiny applied to the original concern. While an infinitely recursive DPIA is highly undesirable, so is one that lacks appropriate reflexivity.

While a DPIA might, potentially and with further clarification, provide a venue for considering trade-offs, this approach has a number of limitations owing in particular to the weakening of certain key provisions in the final text of the GDPR.[101] While the requirement to 'seek the views of data subjects or their representatives' during the DPIA process suggests that those individuals affected can articulate their views about appropriate trade-offs, this obligation is limited. It is only required 'where appropriate' and 'without prejudice to the protection of commercial or public interests'. The exemption from consulting data subjects where it might affect 'the security of the processing operations', presents yet another situation in which protection of data through obscurity could excuse the pursuit of other substantive data protection obligations. As a result of these limitations, such consultations may often in practice constitute a form-filling task, particularly as these views to be sought are not grounded in any particular task or question, and are not (as we discuss further below) required to be published or publicized.

Right to information about privacy architectures

As we have described,[102] we can increasingly locate examples where data subjects' personal data is being processed without the accompanying data subject rights effectively being enabled. Yet it seems rare for data subjects to be informed before the time of collection or processing that such rights will not apply. Where they are, claims seem highly generalized. Apple's Privacy Policy, for example, simply states that they 'may decline to process [access] requests that are frivolous/vexatious, jeopardize the privacy of others, are extremely impractical, or for which access is not otherwise required by local law.'[103] Which data will be 'extremely impractical' to exercise rights over? Which will be considered to 'jeopardize the privacy of others'? Without this information, it seems unclear that a proper evaluation could be made by a data subject as to whether she wishes to entrust her personal data to such a controller.

Must a data controller, explicitly and without request at the time data are obtained, warn a data subject that the rights they might expect do not exist? This would seem critical if, as data protection law expects, data subjects are to play a part in managing the risks in accordance with their own preferences. There is a requirement to provide 'information necessary to ensure fair and transparent processing', including 'the existence of the right to request from the controller access to and rectification or erasure of personal data'.[104] Yet, it is unclear whether this is a provision that requires the existence of these rights in a general sense—an awareness raising measure, as well as one seeking to provide logistical support (eg through pointing to the relevant controller contact details)—or whether this is an existence of these rights in applied context, considering each type of data processed by the controller. We feel that in light of the overarching transparency principle in Article 5(1), linked explicitly to Articles 13–14 in Article 12(1), the latter reading is well-supported. Considering that Article 11(2) contemplates times when there might be no 'existence' of these rights, it makes sense that this requirement would not apply in those cases. Does this

96 See Reuben Binns, 'Data Protection Impact Assessments: A Meta-Regulatory Approach' (2017) 7 International Data Privacy Law 22 <http://dx.doi.org/10.1093/idpl/ipw027>.

97 Recital 75, GDPR.

98 Art 29 Working Party, 'Guidelines on Data Protection Impact Assessment (DPIA) and Determining whether Processing is "likely to Result in a High Risk" for the Purposes of Regulation 2016/679' (WP 248 rev.01, 4 October 2017) 11.

99 Ibid 14.

100 Art 35(7)(d), GDPR.

101 Binns (n 96).

102 See section 'Case studies of rights lost'.

103 Apple Inc., 'Privacy Policy' (19 September 2017). <http://perma.cc/3DC2-M7Z5> accessed 13 November 2017.

104 Arts 13(2)(b) and 14(2)(c), GDPR.

mean that data controllers would have to invert the obligation, and explicitly tell data subjects that their rights will not be honoured? That would be very useful, but it seems less clear. If this was the case, would controllers also have to tell data subjects of the non-existence of automated decision-making, which in Articles 13(2)(f)/14(2)(g) is phrased in a very similar way? Given that 'solely automated', 'significant' automated decisions seem rare in practice,[105] more often than not this would serve to bulk up information notices in a quite meaningless way.

However, it is clear that there is an obligation on data controllers to provide the reasons for their non-fulfilment of a specific access request *ex post*. In contrast to the unclear scope of *ex ante* information requirements, relevant *ex post* information requirements expect controllers to provide more detailed information on when subject rights are not available and why.[106] But there is no explicit hook in Articles 13–14 for a data controller to provide information *ex ante* about the DPbD measures that might be restricting such rights so that a data subject might assess them, nor provide information on the safeguards being applied to their data that might affect their ability to exercise their rights. Given the importance of these rights to the data protection regime as a whole, this seems problematic in relation to the transparency and accountability principles of the GDPR.[107]

We argue however that such a requirement can be read into the obligation to provide 'meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject' of automated decision making.[108] Commentators have historically viewed the potential of automated decision rights in relation to 'algorithmic accountability' discussions, through the lens of 'decisions' individuals encounter in their day-to-day-lives such as credit scoring or behavioural targeting.[109] Yet, in relation to the envisaged removal of fundamental rights using automated processing, we argue that an automated decision ('which may include a measure')[110] could also be considered in relation to processing that happens internally, within a data controller or processor. These rights have been considered strongly restricted by both a restriction to be 'solely' automated, and to trigger 'legal' or 'similarly significant' effects on individuals. There is strong reason to believe that the systems we have been discussing in this paper meet both conditions. Firstly, privacy enhancing technologies rarely have humans in-the-loop after their initial setup—usually this would undermine mechanisms reducing information disclosure—and as a result, we can broadly think of these technologies as 'solely' automated. Secondly, the removal of rights would arguably have both a 'legal effect', in the sense of changing a data subject's position with respect to Article 11, and a 'similarly [significant]' effect, impacting on fundamental rights and freedoms.

As there appear to be grounds to meet this condition, we lastly have to consider whether or not a discernable 'decision' has been made. The most clear indication that it has in the case of DPbD is that Recital 71 specifically includes that the scope of automated decisions 'may include a measure'—the precise terminology in Article 25(1) describing DPbD as 'technical and organisational measures necessary to ensure, for the processing concerned, that this Regulation is implemented'. Some might say that these measures happened at the moment of system design, not at the point of processing, and therefore, not being solely automated nor affecting a single data subject at that point, no information obligation exists. Yet, to apply this reasoning to profiling systems, such as behavioural advertising, would be absurd. While at a mechanical level, visiting a webpage might trigger the application of a pre-built profile to deliver advertising,[111] the 'logic involved' would presumably not (and seemingly not in the eyes of the A29WP)[112] be restricted to the last leg alone—that a user requested online components, which matched a browser fingerprint to a profile accessed a database, and therefore was provided specific content—but would refer to the broader system insofar as it was relevant to the final decision,

105 Lilian Edwards and Michael Veale, 'Slave to the Algorithm? Why a "Right to an Explanation" is Probably Not the Remedy You Are Looking For' (2017) 16 Duke L & Tech Rev 18.

106 Recital 59, GDPR: 'The controller should be obliged to respond to requests from the data subject without undue delay and at the latest within one month and to give reasons where the controller does not intend to comply with any such requests.'; Art 12(4) stating if 'the controller does not take action on the request of the data subject, the controller shall inform the data subject without delay and at the latest within one month of receipt of the request of the reasons for not taking action'.

107 Indeed, it could be argued that fundamental rights are at stake here in some situations: art 8 Charter in particular, but potentially also other rights and freedoms such as non-discrimination (art 21 Charter) or and freedom of expression (Art.11 Charter).

108 Art 13(2)(f)/14(2)(g), GDPR.

109 See eg Lee A Bygrave, 'Minding the Machine: Article 15 of the EC Data Protection Directive and Automated Profiling' (2001) 17 Computer Law & Security Report 17; Mireille Hildebrandt, 'The Dawn of a Critical Transparency Right for the Profiling Era' in J Bus and others (eds.) Digital Enlightenment Yearbook 2012 (IOS Press, 2012); Edwards and Veale (n 105).

110 Recital 71, GDPR.

111 The A29WP indeed contemplate the possibility of advertising meeting the art 22 requirements. See art 29 Working Party (A29WP), 'Guidelines on Automated Individual Decision-making and Profiling for the Purposes of Regulation 2016/679' (WP 251rev.01, 6 February 2018).

112 Ibid.

including the construction of the profile in question. In a similar manner, rights to understand DPbD systems which are applied automatically would presumably have some broader, systemic notion applicable to them as well.

In the case of a 'measure' such as DPbD, it might be hard to imagine what a right to a human 'in-the-loop', the core remedy offered in Article 22 that Articles 13–15 refer to, would look like in this situation. Yet, these information rights are not explicitly fully restricted by the compatibility of the remedy in a separate article. Indeed, we note that not only does the terminology in Articles 13–15 refer to 'automated decision-making', without either of the conditions in Article 22, it also counsels that it is 'at least', not only, in the context of Article 22 that these rights trigger, opening the door for less restrictive judicial interpretations in the future.

The effect of this reading of automated decision information rights on DPbD measures which prevent the effective exercise of other rights would be twofold. It would firstly oblige controllers to provide 'meaningful information about the [...] significance and envisaged consequences' of such processing—the loss of data protection rights. They would have to do this *ex ante*—the Article 29 Working Party has recently taken the view that the 'meaningful information' rights in Articles 13–14 should provide identical information to those in Article 15.[113] At minimum, this provision would have the same effect we argue is present in Article 13(2)(b) and 14(2)(c) above, reinforcing our reading of the GDPR that to inform data subjects of these lack of rights is mandatory. A more generous reading could even see it go beyond this. The 'consequences' of the loss of data protection rights include a loss of control, and as such this might entail a discussion of the re-identification risk were such data to be accessed without authorization. Insofar as 'envisaged' is understood as 'intended'[114] rather than 'foreseen', it could be countered that the data controller does not 'intend' a data breach, and therefore would not be required to inform data subjects about its potential consequences. Yet, given that such a breach could be highly damaging to data subjects, it is likely to trigger the separate 'significance' requirement, even were it to dodge the 'envisaged' one.

The second consequence relates to the 'meaningful information about the logic involved' requirement. This gets us closer to an obligation on the data controller to provide information about the extent, form and structure of relevant safeguards in a way that can be assessed by the data subject—or indeed, given these are *ex ante* information rights not requiring an existing data subject to trigger, by interested parties more broadly. The Oxford English Dictionary defines 'logic' as 'a system or set of principles underlying the arrangements of elements in a computer or electronic device so as to perform a specified task.' Where the task is partial de-identification or some other computational transformation to render such data difficult to single out, and thus deprive and individual data subject of certain data protection rights, this would indicate that a basic—and importantly, a 'meaningful'—schematic would be provided. The 'meaningful' condition, one of the few changes to these rights from the DPD, obliges information about this logic to relate to the data subject in a useful way—but given a lack of detailed requirements written in the GDPR, we will likely have to wait for this right to be tested to understand how far it will take data subjects.

Despite a lack of detailed requirements relating to information rights, following Articles 5(2) and 24(1), data controllers are expected to be able to adequately demonstrate compliance with all GDPR provisions, which includes security and DPbD obligations. As discussed above,[115] DPIAs might be an important venue for demonstrating this compliance and hammering out the trade-offs faced. As it stands however, they do not seem to be a reliable transparency mechanism—there is no obligation to publish these documents under the GDPR, and indeed industry opinion is highly opposed to such an obligation, usually on grounds of their potential to contain trade secrets and proprietary information.[116] When passed to a DPA as part of prior

113 Ibid 15. For commentary on how this plays into controversy on the issue, see Michael Veale and Lilian Edwards, 'Clarity, Surprises, and Further Questions in the Article 29 Working Party Draft Guidance on Automated Decision-Making and Profiling' (2018) 34(2) Computer Law & Security Review.

114 The German version of the law is perhaps better translated in this way. See Sandra Wachter, Brent Mittelstadt and Luciano Floridi, 'Why a Right to Explanation of Automated Decision-making does not exist in the General Data Protection Regulation' (2017) 7 International Data Privacy Law 2, 84.

115 See above at section 'Making trade-offs with Data Protection Impact Assessments'.

116 See the responses to the draft version of A29WP (n 98) in the response to a Freedom of Information request from the European Commission, DG Justice and Consumers (9 August 2017) <https://www.asktheeu.org/en/request/a29wp_data_protection_impact_ass> accessed 15 November 2017, in particular the enclosed response from DIGITALEUROPE expressing that view, among others. Indeed, trade secrets or intellectual property have been a traditional carve-out in the area of rights to 'logic of the processing'. Yet, according to Malgieri, 'if a conflict should arise between privacy rights of individuals and trade secret rights of businesses, privacy rights should prevail on trade secret rights.' See Gianclaudio Malgieri, 'Trade Secrets v Personal Data: A Possible Solution for Balancing Rights' (2016) 6 International Data Privacy Law 2, 103. Recital 63 GDPR does acknowledge access rights should not adversely affect 'trade secrets or intellectual property and in particular the copyright protecting the software,' but also mentions that such arguments cannot be (ab)used to refuse access altogether. Similarly, the EU Trade Secrets Directive also notes that its provisions 'should not affect the rights and obligations laid down' in the DPD. See Recital 35, Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (trade

consultation,[117] the documents may become subject to local Freedom of Information laws, but given that data controllers can avoid prior consultation by claiming they have mitigated the risk, it is yet to be seen how common prior consultation will be in practice.

The problem with the lack of publishing of this information does not relate to an imaginary world where engaged data subjects pore over the minutiae of DPIAs, but in general the lack of rigorous scrutiny expected of a pluralist society afforded to organizational or technical approaches to privacy. Oversight is unlikely to be useful if only provided at an individual level. Just as individuals suffer from consent fatigue, many of the solutions for the increasingly complex processing ecosystem today risk of a 'transparency fallacy', where the responsibility for obtaining and digesting complex information about computational systems falls, unhelpfully, on the data subject.[118] Instead, having third parties placed as beneficiaries of some information rights would be a useful future step. While DPAs have significantly increased powers to investigate data controllers, this usually happens only after a complaint has been raised.[119] It is difficult to raise a complaint about improper or ineffective applications of privacy or data protection by design without some insight into the system infrastructure—that is, until such systems fail in a large and noticeable way, at which point transparency is hardly a helpful remedy. In theory, high-risk processing where risks cannot be sufficiently mitigated must involve consultation and prior authorization with the responsible DPA.[120] Yet, the precise risks we have been discussing occur during this mitigation process, and the trade-offs are commonly ignored—because they present separate risks from the initial high-risk processing, we are not optimistic they will be flagged and identified. Article 80(1) and 80(2) envisage some role for bodies to exercise rights either mandated by a data subject or (optionally, subject to member state derogations) without them, but these rights do not include those concerning information provision. Aligning information rights with those that can understand, investigate and report potential breaches, or simply to publicly highlight the existence of state-of-the-art technologies that can better make the trade-offs between different aspects of privacy and data protection in the context of DPbD, seems like a requirement for the future, particularly as systems become more pervasive, invisible and complex.

Conclusions

Privacy by design was initially defined as a holistic concept. Amidst the vigorous and welcome research in how technological approaches can help us achieve it, its well-rounded nature has been somewhat lost. PbD, and the DPbD now mandated by law, is seen increasingly as a synonym for the formal privacy enhancing technologies literature that take reducing unwanted information disclosure as their sole goal—partly, at least, as it is a mathematically tractable, single optimization target. These literatures undoubtedly provide useful tools for both data subjects and controllers—we do not contest that—but they are not designed with data protection in mind, when we are sorely in need of such technologies to help us uphold data protection principles in today's data-rich world.

Because data protection does not take sole aim at information disclosure, but a framework of rights and obligations intended to strike fair balances between a wide array of societal aims, fundamental rights and personal freedoms, we argue that the way that deployed PbD solutions trade-off against these rights while leaving significant residual risk to data subjects is problematic. We accept it is often impossible to have everything at once, but believe that this means decisions about which rights and risks to prioritize over others must be openly discussed and decisions rendered accountable. At a high level, the GDPR's transparency and accountability principles would appear to necessitate it. DPIAs are a good venue to make trade-offs, but are intransparent from a lack of publishing requirements, and recent guidance around them has omitted obligations to consider DPbD specifically.

This is particularly important as controllers do have economic incentives to minimize obligations to fulfil data subject requests. While we believe that they are welcome to maximize their economic logics within the boundaries of the law, and that doing so is not a nefarious aim, there is a danger that a range of practices

secrets) against their unlawful acquisition, use and disclosure, 2016 OJ (L 157) 1.

117 Art 36(3)(g), GDPR.

118 Edwards and Veale (n 105).

119 Indeed, given the resource limitations of DPAs, it is hard to see proactive investigations affecting anything but the most high profile actors. The UK Information Commissioner has noted that her office has a history of taking forward complaints even where there is no data subject mandating them, in relation to national debate around whether the UK makes a derogation to incorporate Art 80(2), but even this remains very different from solo investigation. See Information Commissioner's Office, The Information Commissioner's Office (ICO) response to DCMS General Data Protection Regulation (GDPR) derogations call for views. (ICO 2017) para 113.

120 Art 36, GDPR.

emerge that go unscrutinized as a result. Scrutinizing the effectiveness of technical privacy strategies is important to ensure that trade-offs made, implicitly or explicitly, deliberately or not, are compatible with both overarching data protection principles and Article 8 of the Charter. If the form of DPbD is itself shrouded in secrecy, it seems difficult to believe that meaningful oversight is possible.

We urge DPAs and other relevant actors to update their guidance to ensure that *ex ante* transparency rights are enforced to include specific information as to where and why data subjects can expect to lose their rights to DPbD trade-offs. While it is clear in the GDPR that data subjects are entitled to this information *ex post*, when an attempt to use a right such as access, erasure or objection has been refused, we have argued that the Article 13–14 *ex ante* information rights may be more powerful than considered in this domain. Not one but two sources of the GDPR can be drawn upon to support this claim—notification of the existence of rights (Articles 13(2)(b)/14(2)(c)), and meaningful information about significant solely automated decisions and measures (Articles 13(2)(f)/14(2)(g)), both of which to differing degrees appear to oblige data controllers to lay out beforehand which rights and obligations are not being provided, and why.

We urge technical communities, civil society and regulators to support the development of re-identification technologies in order to enable data subjects to be located in partly de-identified datasets so they can better manage their own risks. Re-identification tools are constantly developed by both academic researchers as well as more nefarious actors, yet the codebases for these tools are often too scrappy for easy use by controllers intending to be legally compliant. We argue that making tools to enable rights available will increase data subject agency (by increasing the number of 'state-of-the-art' technologies that are 'available' to controllers) more than it will meaningfully increase the capabilities of adversaries, who, as adept users of scrappy code, are core beneficiaries of the current imbalance.

We also urge the technical community developing PETs to consider, in addition to their current research, how to get more out of the current trade-offs between control and confidentiality in the solutions they engineer. These trade-offs are only likely to get more common, in particular with the growth in technologies such as secure classification and multi-party computation,[121] which are likely to tie controllers' hands in new and interesting ways, or even call our current understanding of a 'controller' and their competencies into question. In calling for this greater examination, we are not seeking to echo Australian Prime Minister Malcolm Turnbull's much-mocked recent proclamation that whilst the 'laws of mathematics are very commendable [...] the only law that applies in Australia is the law of Australia'.[122] But such practical difficulties data controllers face in trying to accommodate data subject rights do form part of a broader disconnect between technical and legal definitions/interpretations of key data protection notions. Indeed, a 2012 Enisa Report highlighted a mismatch between the right to erasure in the law and in practice.[123] There may well be times where legal requirements are technically or mathematically intractable, but without interdisciplinary research and funding challenges to understand the true limits of attempts to maximize these trade-offs, we will not know. While not all PETs research should focus in this way, and there is a lot to be gained from even deeper research into how to reduce information disclosure further in more complex application areas, the lack of research into this area is stark and sorely in need of rectification.

In other cases, such as the WiFi analytics described above, the issue may be both technical and due to a lack of coordination in the use of privacy-enhancing tools in an untrusted environment. Forcing users to make their own devices difficult to track, such as through MAC randomization, serves to chastise them for sensible precaution by stripping them of their rights to manage data that is still risky and fraught with re-identification potential. Recent legislative moves, such as those in the proposed ePrivacy Regulation to make Do Not Track signals from browsers and devices legally binding[124] have some promise in this area, but depending on an individual's threat model, may be of little use. An individual may trust established data controllers, but be using privacy enhancing technologies, such as MAC address randomization, to prevent 'cowboys' with little regard for data protection law scraping their passively emitted data. This type of problem appears

121 See eg Raphael Bost and others, 'Machine Learning Classification over Encrypted Data', *Proceedings 2015 Network and Distributed System Security Symposium* (Internet Society 2015).

122 Chris Duckett and Asha McLean, 'The Laws of Australia Will Trump the Laws of Mathematics: Turnbull' *ZDNet* (14 July 2017) <http://www.zdnet.com/article/the-laws-of-australia-will-trump-the-laws-of-mathematics-turnbull/> accessed 12 November 2017.

123 See Peter Druschel, Michael Backes, and Rodica Tirtea, *The Right to Be Forgotten – between Expectations and Practice* (ENISA, November 2012), https://www.enisa.europa.eu/activities/identity-and-trust/library/deliverables/the-right-to-be-forgotten/ (surrounding difficulties in identifying and locating specific sets of data, corruption of databases, endangering the integrity of backups—and in the extreme, the cost and difficulties surrounding physically destroying the storage device).

124 Committee on Civil Liberties, Justice and Home Affairs (n 51).

to be difficult to solve, and likely deserves more legal, social, and technical examination than it has currently been afforded.

Emerging problems with characteristics such as those we are describing would likely benefit from broad interdisciplinary engagement, including Human-Computer Interaction (HCI) and Responsible Research and Innovation (RRI) which have significant experience in this field. Like all fields filled with trade-offs, they cannot be 'solved'—but we are confident that they can be better navigated and managed. Data protection law can surely help with the first step—acknowledging, both internally and externally, that these trade-offs exist. Making 'better' trade-offs promises to be considerably harder, but surely an important task for those spanning roles, disciplines and sectors in the years to come.

doi:10.1093/idpl/ipy002
Advance Access Publication 4 April 2018