# PRIVACY MINDSET, TECHNOLOGICAL MINDSET

Michael Birnhack,* Eran Toch,** Irit Hadar***

Draft, September 28, 2014

55 JURIMETRICS: JOURNAL OF LAW, SCIENCE & TECHNOLOGY _ (forthcoming, 2014)

*Policymakers around the world constantly search for new tools to address growing concerns as to informational privacy (data protection). One solution that has gained support in recent years among policy makers is Privacy by Design (PbD). The idea is simple: think of privacy ex ante, and embed privacy within the design of a new technological system, rather than try to fix it ex post, when it is often too late. However, PbD is yet to gain an active role in engineering practices. Thus far, there are only a few success stories.*

*We argue that a major obstacle for PbD is the discursive and conceptual gap between law and technology. A better diagnosis of the gaps between the legal and technological perceptions of privacy is a crucial step in seeking viable solutions. We juxtapose the two fields by reading each field in terms of the other field. We reverse engineer the law, so as to expose its hidden assumptions about technology (the law's technological mindset), and we read canonical technological texts, so as to expose their hidden assumptions about privacy (technology's privacy mindset). Our focus is on one set of informational privacy practices: the large corporation that collects data from individual data subjects.*

*This dual reverse engineering exercise indicates substantial gaps between the legal perception of informational privacy, as reflected in the set of principles commonly known as Fair Information Privacy Principles (FIPPs) and the perceptions of the engineering community. While both information technology and privacy law attempt to regulate the flow of data, they do so in utterly different ways, holding different goals and applying different constraints. The gaps between law and technology point to potential avenues to save PbD.*

* Professor of Law, Faculty of Law, Tel Aviv University. birnhack@post.tau.ac.il

** Senior Lecturer, Faculty of Engineering, Tel Aviv University. erant@post.tau.ac.il

*** Senior Lecturer, Department of Information Systems, University of Haifa. hadari@is.haifa.ac.il

We acknowledge the support of ISF Grant 1116/12, and the first author acknowledges the support of the Israeli Ministry of Science & Technology, Grant 3-9770. We thank Lisa Austin, Julie Cohen, Lilian Edwards, Sue Glueck, Seda Gürses, Natalie Helberger, Susan Landau, Avner Levin, Paul Ohm, Joel Reidenberg, Ira Rubinstein, Omer Tene, Tal Zarsky, and participants at the seventh Privacy Law Scholars Conference (Washington DC, June 2014) for helpful comments, and our team members, Oshrat Ayalon, Arod Balisa, Tomer Hasson, and Sofia Sherman.

Table of Contents

## I. INTRODUCTION

Embed privacy within a technological system as an integral part of the design, and do so ex ante and throughout the technological lifecycle, rather than try to fix it ex post, when it is often too late and expensive. This is the core meaning of Privacy by Design (PbD). Translated into engineering language, PbD insists that privacy is to be considered as a threshold system requirement and should not be traded-off without significant consideration. The idea is attractive, but apparently, difficult to apply. This article explores one understudied explanation for the difficulties in implementing PbD: the deep discursive gaps between the legal field and the field of engineering. We ask what is the law's underlying understanding of technology, and from the other side, how does the field of engineering conceive of privacy? We focus on the classic informational privacy paradigm: the large corporation that collects data from its customers, the data subjects. Accordingly, we offer a close reading of canonical texts in the field of data warehousing, the predecessor of big data, and data science – the analysis of high volume datasets, such as data warehouses and big data. In a nutshell, we find that whereas for lawyers PbD seems an intuitive and sensible policy tool, for information systems developers and engineers it is anything but intuitive, as it goes against the grain of several well-established principles of information systems engineering.[1]

Policymakers around the world search for new tools to address growing privacy concerns. Regulatory options range from market-based solutions on the one end, to intense regulation on the other end. PbD has gained broad support in recent years as a possible regulatory mode. However, thus far, there are very few PbD success stories. Despite the regulatory enthusiasm, PbD is yet to gain an active role in engineering practices.

PbD faces many challenges. The *first* challenge is conceptual: the very concept of privacy is controversial, contested, and unstable. While lawyers and engineers may agree that privacy should be designed into new systems, there is little agreement as to what privacy means. This difficulty is enhanced on a global scale: James Whitman aptly characterized the legal understanding of privacy in the United States as one of liberty, and the European understanding of privacy as one of dignity.[2] A *second* challenge is ideological. Technologists might reject PbD for

---

[1] *Cf.* Professor Edward Felten's observation, that "[i]n technology policy debates, lawyers put too much faith in technical solutions, while technologists put too much faith in legal solutions." Quoted in Paul Ohm, *Breaking Felten's Third Law: How Not to Fix the Internet*, 87 DENV. U. L. REV. ONLINE 50 (2010). Here, we examine the possibility of law and technology joining hands, to achieve a shared goal.

[2] *See* James Q. Whitman, *The Two Western Cultures of Privacy: Dignity Versus Liberty*, 113 YALE L.J. 1151 (2004).

its purported intervention in the technological process, and libertarians might consider it an interference with the market, at least inasmuch as it is required by law. A *third* challenge, and our main focus here, is technological: implementing privacy might require considerable resources and goes against principles of technological design. The challenge is not merely a technical hurdle. As Helen Nissenbaum observed in assessing the difficulties of gaining support for a specific PbD technology, the greatest barrier was "a cultural mythology of innovation, incredibly powerful in the context of the Internet and web."[3] Privacy is framed as an impediment to innovation.

In this article we explore the gaps between privacy law and technology in the contexts of data warehousing and data science. Data warehousing is an engineering field that focuses on collecting, integrating, and analyzing large quantities of data from heterogeneous sources over longitudinal periods of time. Data science is an emerging field and applies sophisticated algorithms to mine big datasets, find patterns, and use them to predict behavior. Data science builds on data warehousing, data mining, and other practices related to collecting, managing, and analyzing data, from small to large scale datasets (commonly known as big data.) These two fields increasingly represent contemporary information systems. First, the distinction between contemporary analytical systems and operational systems is increasingly blurred. For instance, recommendation systems that suggest products to consumers in electronic commerce (such as Amazon's familiar recommendations feature) merge analytical and operational elements. Second, as the volume of data grows exponentially, so does the interest in analytical systems, especially regarding big data analysis. Third, large corporations almost necessarily have a data warehouse in place. These organizations collect data about their customers and hence raise issues of informational privacy.

We examine two leading books on data warehousing and one book on data science: Ralph Kimball and Margy Ross, *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling* (3rd ed., 2013) (hereafter: *DW Toolkit*),[4] William Inmon, *Building the Data Warehouse* (4th ed. 2005) (hereafter: *Building DW*),[5] and Foster Provost and Tom Fawcett, *Data Science for*

---

[3] Helen Nissenbaum, *From Preemption to Circumvention: If Technology Regulates, Why Do We Need Regulation (And Vice Versa)?*, 26 BERKELEY TECH. L.J. 1367, 1384 (2011).

[4] Ralph Kimball & Margy Ross, *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling* (3rd ed., 2013) (hereafter: *DW Toolkit*).

[5] William Inmon, *Building the Data Warehouse* (4th ed. 2005) (hereafter: *Building DW*).

*Business: What You Need to Know About Data Mining and Data-Analytic Thinking* (2013) (hereafter: *DS Business*).[6]

These are books that developers study from and turn to for guidance, among other sources. These books both constitute the developers' knowledge and serve as a reflection, a snapshot if you wish, of contemporary engineering perceptions.[7] We search for the underlying conception that the field of engineering holds as to informational privacy. The reading does not end with the text, but seeks to expose its subtext. Our goal is to uncover *technology's mindset* as to privacy that is embedded in these books: how and what do they think of privacy? We use the term *mindset* to refer to the overall doctrine that emerges from the texts (the law or the engineering books), which has its own objectives, language, and characteristics. The mindset encapsulates the type of systems that the doctrine finds useful, legitimate, and desirable.

This instrumental reading, joined by initial empirical evidence about developers' perceptions in a related research,[8] indicate substantial gaps between the legal perception of informational privacy, as reflected in the set of principles commonly known as Fair Information Privacy Principles (FIPPs) and the engineering community's perception of privacy.[9] While both technological systems and privacy law attempt to regulate the flow of data, they do so in utterly different directions, holding different goals and applying different constraints. A diagnosis of the gaps between the legal and technological perceptions of privacy is a vital step in seeking viable solutions.

Part II provides a background of PbD, outlining its origins and its rise in legal circles. We characterize PbD as *code*, namely a technological solution for a socio-legal problem caused by technology, and offer a typology that differentiates it from Privacy Enhancing Technologies (PETs). We then query PbD's lack of success, and point to several challenges it faces. Parts III

---

[6] Foster Provost & Tom Fawcett, *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking* (2013) (hereafter: *DS Business*).

[7] We use the term developers to refer to those who design, develop, or build information systems, and reserve the term engineers to refer to those active in the professional field that surrounds information systems.

[8] Irit Hadar, Tomer Hasson, Oshrat Ayalon, Sofia Sherman, Eran Toch, Michael Birnhack, Arod Balisa, *Are Designers Ready for Privacy by Design? Examining Perceptions of Privacy among Information Systems Designers* (work in progress, 2014).

[9] FIPPs originated in the United States, in a governmental report following Watergate. *See chapter III of the Report of the Secretary's Advisory Committee on Automated Personal Data Systems* (1973), known as the Ware Report, available at http://epic.org/privacy/hew1973report/foreword.htm. FIPPs are now the common grounds between different approaches to informational privacy, namely the American and European approaches.

and IV have parallel tasks. Both law and technology attempt to regulate data flows, and both are not void of social values. The law necessarily holds hidden assumptions as to technology, even when the law tries to use technology-neutral language. We need to decipher these assumptions, which accumulate to form the law's *technological mindset*. This is the task of Part III. Technology too is designed based on hidden assumptions. In Part IV, we strive to unearth these assumptions, and expose technology's *privacy mindset*, namely, how does the technological field think about privacy? Juxtaposing the two mindsets – the law's technological mindset and technology's privacy mindset takes us back to PbD, and indicates the depth of the gap between the two. The conclusion points to potential avenues for solutions.

## II. PRIVACY AND DESIGN

There is some enthusiasm among policymakers about PbD on both sides of the Atlantic. There are different variations and nuances of PbD, which require clarification so to set common grounds. This Part begins with a broad-brush outline of PbD's development and its legal anchors, characterizes it as a mode of regulation by *code*, and differentiates it from PETs. We conclude by raising some doubts as to PbD's success.

### A. Origins[10]

The idea of privacy by design—the term not yet used—emerged in the 1990s, with the convergence of several factors: (1) the fast diffusion of digital networks; (2) the maturation of data protection law; (3) the understanding of both law and technology as reflecting social values; and (4) the search for new modes of regulation.

*First*, the increasing importance of digital networks, namely the Internet, in our lives enabled easier, cheaper and faster collection, processing, and transfer of data about end-users. As our lives go digital, we face new forms of processing of vast amounts of data so to offer targeted advertising, making predictions about users' behavior, and much more.

*Second*, the legal maturation of informational privacy, or data protection in European terms, conveniently marked by the 1995 European Data Protection Directive, meant that personal

[10] This section provides a broad-brush outline of PbD. The political history of PbD is yet to be studied.

data deserved closer regulatory attention than ever before.[11] The Directive signaled a shift from local regulations (such as in Sweden and Germany), and from soft law (i.e., non-binding rules) in the form of the OECD's 1980 Guidelines,[12] to hard law (i.e., binding rules)—at least for the EU Member States. The Directive soon became a legal engine for spreading data protection law around the world.[13]

*Third*, by the mid-1990s, philosophers and sociologists of technology have made their case loud and clear, that technology is not merely a technical tool, but a human creation that necessarily reflects a certain set of values.[14] The study of the social aspects of technology (STS – Science, Technology & Society) taught us that designers embed certain values in the technology and that users construct the social meaning of technology over time. Accordingly, the social design of technology is a matter of choice.[15]

*Fourth*, the initial enthusiasm with the Internet was soon replaced with a concern for its downsides. Pornography and its negative affect on children drew most of the attention in the 1990s, especially in the United States,[16] with other concerns added, such as copyright infringements, defamation, terrorists' use of the network, and more.[17] It became clear that traditional, conventional law was inadequate to address such new issues. There was also a growing

---

[11] Council Directive 95/46/EC, On the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data, 1995 O.J. (L 281) (hereafter: Data Protection Directive).

[12] *See* OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data of 1980 (hereafter: 1980 OECD Guidelines).

[13] *See* Michael D. Birnhack, *The EU Data Protection Directive: An Engine of a Global Regime*, 24 COMP. L. & SEC. REP. 508 (2008); Graham Greenleaf, *Sheherezade and the 101 Data Privacy Laws: Origins, Significance and Global Trajectories,* 23(1) J.L. INFO. & SCI. (2014), available at http://ssrn.com/abstract=2280877.

[14] *See* e.g., Helen Nissenbaum, *Accountability in a Computerized Society*, *in* HUMAN VALUES AND THE DESIGN OF COMPUTER TECHNOLOGY 41 (Batya Friedman ed., 1997).

[15] In the context of technologies of identification, Jeffrey Rosen writes that "Nearly all of these technologies can be designed in ways that strike better or worse balances between liberty and security." *See* JEFFREY ROSEN, THE NAKED CROWD: RECLAIMING SECURITY AND FREEDOM IN AN ANXIOUS AGE 100 (2004).

[16] Culminating in the enactment of the Communications Decency Act of 1996, which was then struck down by the Supreme Court. *See* Reno v. ACLU, 521 U.S. 844 (1997).

[17] *See* e.g., THE OFFENSIVE INTERNET: SPEECH, PRIVACY, AND REPUTATION (Saul Levmore & Martha C. Nussbaum, eds., 2012).

interest in other forms of regulation, such as private ordering, or for our purposes, regulation by technology, which Joel Reidenberg called *Lex Informatica*,[18] and Lawrence Lessig called *Code*.[19]

With the above factors converging, it was only a matter of time until the idea of designing technology was applied to privacy. A few scholars pointed to this avenue. Joel Reidenberg observed as early as 1993 that "[t]echnical choices lead to normative decisions about fair information practice standards."[20] In 2004, Daniel Solove wrote that "privacy must be protected by reforming the architecture."[21]

The first technological response to privacy concerns was to counter privacy threatening technologies with PETs. We shall return to the relationship between PETs and PbD in section D. The first official discussion of PbD, albeit an indirect one, appeared in a 1995 joint publication of the Information and Privacy Commissioner of Ontario, Canada and the Dutch data protection authority.[22] The report placed much emphasis on the anonymization of personal data as a key to protecting privacy, and its innovative approach was to treat PETs as a policy tool, rather than just a technological one. Accordingly, for transactions that require identification, the report articulated the question that data controllers should ask: "how much personal information/data is truly

---

[18] Joel R. Reidenberg, *Lex Informatica: The Formulation of Information Policy Rules through Technology*, 76 TEX. L. REV. 553 (1998).

[19] LAWRENCE LESSIG, CODE AND OTHER LAWS OF CYBERSPACE (1999).

[20] *See* Joel R. Reidenberg, *Rules of the Road for Global Electronic Highways: Merging the Trade and Technical Paradigms*, 6 HARV. J. L. & TECH. 287, 301 (1993). *See* also at 303, writing that "The technical paradigm locates control of information practices in the network infrastructure." Later on, Reidenberg was more explicit, writing that "The same information infrastructure that creates the privacy dilemma may also offer opportunities to develop and implement fair information practices rules that preserve citizens' rights while further enhancing economic value." *See* Joel R. Reidenberg, *The Use of Technology to Assure Internet Privacy: Adapting Labels and Filters for Data Protection*, CYBERNEWS III:6 (1997). On the engineering side, the Platform for Internet Content Selection (PICS), a working group within the World Wide Web Consortium (W3C) applied PICS to privacy, resulting in the Platform for Privacy Preferences (P3P). *See* http://www.w3.org/P3P/, and Lorrie Faith Cranor, *The Role of Data Protection Authorities in the Design and Deployment of the Platform for Privacy Preferences*, XXIII International Conference of Data Protection Commissioners (2001), available at http://lorrie.cranor.org/pubs/paris-talk0901.html.

[21] DANIEL SOLOVE, THE DIGITAL PERSON: TECHNOLOGY AND PRIVACY IN THE INFORMATION AGE 100 (2004).

[22] Tom Wright & Peter Hustinx, *Privacy-Enhancing Technologies: The Path to Anonymity* (Volume I) (August, 1995), available at http://www.ipc.on.ca/english/Resources/Discussion-Papers/Discussion-Papers-Summary/?id=329. Hustinx credits Ontario's Assistant Commissioner at the time (later commissioner) Ann Cavoukian, and John Borking on the Dutch side. *See* Peter Hustinx, *Privacy by Design: Delivering the Promises*, 3 IDIS 253 (2010).

An earlier use of the term appeared in a 1973 report by the County Council of Essex in the UK, which provided architectural guidelines for residential areas. The report seems to have used the term intuitively. *See* County Council of Essex (Planning Department), *A Design Guide for Residential Areas* 117 (1973) ("High level windows on first floor rear elevation makes privacy by design possible on private side, for dwellings opposite and adjunct.")

required for the proper functioning of the information system involving this transaction?"[23] The report continued: "This question must be asked at the outset - prior to the design and development of any new system."[24] This question reflects the important privacy principle known as data minimization: collect only the minimum data needed for a legitimate purpose.[25] In discussing what the report called the Identity Protector, the authors wrote: "A simple guideline for designers of new information systems is to minimize the identity domain wherever possible and maximize the pseudo domain."[26] The essence of these answers was later renamed as privacy by design.

Thereafter, PbD deserved some more occasional discussion in legal literature. In a 1996 presentation, Herbert Burkert suggested a thoughtful and comprehensive taxonomy of PETs, focusing on identity.[27] Burkert discussed various design options, with a keen social understanding of their role: PETs, he argued, follow normative decisions about the technological design.[28] Julie Cohen, rehearsed in STS studies, wrote in 2000 that "Currently, technologies designed to measure consumer preferences permit retrieval and matching of data with names and other identifying characteristics. Systems could be designed quite differently. They could, for example, allow aggregate profiling of groups of consumers without generating personally-identified or identifiable data."[29] Cohen mentioned PETs, the term used at the time, but essentially, referred to PbD: "At minimum, however, law can and should establish a new set of institutional parameters that supply incentives for the design of privacy-enhancing technologies to flourish. Legal protection alone cannot create or guarantee informational privacy. But it is a place to begin."[30]

PbD gained momentum. In the late 1990s, the World Wide Consortium (W3C) began working on a Platform for Privacy Preferences (P3P), based on the Platform for Internet Content

---

[23] Wright & Hustinx, *id.* at s. 1.3.

[24] *Id*. *See* also at s. 1.7.5.

[25] *See* Data Protection Directive, art. 6(1)(c).

[26] Wright & Hustinx, *supra* note 22, at s. 1.6.

[27] The presentation was published as an article two years later. *See* Herbert Burkert, *Privacy-Enhancing Technologies: Typology, Critique, Vision*, *in* TECHNOLOGY AND PRIVACY: A NEW LANDSCAPE 125 (Philip E. Agre & Marc Rotenberg, eds., 1998).

[28] *Id.* at 130.

[29] Julie E. Cohen, *Examined Lives: Informational Privacy and the Subject as Object,* 52 STAN. L. REV. 1373, 1401 (2000).

[30] Cohen, *id.* at 1437-38.

Selection (PICS).[31] A 2001 OECD forum discussed PbD, treating it under the headline of PETs.[32] The suggestion to OECD Member States was to "Actively encourage developers of systems and software applications to incorporate privacy into the design of information technologies."[33] In the same year, computer scientist Marc Langheinrich suggested—under the explicit heading of Privacy by Design—six principles for guiding the design of ubiquitous systems: notice, choice and consent, proximity and locality, anonymity and pseudonymity, security, and access and recourse.[34] However, few scholars followed.

Over time, the concept of PbD has evolved, promoted by Ann Cavoukian, Ontario's Information and Privacy commissioner. Today, Cavoukian advocates PbD as a comprehensive "philosophy and approach." Her version of PbD now covers not only technology, but business practices and physical design,[35] and contains broader privacy principles. She maintains that the objectives of PbD are "ensuring privacy and gaining personal control over one's information and, for organizations, gaining a sustainable competitive advantage," and lists seven "foundational principles," which are: "(1) proactive not reactive; preventive not remedial; (2) privacy as a default setting; (3) privacy embedded into design; (4) full functionality – positive sum, not zero-sum; (5) end-to-end security – full lifecycle protection; (6) visibility and transparency – keep it open; (7) respect for user privacy – keep it user-centric."[36] Here, we focus mostly on the technological aspect of PbD.[37] PbD has now become somewhat of a brand. The question here is about its substance.

---

[31] *See supra* note 20.

[32] Privacy Online, *OECD Guidance on Policy and Practice*, 21 (2003) (reporting the 2001 forum, one of its items being "the challenges of, and methods for, educating business about the importance of privacy by design and the use of PETs.") Interestingly, the discussion originated from the industry, with Stephanie Perrin, then Chief Privacy Officer of Zero-Knowledge Systems Inc., a Canadian company that developed PETs.

[33] *Id.* at 30.

[34] Marc Langheinrich, *Privacy by Design: Principles of Privacy-Aware Ubiquitous Systems*, 3 UBICOMP PROCEEDINGS 273 (2001).

[35] *See* e.g., Ann Cavoukian, *Privacy by Design* (January 2009).

[36] *See* http://www.privacybydesign.ca/index.php/about-pbd/7-foundational-principles/.

[37] For a discussion of the organizational aspects of PbD, *see* e.g., Julie Smith David & Marilyn Prosch, *Extending the Value Chain to Incorporate Privacy by Design Principles*, 3 IDIS 295 (2010).

## B. Legal Anchors

In recent years, we witness another step in the evolution of PbD: from a rather abstract idea and campaign, it moves towards becoming a binding legal requirement, at least in the EU. This section surveys and analyzes the main points on PbD's legal timeline. A first bud, then not yet classified as PbD, is found in the EU Directive. Article 17(1) sets the principle of "security of processing" of personal data. It instructs that –

> Member States shall provide that the controller must implement appropriate technical and organizational measures to protect personal data against accidental or unlawful destruction or accidental loss, alteration, unauthorized disclosure or access, in particular where the processing involves the transmission of data over a network, and against all other unlawful forms of processing.
>
> Having regard to the state of the art and the cost of their implementation, such measures shall ensure a level of security appropriate to the risks represented by the processing and the nature of the data to be protected.

Note that article 17 imposes a duty on the data controller,[38] it focuses on the data itself (rather than on the source of the data, the data subject or other nodes in the flow of data), with an emphasis on different aspects of data security, including its integrity (that it is not altered), confidentiality (that it is not disclosed), and security vis-à-vis third parties (preventing unauthorized access). The requirement is phrased as an open standard (rather than a precise rule), imposing a duty to implement technologies, but without elaborating on what kinds of technologies, or when should they be implemented. The second paragraph delegates the risk assessment to the data controller. The controller is to decide which measures to use. In other words, this early version of PbD referred only to one aspect of the overall protection of personal data—data security; it was ancillary to the legal toolkit of data protection and was meant to support it from the outside.

A substantial boost to PbD came about with the 2010 Jerusalem Resolution – a joint statement by a group of data protection commissioners from around the world.[39] The Resolution

---

[38] "'Controller' means the natural or legal person, public authority, agency or any other body which alone or jointly with others determines the purposes and means of the processing of personal data; . . ." Data protection Directive, art. 2(d).

[39] 32nd International Conference of Data Protection and Privacy Commissioners, Resolution on Privacy by Design (2010), available at http://www.justice.gov.il/NR/rdonlyres/F8A79347-170C-4EEF-A0AD-155554558A5F/26502/ResolutiononPrivacybyDesign.pdf.

placed PbD under the global spotlight. It recognized PbD "as an essential component of fundamental privacy protection," but it did not elaborate on the scope, coverage or meaning of the principle, other than by the resolution's encouragement and adoption of Cavoukian's seven principles. Thus, PbD left the sideline position of the data protection field, and stepped into the center, but still, it remained rather general, and non-binding.[40]

The next step was to try and anchor PbD into the law itself, as a binding legal requirement. Following a 2009 European Consultation which endorsed PbD and suggested that it should be binding for technology designers,[41] in January 2012, the European Parliament published a comprehensive proposal to replace the 1995 Directive with a General Data Protection Regulation (GDPR).[42] Article 23 of the proposed GDPR, under the heading "data protection by design" suggested the addition of several dimensions: time, scope, subject matter, and substantive principles. As for the temporal dimension, according to the GDPR, the technological measures should be applied at the initial design and throughout the lifecycle of the processing. As for scope, whereas the Directive refers only to data security, the GDPR refers to the entire basket of the requirements set out in the GDPR, which is an updated set of FIPPs. The subject matter and focus are on the data subject. The proposal incorporated the data minimization principle not only as a general obligation, but as a positive requirement. Thus, the GDPR allocated a far greater role for

---

[40] Subsequent declarations of the same forum mentioned PbD in passing, or not at all, *see* Mexico City Declaration, 33rd International Conference of Data Protection and Privacy Commissioners (2011), available at http://privacyconference2011.org/htmls/adoptedResolutions/2011_Mexico/Mexico_City_Declaration_ENG.pdf (no mention of PbD); Punta del Este Resolution on Cloud Computing, 34th International Conference of Data Protection and Privacy Commissioners (2012), available at http://privacyconference2012.org/wps/wcm/connect/92d083804d5dbb9ab90dfbfd6066fd91/Resolutionon_Cloud_Computing.pdf?MOD=AJPERES. The 35th international meeting focused on mobile application. The commissioners found that "App developers are often unaware of the privacy implications of their work and unfamiliar with concepts like privacy by design and default." *See* Warsaw Declaration on the 'appification' of society, 35th International Conference of Data Protection and Privacy Commissioners (2013), available at https://privacyconference2013.org/web/pageFiles/kcfinder/files/ATT29312.pdf.

[41] Article 29 Data Protection Working Party, WP 168, *The Future of Privacy* (2009), at sec. 46, available at http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2009/wp168_en.pdf.

[42] *See Commission, Proposal for a Regulation of the European Parliament and of the Council on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation)*, COM (2012) 11 final (Jan. 25, 2012), *available at* http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf (hereafter: GDPR.)

PbD. The proposal that PbD would become an all-encompassing principle provides an additional necessary safeguard to the entire regulatory basket.[43]

The GDPR is currently debated in the European legislative bodies, and in March 2014, the European Parliament adopted several amendments, based on proposals of the Committee on Civil Liberties, Justice and Home Affairs (known as LIBE), though the proposals were not fully adopted.[44] The current GDPR text adopts a risk assessment consideration, namely, that the data controller should apply technological measures that are proportionate to the risk; it requires the data controller and processor to implement "appropriate and proportionate" technical and organizational measures throughout the entire lifecycle of the system, and it requires that PbD follows the basic principles of the proposed GDPR: accuracy, confidentiality, integrity, physical security and deletion of personal data, purpose limitation, and data minimization.[45]

In the meantime, the idea of PbD has crossed the border from Canada to the United States. The Federal Trade Commission (FTC) endorsed PbD as an important element for a new informational privacy legal regime, but in a different way than the European approach. The FTC published a final report in March 2012.[46] It aims first and foremost at the business sector rather than policymakers, and prefers self-regulation to governmental regulation, although the FTC does join the call to consider the enactment of "baseline privacy legislation." The baseline principle is defined as "Companies should promote consumer privacy throughout their organizations and at

---

[43] Article 37 of the GDPR further strengthens the PbD, by listing the tasks of a data protection officer, a new requirement set in Article 35.

[44] *See* Amendment 118, regarding Article 23 of the European Parliament legislative resolution of 12 March 2014 on the proposal for a regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation) (COM(2012)0011 – C7-0025/2012 – 2012/0011(COD)) (Ordinary legislative procedure: first reading) (hereafter: LIBE Amendments), available at http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+20140312+ITEMS+DOC+XML+V0//EN&language=EN#sdocta5. For LIBE's proposals, *see*: Committee on Civil Liberties, Justice and Home Affairs, Report on the proposal for a regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation) (COM(2012)0011-C7-0025/2012-2012/0011(COD)) (November 22, 2013).

[45] Id.

[46] FTC REPORT, PROTECTING CONSUMER PRIVACY IN AN ERA OF RAPID CHANGE – RECOMMENDATIONS FOR BUSINESSES AND POLICYMAKERS (2012). For a critique of the FTC's privacy policies, *see* Randal C. Picker, *Unjustified By Design: Unfairness and the FTC's Regulation of Privacy and Data Security* (draft, 2013), available at http://www.masonlec.org/site/rte_uploads/files/PickerGMUDraft.pdf.

every stage of the development of their products and services."[47] PbD is designated a prominent place in the recommendations to businesses. The benefits of PbD are explained as shifting the burden from consumers (the FTC does not use the broader European term of data subjects, and remains within its mandate to regulate trade, referring to consumers) to the businesses, the latter are required to "treat consumer data in a responsible manner."[48] Importantly, the Report explicitly discusses the scope of PbD – which principles should it cover? The answer is the entire FIPPs toolkit, as updated and modified in other sections of the report. PbD should refer to data security, reasonable collection limits,[49] retention practices,[50] and data accuracy.[51] The proposed PbD principle is accompanied by an organizational principle, that "Companies should maintain comprehensive data management procedures throughout the life cycle of their products and services."[52] Thus, like Cavoukian, the FTC wishes to apply PbD not only to technology, but also to organizational procedures. However, although it is broader in scope, it is also weaker, in that it is a recommendation to businesses, rather than a binding legal duty.

In January 2012, while the FTC's process was pending, the White House published its own report, *Consumer Data Privacy in a Networked World*.[53] The report proposed the legislation of a Consumer Privacy Bill of Rights, to be implemented by codes of conduct that would be developed by government, industry and consumer advocates, and enforced by the FTC. The White House report listed several privacy principles: individual control, transparency, respect for context, security, access and accuracy, focused collection, and accountability. PbD was not listed.

An indirect reference to PbD is found in the context of data security. The Consumer Privacy Bill of Rights proposes that consumers have a right to secure and responsible handling of

---

[47] FTC REPORT, *id*. at 22. Interestingly, the Report refers to the "broad international recognition and adoption of privacy by design," *id*.

[48] *Id.* at 23.

[49] The FTC explained 'reasonable limitations,' by pointing to the tension between business needs for flexibility and innovation in using consumers' data for new purposes on the one hand, and consumer privacy on the other hand. *Id.* at 25-26.

[50] The report adds also the disposal of data, *id.* at 27-29.

[51] *Id.* at 23-24.

[52] *Id.* at 30-32.

[53] *See Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy* (The White House, 2012), available at http://www.whitehouse.gov/sites/default/files/privacy-final.pdf.

their personal data, and then explains the correlative duty imposed on the companies, to conduct a risk assessment and maintain "reasonable safeguards to control risks such as loss; unauthorized access, use, destruction, or modification; and improper disclosure."[54] An explanatory note comments that "Technologies and procedures that keep personal data secure are essential to protecting consumer privacy," but also clarifies that the security principle "gives companies the discretion to choose technologies and procedures that best fit the scale and scope of the personal data that they maintain."[55] Finally, the report cautions against "Prescribing technology-specific means of complying with the law's obligations."[56] This should not be read as opposition to PbD, but as a call to pursue technologically-neutral means, and more importantly, leaving it to the market to decide.

Another notable legal anchor of PbD, albeit a soft law one, is the 2013 OECD Guidelines governing the protection of privacy and transborder flows of personal data, which update the 1980 OECD Guidelines.[57] While the guidelines do not bind any country, they do signal an overall approach to privacy among developed countries. PbD is not explicitly mentioned in the 2013 Guidelines (nor were they mentioned in the 1980 Guidelines). However, article 15 suggests that a data controller should have in place a management program that, *inter alia*, "provides for appropriate safeguards based on privacy risk assessment." The explanatory notes suggest that "privacy management programme can also assist in the practical implementation of concepts such as 'privacy by design,' whereby technologies, processes, and practices to protect privacy are built into system architectures, rather than added on later as an afterthought."[58]

To summarize, PbD is yet to become a binding legal rule. Its first appearance in the EU Directive was relatively minor, limited to data security, and delegated to the discretion of the data controller. Later suggestions, especially the proposed GDPR, substantially boosted and strengthened PbD in several dimensions. On the American side, PbD is supported by the FTC, but

---

[54] *Id.* at 19.

[55] *Id*. Additional indirect references to privacy protection by technologies are found in reference to the proposed multistakeholder process, *id.* at 24.

[56] *Id.* at 35.

[57] *See* OECD Recommendation of the Council concerning Guidelines governing the Protection of Privacy and Transborder Flows of Personal Data (2013) [C(80)58/FINAL, as amended on 11 July 2013 by C(2013)79].

[58] *Id*. Art. 15(a)(iii) and notes at 24.

only as a self-regulation, non-binding tool. The OECD Guidelines reflect a similar position: PbD can be one way for the data controller to manage its duties, but it is not required.

### C. Code

The overview of PbD's legal anchors sheds light on its unique character as a form of regulation. Unlike conventional legal regulation that determines the right and wrong or provides incentives to act in a certain way or avoid action altogether, PbD is regulation by technology itself. This characteristic is an important part in its attractiveness, but also its vulnerability.

Lessig famously pointed to four modalities of regulation: other than the law, he pointed to market norms, social norms, and architecture – physical and technological.[59] He concluded that architecture, or code, is law: it regulates what we can or cannot do, no less than law. The way technology is designed enables us to use it in certain ways but not others, thus shaping our behavior. Framed in these terms, PbD is a deliberate attempt to use technology to serve social and legal goals.

Not being formal law, PbD means that libertarians should be happy with it: if PbD is successful, it would obliterate the need for governmental intervention in the market. Cyber-libertarians should also be happy with it: the less the government interferes with technology, the better. The private ordering nature of PbD explains the differences between the European and American approaches in this regard. The Europeans trusted their governments to regulate personal data, with the result of an extensive legal regime. Regulating technology, including imposing it as a legal duty, does not scare them. The American distrust in government, by contrast, is a foundational principle. This is the basis of the constitutional system of checks and balances. PbD is welcome, as long as it is adopted as a measure of self-regulation. The FTC report discussed above is the clearest in this approach.

Regulation of technology raises further challenges, such as the pace of technological development. The law is slow to respond, and is difficult to amend. The typical legal response to this difficulty is to enact technology-neutral laws. Later on, in Part III, we shall argue that this is by and large a myth, as there is always a hidden technological mindset that limits our cognition and limits the law.

---

[59] LESSIG, CODE, *supra* note 19.

To sum up, PbD might be intuitive to lawyers (well, once they think about it) and attractive, but as a duty imposed onto data controllers, it is a blunt interference with technology and with business practices, and runs into legislative challenges. As a form of self-regulation, it can avoid these difficulties, but then there is less of an incentive to adopt it.

### D. PETs and PbD

As we saw earlier, PbD stemmed from PETs.[60] The literature often treats the two alike, or as closely related,[61] but they have partially diverged, and we join Ira Rubinstein in arguing that it is useful to distinguish between them more clearly.[62] Importantly, the differentiation that we offer here is more pedagogical than descriptive. It is often the case that PETs and PbD can be integrated, and in any case, they are not mutually exclusive. Our purpose is to better understand PbD and its challenges.

Initially, PETs were defined broadly, referring to any technology that protects privacy or enhances it in some way.[63] The intention was to embed at least some FIPPs within these technologies. Initial emphasis was on anonymization, following David Chaum's first anonymization technology.[64] It took policymakers a while to catch up with the new technological thread, but by the mid-1990s the term PETs was coined and soon became a policy goal. For example, the 1995 Dutch-Canadian study emphasized identification and its flipside –

---

[60] *See* Burkert, *Privacy-Enhancing Technologies*, *supra* note 27; Ann Cavoukian, *Privacy and Radical Pragmatism: Change the Paradigm*, *in* PRIVACY BY DESIGN 15, 24 (2008) ("This concept [PETs] also includes the design of the information system architecture.") Simon Davies attributes PbD's origins, to cryptographic techniques and only later associated with PETs. *See* Simon Davies, *Why Privacy by Design is the Next Crucial Step for Privacy Protection* 3 (2010), available at http://www.i-comp.org/wp-content/uploads/2013/07/privacy-by-design.pdf.

[61] Hustinx, *Privacy by Design*, *supra* note 22, at 254 writes: "It is also clear that the concept of PET is closely related to the principle of 'data minimization' that is now widely used, and gradually developed into the principle of 'Privacy by Design.'"

[62] *See* Ira S. Rubinstein, *Regulating Privacy by Design*, 26 BERKELEY TECH. L.J. 1409, 1411-12 (2011) (arguing that "PETs are applications or tools with discrete goals that address a single dimension of privacy . . . In contrast, privacy by design is not a specific technology or product but a systematic approach to designing any technology that embeds privacy into the underlying specifications or architecture.")

[63] *See* e.g., COLIN J. BENNETT & CHARLES D. RAAB, THE GOVERNANCE OF PRIVACY: POLICY INSTRUMENTS IN GLOBAL PERSPECTIVE 179 (2006) (explaining how technology can become a policy instrument in this context.)

[64] *See* David L. Chaum, *Untraceable Electronic Mail, Return Addresses, and Digital Pseudonyms*, 24 COMMUNICATIONS OF THE ACM 84 (1981). For a survey of anonymization technologies, *see* Chris Nicoll, *Concealing and Revealing Identity on the Internet*, *in* DIGITAL ANONYMITY AND THE LAW: TENSIONS AND DIMENSIONS 99, 122-26 (C. Nicoll, J.E.J. Prins, M.J.M. Dellen, eds., 2003).

anonymization, and accordingly suggested various anonymization technologies.[65] Indeed, technologies that mask a user's identity protect her privacy, at least as long as re-identification is unavailable or prohibited.[66]

Over the years, numerous technologies offered various privacy-related services. Anonymization technologies are prominent in this list, ranging from Chaum's MIX to web-anonymizers, remailing services, to TOR.[67] Other technologies provide data security, with encryption technologies leading this thread. Yet another kind of privacy technologies offer management tools, such as P3P, which sought to match users' privacy preferences with the websites' self-declared policies.[68] Note that in practice, P3P acts behind the scenes, and does not require users to take active steps.

Various classifications of PETs were suggested in the literature,[69] to the extent that in 2008 the British Information Commissioner (ICO) admitted that "there is no widely accepted definition for the term privacy enhancing technologies," but pointed to its core principles: reducing the risk of contravening privacy principles; minimizing personal data held about data subjects; and empowering individuals to maintain control over their data.[70]

A few commentators pointed to differences between PETs and PbD.[71] Davies wrote that "Where PETs focused us on the positive potential of technology, *Privacy by Design* prescribes that we build privacy directly into the design and operation, not only of technology, but also of

---

[65] *See supra* note 22.

[66] Paul Ohm argued that because de-anonymization has become easier, the legal criterion of non-identifiability collapses. *See* Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701 (2010). For a critique, *see* Felix T. Wu, *Defining Privacy and Utility in Data Sets*, 84 U. COL. L. REV. 1117 (2013).

[67] *See* TOR Project, at https://www.torproject.org/, and its academic origin: Roger Dingledine, Nick Mathewson & Paul Syverson, *Tor: The Second-Generation Onion Router*, 13 USENIX SECURITY SYMPOSIUM (2004).

[68] *See* Platform for Privacy Preferences (P3P) Project, at http://www.w3.org/P3P/.

[69] *See* e.g., Burkert, *Privacy-Enhancing Technologies*, *supra* note 27; BENNETT & RAAB, THE GOVERNANCE OF PRIVACY, *supra* note 63, at 181-97 (classification based on the agents involved and the technologies' feature as policy instruments); Rubinstein, *Regulating Privacy by Design*, *supra* note 62, at 1422 (distinguishing front-end software development that addresses design processes for customers, and back-end practices, which are data management policies.) For a thorough review, *see* Lothar Fritsch, *State of the Art of Privacy-Enhancing Technology (PET)*, PETWeb Project (2007), available at http://publications.nr.no/4589/Fritsch_-_State_of_the_Art_of_Privacy-enhancing_Technology.pdf.

[70] Information Commissioner's Office, Privacy by Design 8 (2008), available at http://ico.org.uk/for_organisations/data_protection/topic_guides/~/media/documents/pdb_report_html/PRIVACY_BY_DESIGN_REPORT_V2.ashx.

[71] *See* also Rubinstein, *Regulating Privacy by Design*, *supra* note 62.

operational systems, work processes, management structures, physical spaces and networked infrastructure."[72] This observations point to PbD's divergence from PETs, in that PbD now refers not only to technology, but to organizations as well. PETs have also expanded, to cover what Hustinx calls privacy enforcing and privacy enabling technologies.[73]

Diaz, Tene and Gürses advocate embracing PETs.[74] They focus on PETs that enable "individuals to engage in online activities *free from surveillance and interference*,"[75] and distinguish between three kinds of PETs: first, PETs that are implemented by the data controller.[76] Examples are a system that enables the European Electronic Toll Service to collect fees without receiving locational data, and protocols suggested for smart metering.[77] A second category is PETs implemented on the user's side, within a service offered by the data controller.[78] An example is encryption that enables users to communicate on a social network or email service, without the platform being able to access the communication. The third category is collaborative applications which do not involve the data controller,[79] the TOR network being the leading example.

Based on the PETs and PbD literature, we suggest the following conceptualization: both PbD and PETs share the same goal, of promoting informational privacy, by using technology, with the purpose of complying with FIPPs, but they diverge in the way they try to do so, and subsequently, in the kind of protection they can offer. The term PETs is better reserved for third party technologies, to be used in conjunction with the application technology, and thus usually come into operation *ex post*, after the application had already been deployed. PbD, by contrast, is better reserved for technological measures embedded in the application technology itself, and thus, often comes into operation *ex ante*, before or during the technological design. Put differently,

---

[72] Davies, *Why Privacy by Design, supra* note 60, at 3.

[73] Hustinx, *Privacy by Design*, *supra* note 22, at 253.

[74] Claudia Diaz, Omer Tene & Seda Gürses, *Hero or Villain: The Data Controller in Privacy Law and Technologies*, 74 OHIO ST. L.J. 923 (2013).

[75] *Id.* at 924.

[76] *Id.* at 944.

[77] *Id.* at 959. The authors argue that "policymakers should *incentivize* and, in appropriate cases, *require* implementation of PETs into the design of infrastructures, products, and services."

[78] *Id.* at 950.

[79] *Id.* at 953.

PETs are applied to the outer layer of the system where it interfaces with the outside world, whereas PbD aims at an inner layer of the system.[80]

This characterization is based on two criteria: the party that initiates and operates the privacy-related technology, and the timing of its implementation. The first criterion enables us to better assess the issue of trust. PETs do not rely on the developer of the application technology, while PbD expects—or demands—that the cat guards the milk. Both options have their advantages and disadvantages in this respect. PETs carry more trust than PbD, but the external PET designer is likely to be less familiar with the system for which the PET is intended, and might not be able to access all the data she needs in order to provide the best technological protection possible. A PbD designer has better knowledge of the system and full access, but users might be suspicious about the privacy protection system. Context, of course, matters here. The latter PbD deficiency can be answered by regulatory review, peer review within technological circles, or the market: if a company applied PbD which turns out not to provide sufficient privacy, users might react,[81] as well as the regulator in some cases, for example, the FTC might investigate a company for deceptive practices.[82]

The second criterion that distinguished PETs from PbD is the temporal dimension, the former being applied *ex post* and the latter applied *ex ante*. PbD, as advocated, places much importance on the timing, and rightly so. Depending on the system, redesigning it so to better protect privacy interests might be expensive. For example, when Congress required the manufacturers of backscatter body scanners deployed in American airports, to install Automated Target Recognition (ATR) technology that produce a generic figure rather than the naked image of passenger, one manufacturer did not meet the legal-technological requirement – these scanners were pulled out of the airports.[83]

---

[80] *See* Sarah Spiekermann & Lorrie Faith Cranor, *Engineering Privacy*, 35 IEEE TRANSACTIONS ON SOFTWARE ENGINEERING 67 (2009). We are indebted to Julie Cohen for suggesting this description to us.

[81] Several well-documented technological and business changes to privacy policies and practices resulted in a public outcry, and the companies retreating. *See* e.g., the Google Buzz fiasco, which resulted also in a class action: *In re* Google Buzz Privacy Litigation, No. 5:10-CV-00672-JW (N.D. Cal. Sept. 3, 2010) *available at* http://www.archive.org/download/gov.uscourts.cand.224341/gov.uscourts.cand.224341.41.1.pdf.

[82] The FTC has power to enforce the prohibition of "unfair or deceptive acts or practices in or affecting commerce." *See* §5 of the Federal Trade Commission Act, codified as 15 U.S.C. §45.

[83] *See* Yofi Tirosh & Michael Birnhack, *Naked in Front of the Machine: Does Airport Scanning Violate Privacy?*, 74 OHIO ST. L.J. 1263 (2013).

Two more differences derive from the above distinctions. First, PETs are installed and used by the user herself, whereas PbD is installed by the data controller, perhaps even without the end user's knowledge.[84] Again, each approach has both advantages and disadvantages. PETs empower the data subject and provide her with means to control her own data, but this requires that the data subject is aware of the data collection and use, that she understands the implications thereof, cares enough to act, and knows what to do. Thus, there are informational, cognitive, and technological literacy barriers to using some kinds of PETs. For example, how many users who care about their anonymity use TOR? PbD is more user-friendly in this sense, as the user need not take any active step, yet her privacy interests are protected. However, once again this means that the cat guards the milk.

Second, a PbD technology is usually specific, and requires tailor-made design. PETs, in contrast, may be applied at a network level, or be more generic, thus applying to many specific technologies and applications. For example, P3P was embedded in internet browsers, but it still required cooperation of the visited websites.[85]

These differences matter: who designs the technology to begin with and when, who needs to take active steps and of what sort, and the potential scope of the use – justify separating PETs from PbD. In this, we diverge from Cavoukian's portrayal of PbD as a subset of PETs, and wish to reserve PbD for the first category that Diaz et al suggested (technology implemented by the data controller), and reserve their other two categories to PETs (technologies implemented on the user's side, and collaborative applications).

Based on this distinction, we can note the differences in scope and power of each kind of privacy-related technology. PETs are used as an additional layer on top of existing software, for example an anonymizer service. If successful, PETs can mask the user's identity, thus providing full privacy protection (unless and until re-identification is possible using other sources). PETs can alert a user to notices and consent requests, or enable easier opting-out. PETs can provide data security for the data on the user's side. However, PETs do not provide the user with control beyond

---

[84] Or, framed in Rubinstein's terms, we would say that PETs are front-end processes, whereas PbD refers to back-end processes (though PbD need not exclude front-end processes). *See* Rubinstein, *Regulating Privacy by Design*, *supra* note 62.

[85] *See* Lorrie Faith Cranor, Serge Egelman, Steve Sheng, Alecia M. McDonald, Abdur Chowdhury, *P3P Deployment on Websites*, 7(3) ELECTRONIC COMMERCE RESEARCH AND APPLICATIONS 274 (2008) (finding, in 2007, that P3P was deployed on 10% of the sites in the top-20 results of a typical search.)

the initial meeting point between the user and the data controller.[86] PbD, on the other hand, may provide anonymous data collection, but it can also minimize the data collected to begin with, monitor controllers' (and their employees') access to the users' data, or provide data security along the data flow. However, PbD requires the user's trust, as we saw earlier.

Once again, PETs and PbD are not mutually exclusive. They can work in tandem, for example, a possible privacy design can leave enough leeway for end-users to choose how they wish to manage their privacy, including by using external PETs. The distinction we offered here emphasizes the power that the designers of the technology have. Hence, the attractiveness of PbD, but also its limitations. PbD expects the developer to implement privacy in the technology, but other than this statement, the developer is left to her own devices to figure what this actually means.

**E. Success?**

There has been much talk about PbD in recent years: the increasing understanding of risks to our privacy, the continuous search for innovative modes of regulation, and the attractive features of PbD explain this interest. However, it seems that the initial enthusiasm is fading a bit and is replaced with some sobering up. Being regulation by code, PbD poses further regulatory challenges, as a legislature can, at most, require a PbD procedure, rather than the deployment and design of specific technologies. Perhaps it is no surprise that the emphasis has slightly shifted from technology to organizational requirements.[87] The comparison to PETs further illuminates some of PbD's inherent shortcomings: it should be adopted by the designer of the technology who might not have an incentive to do so; there is no external review of the procedure (but the market might serve as a check), and it needs to be specific to the application technology at stake.

Scholars began raising questions about PbD. Diaz et al note that "Even the concept of 'privacy by design,' which some initially thought was meant to embed principles of data minimization and anonymization into product engineering, is increasingly translated to introducing FIPPs compliance into organizational processes."[88] Davies argues that "Presently,

---

[86] For the notion of meeting points, *see* Michael Birnhack & Niv Ahituv, *Privacy Implications of Emerging & Future Technologies* (PRACTIS Project, 2013), available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2364396.

[87] *See supra* note 52.

[88] Diaz et al, *Hero or Villain*, *supra* note 74, at 931.

Privacy by Design is more a concept than a technique."[89] He further observes, referring to Ontario's Information and Privacy Commissioner's promotion of PbD that "Ontario offers a superb motivational platform, but provides little substantive engineering advice,"[90] and concludes that "PbD has become a fashionable idea, and in the wake of fashion came the pretenders that falsely claim their organisations or products have a genuine commitment to the PbD process."[91] Rubinstein and Good similarly observe that Ontario's "seven principles are more aspirational than practical or operational."[92] Brown and Marsden comment that while the idea of PbD has been discussed since the 1990s, "it has taken the threat of enforcement action to persuade some companies to take these principles seriously."[93]

Indeed, most PbD examples mentioned thus far in the literature are about what can or should be done,[94] hypothetical case studies,[95] what could have been done but was not done,[96] or, in Canada, there are PbD examples as to public or regulated bodies.[97]

Why aren't there more success stories? The answer lies in several domains. *First*, privacy itself is a contested, controversial and complicated concept. There is no one agreed-upon definition, nor is there an agreement about the precise composition of FIPPs, or their meaning. For

---

[89] Davies, *Why Privacy by Design, supra* note 60, at 4.

[90] *Id.* at 6.

[91] *Id.* at 9.

[92] Ira S. Rubinstein & Nathaniel Good, *Privacy by Design: A Counterfactual Analysis of Google and Facebook Privacy Incidents*, 28 BERKELEY TECH. L.J. 1333, 1338 (2013).

[93] IAN BROWN & CHRISTOPHER T. MARSDEN, REGULATING CODE: GOVERNANCE AND BETTER REGULATION IN THE INFORMATION AGE 66 (2013).

[94] *See* e.g., Yang Wang, Pedro Giovanni Leon, Kevin Scott, Xiaoxuan Chen, Alessandro Acquisti, Lorrie Faith Cranor, *Privacy Nudges for Social Media: An Exploratory Facebook Study*, PSOSM (2013) (examining privacy nudges developed for Facebook, to help users avoid regrettable postings); Jennifer King, *"How Come I'm Allowing Strangers to Go through My Phone?"— Smartphones and Privacy Expectations*, SOUPS (2013) (discussing PbD in the context of smartphones); Marc van Lieshout & Linda Kool, *Privacy Implications of RFID*, 262 IFIP 129 (2008) (RFID tags in railway systems); Dorothy J. Glancy, *Privacy in Autonomous Vehicles*, 52 SANTA CLARA L. REV. 1171, 1226 (2012) (suggesting PbD in the context of autonomous vehicles).

[95] *See* e.g., a proposed anonymous e-petition and a system of electronic toll pricing, discussed in Seda Gürses, Carmela Troncoso & Claudia Diaz, *Engineering Privacy by Design*, 4th International Conference on Computers, Privacy & Data Protection (2011).

[96] For example, Rubinstein and Good discuss ten counterfactual cases of Google and Facebook services and show how the privacy incidents might have been avoided if engineering principles that reflect privacy were put in place. *See* Rubinstein & Good, *Privacy by Design*, *supra* note 92, at 1377-1406.

[97] *See* Ann Cavoukian, Privacy and Video Surveillance in Mass Transit Systems: A Special Investigation Report (2008); Independent Electricity System Operator and Information and Privacy Commissioner, Building Privacy into Ontario's Smart Meter Data management System: A Control framework (2012).

example, even if all would agree that notice is an important privacy principle, that the data subject is notified about data collection, its purpose, and use before collection begins, there are many very different ways to implement consent: for example, should it be in the form of opt-in or opt-out? The difference matters, as default rules tend to stick.[98]

*Second*, the corporations' incentives do not favor PbD. The process of implementing privacy into technology might be expensive in itself. In the absence of a binding legal requirement to design privacy into technology, businesses are not obliged to engage in such a practice, and at least within permissive legal regimes, such as the American one, the privacy threshold to be met is quite low to begin with. In addition, there seems to be a weak consumer demand for such technologies.[99] Moreover, companies might have an incentive not to apply PbD. In an age of big data, businesses are interested in collecting as much data as they can, from a variety of sources, and use it in ways they are not fully aware of at the time of collection.[100] A big data organizational mindset directly conflicts with fundamental data protection principles such as data minimization, purpose specification, and accordingly, the notice and consent requirements. We shall see more of this conflict later on, in Part IV. Moreover, within the organization, there may be internal challenges,[101] such as who initiates a PbD approach? Does management succeed in conveying its message to the developers?

*Third*, and the most relevant for our current exploration for the explanation of lack of PbD success stories, is the technological challenge: how can the concept of privacy be translated into concrete requirements?[102] Early in the day, Burkert, using the term PETs but in fact referring to what later became known as PbD, argued that "[t]he attraction of PET concepts, and perhaps also one of their main purposes, is . . . that they take the system designers' view of the world and talk

---

[98] Although notice reflects the user's control over the collection of her personal data, it has by and large failed to serve its purpose. *See* Kirsten Martin, *Transaction Costs, Privacy and Trust: The Laudable Goals and Ultimate Failure of Notice and Choice to Respect Privacy Online*, 18(12) FIRST MONDAY (2013).

[99] *See* Rubinstein, *Regulating Privacy by Design*, *supra* note 62, at 1436.

[100] For a discussion of the challenges big data challenges poses to the law, *see* Omer Tene & Jules Polonetsky, *Big Data for All: Privacy and User Control in the Age of Analytics*, 11 NW. J. TECH & INTELL. PROP. 239, 256-63 (2013). The authors then propose legal changes to better fit big data.

[101] *See* e.g., Davies, *Why Privacy by Design, supra* note 60, at 6. The organizational challenges increase if one focuses on PbD in its organizational aspect rather than the technological one. *See* Kenneth A. Bamberger & Deirdre K. Mulligan, *Privacy on the Books and on the Ground*, 63 STAN. L. REV. 247 (2011).

[102] *See* also Nissenbaum, *From Preemption to Circumvention*, *supra* note 3 (explaining the barriers in implementing a PbD system).

to the designers in their own language. The system designers' view, however, is one of abstraction followed by formalization."[103] What was perceived at the time as an attraction, turned out to be one of the main barriers of PbD: deep gaps between lawyers and designers. Gürses et al concluded that "Despite its comprehensiveness, it is not clear from Cavoukian's document, what 'privacy by design' actually is and how it should be translated into the engineering practice."[104] Le Métayer reached a similar conclusion: "One must admit however that the take-up of privacy by design in the industry is still rather limited." He then pointed to the absence of binding legal rules requiring PbD and lack of incentives as possible explanations, and then offered a general methodology.[105] As Rubinstein and Good argued, "Privacy by design requires the translation of FIPs into engineering and design principles and practices."[106] Another example for this challenge is Ian Brown's discussion of the implementation of smart meters in the UK. He found that little attention was paid to privacy in the programs' early stages, and by the time privacy entered the discussion, key decisions had already been made.[107] In exploring the reason for this late introduction of privacy into the discussion, Brown explains that those involved in the process ridiculed PETs and PbD. Thus, the gap between the law and engineers is not just one of lack of information, it goes much deeper.

Recall Ontario's seven principles – they are general statements, but important as they are, they do not (nor do they intend to) provide the designer with a useable working check-list.[108] The idea of PbD requires concretization. A few scholars attempted to answer the technological challenge, based on current data protection law. Recall Langheinrich's principles for the design of ubiquitous systems.[109] Hoepman suggested a PbD strategy, with eight points: minimize, hide, separate, aggregate, inform, control, enforce, and demonstrate.[110] Hartzog and Stutzman argue

---

[103] Burkert, *Privacy-Enhancing Technologies*, *supra* note 27, at 133.

[104] Gürses et al, *Engineering Privacy by Design*, *supra* note 95, at *3.

[105] Daniel Le Métayer, *Privacy by Design: A Formal Framework for the Analysis of Architectural Choices*, 5 (Research Report 8229, February 2013).

[106] Rubinstein & Good, *Privacy by Design*, *supra* note 92, at 1341.

[107] Ian Brown, *Britain's Smart Meter Programme: A Case Study in Privacy by Design*, 28 INT'L REV. L. COMP. & TECH. 172 (2014).

[108] *See supra* note 36.

[109] Langheinrich, *Privacy by Design*, *supra* note 34.

[110] Jaap-Henk Hoepman, *Privacy Design Strategies*, arXiv:1210.6621v2 (2013).

that privacy itself is too broad and opaque a concept and more specifically, that PbD is unfit for the social web. Instead, they offer obscurity as a more helpful concept to protect users' privacy.[111]

Each of the above challenges needs to be addressed. Here, we focus on the technological challenge. Rubinstein and Good articulated the challenge: "FIPs must be translated into principles of privacy engineering and usability and that the best way to accomplish this task is to review the relevant technical literature and distill the findings of computer scientists and usability experts."[112] But the fault is not only with engineers, it is also with the law, to which we now turn.

## III. READING THE LAW

The law often uses technology-neutral language, namely, it does not refer explicitly to any particular technology. However, despite the neutral language, the law cannot be oblivious to its subject matter: regulating any activity requires that the law has some prior perception of how that activity operates or how it should operate. This general proposition applies to informational privacy law as well. This body of law attempts to regulate the flows of personal data. In so doing, the law has an underlying view of information and of the technological systems that enable the collection, processing, and transfer of data. This Part attempts to figure out informational privacy law's underlying attitude as to the technology at stake. This is the law's technological mindset. We read the law by applying an interpretive mode which we call the reverse engineering of the law,[113] and apply it to informational privacy law. We look at both American and European law.

### A. Reverse Engineering the Law

The law is a social tool that executes the political community's choices. In order to achieve this goal, the legislature should understand what is at stake. This applies also to the interpreter of the law (a judge, an attorney, or a citizen interested in knowing the law): to better interpret the law, the interpreter should understand the legislative context.

---

[111] Woodrow Hartzog & Frederic Stutzman, *Obscurity by Design*, 88 WASH. L. REV. 385 (2013).

[112] Rubinstein & Good, *Privacy by Design*, *supra* note 92, at 1341-42. They conclude that "the most reliable way to incorporate privacy design into product development is to include privacy considerations in the definition of software 'requirements' or specifications." *Id.* at 1353.

[113] For a detailed analysis, *see* Michael D. Birnhack, *Reverse Engineering Informational Privacy Law*, 15 YALE J.L. & TECH. 24 (2012).

Such a context always exists. A body of law that regulates a certain activity necessarily holds a view as to how that activity operates or should operate. Contract law, for example, determines how a contract is formed, executed, and enforced. Without an underlying vision as to the way in which people and other legal entities interact, contract law would be at most an arbitrary set of incoherent rules. This is true of any legal field: The law necessarily holds a view as to its subject matter. This view might be explicit and stated in the law itself,[114] in accompanying notes,[115] or in the legislative history: parliamentary (or Congressional) reports, testimonies, and similar sources. But on occasion, a law's conception of its subject matter is hidden, not fully articulated (or not articulated at all), or it has changed over time, so that legislative history is no longer helpful. In such cases, we need to decipher the law and expose its hidden assumptions and underlying conception of its subject matter. This is the reverse engineering of the law.

Legislation often aims for the general, inclusive and flexible standard, rather than the concrete, well-defined and precise rule.[116] One way to do so is to use neutral language in addressing technology. Legislatures make a special effort to use technology-neutral language. The idea is to maintain the law as flexible as possible so as to cover new technologies that might emerge, without the need to constantly amend the law.[117] Such flexibility attempts to answer the popular complaint that the law lags behind technology.

For example, in the context of unauthorized interception of another's conversations, instead of simply saying "telephone," "electronic mail," or the like, the Electronic Communications Privacy Act (ECPA) refers to a "wire communication."[118] Note that this is only partially a technology-specific language, as it still refers to "wire," excluding, for example, print or handwritten messages.

---

[114] *See* e.g., Genetic Information and Nondiscrimination Act of 2008 (GINA), Pub. L. 110-233, 122 Stat. 881, which includes Congressional findings, at §2.

[115] European Directives, for example, are preceded by lengthy "recitals," which typically explain the background of the law. *See* e.g., the 72 recitals of the Data Protection Directive.

[116] For the rules-standards ongoing discussion, see for example Hart's classical example about a law that prohibits vehicles in the park, which seems at first like a clear rule, but turns out to be a standard. *See* H.L.A. HART, THE CONCEPT OF LAW 125-26 (2d ed. 1994); and a critical analysis in Frederick Schauer, *A Critical Guide to Vehicles in the Park*, 83 N.Y.U. L. REV. 1109 (2008).

[117] Bert-Jaap Koops aptly called this feature "futureproofing." *See* Bert-Jaap Koops, *Should ICT Regulation be Technology-Neutral*, *in* STARTING POINTS FOR ICT REGULATION: DECONSTRUCTING PREVALENT POLICY ONE-LINERS 77 (Bert-Jaap Koops, Miriam Lips, Corien Prins & Maurice Schellekens, eds., 2006).

[118] Electronic Communications Privacy Act, codified as 18 U.S.C. §2510(1).

Another example is the Video Privacy Protection Act (VPPA), which regulates data about consumers' video content consumption habits. The VPPA uses the term "video tape," alongside a definition of a service provider, which refers to "prerecorded video cassette tapes or similar audio visual material."[119] Congress enacted the VPPA following publications about Judge Robert Bork's video rental habits (nothing sensational there).[120] The VPPA is rather specific in its technological choice. Thus, faced with a legislative reference to a "video tape," the interpreter is required to figure out whether this can cover also a DVD,[121] or streaming services. The Northern District Court in California replied positively to the latter, concluding that "Congress was concerned with protecting the confidentiality of private information about viewing preferences regardless of the business model or media format involved."[122]

Congress was aware of analogue technologies: the VPPA regulates data about users who rent content that is stored in a physical object, a videotape. The technological mindset was particular: the law assumed a technological structure in which there is a meeting point between the consumer and the vendor at the time of renting the videotape. That is when the data about the transaction is created: who rented which content, when, and where. This analogue mindset survived also when digital storage replaced the analogue, namely, with the introduction of the DVD. The technology changed, but the informational context has not. Renting a DVD does not produce more data than renting a video tape. Hence, even though the Act did not mention DVD, the technological mindset could encompass the new technology. Once technology developed yet again, and instead of renting a DVD users increasingly view content online, on demand, by way of streaming, the informational context changes: now the service provider knows not only which content was borrowed and when, but also when it was actually used, at which physical location, and whether the viewer paused, replayed,[123] or quit viewing in the middle. This is a much richer informational context, but the court found that Congress anticipated the technological developments and chose sufficiently neutral legislative language, even though it did not know what

---

[119] Video Privacy Protection Act, 18 U.S.C. §2710 (2006).

[120] *See* Paul M. Schwartz, *Preemption and Privacy*, 118 YALE L.J. 902, 935-36 (2009).

[121] For a positive answer as to DVD, *see* Schwartz, *Preemption and Privacy*, *id*. at, 912.

[122] *See In re* Hulu Privacy Litig., No. 11-03764, 2012 WL 3282960 at *6 (N.D. Cal. 2012).

[123] When Janet Jackson's breast was exposed in the 2004 Super Bowl, TiVo was able to report exactly how many viewers replayed the scene. *See* Ben Charny, *Jackson's Super Bowl Flash Grabs TiVo Users*, CNET NEWS (February 2, 2004, 3:22 PM PST). We may assume that such systems can easily identify the viewers.

kind of change.[124] The seemingly technological neutral language was anchored in a particular technological mindset, but left sufficient leeway for courts to apply it to new technological situations.

Accordingly, in order to better understand the law and its interaction with technology, we need to figure out the law's technological mindset. Computer scientists and engineers often reverse engineer software and end-products, with the intention of figuring out how the product was designed and what its internal logic, mechanism, and underlying ideas are. Reverse engineering begins with the end-product rather than with the original code or design and works backwards. Applying reverse engineering as a way to read the law means that we begin with the text of the legislation and work backwards, in order to decipher its conceptual building blocks. The proposed reading is interested in the legislation's attitude as to technology: what is the law's underlying technological mindset?

Realizing the law's technological mindset can indicate whether the law is still valid, or whether it needs to be updated. This reading can foster better collaboration between legislators and developers, to produce better laws. This is a legislative advantage. Figuring out the law's technological mindset can assist courts in interpreting the law, and instruct them as to whether they can legitimately interpret a law so as to cover a new technological situation, even if it is not enumerated in the law, or that they should refer the matter to the legislature or authorized executive agency. This is an interpretive advantage. Reverse engineering the law can illuminate the complex relationship between law and technology, and show how each of the two affects the other in a dialectical manner. This is an advantage for those who wish to use technology as code, namely, as a regulatory mode. This is the case of PbD. We can now place informational privacy law under the spotlight.

### B. Reverse Engineering Informational Privacy Law

Informational privacy law has many facets. There is a diversity of legal regimes that apply to personal data, ranging from comprehensive regulation in the European Union, Canada, and a

[124] The situation can be analyzed using Helen Nissenbaum's framework of contextual integrity. *See* HELEN NISSENBAUM, PRIVACY IN CONTEXT: TECHNOLOGY, POLICY, AND THE INTEGRITY OF SOCIAL LIFE (2010). Such an analysis would indicate that there has been a change in the flow of information. Our analysis offers one possible explanation for the change of the context: the hidden technological mindset no longer fits the new technology.

growing number of countries that follow the European model on one end,[125] to no-regulation at all in other countries on the other end, and many points in between. The legal regimes vary in their scope, structure, and mechanisms. However, despite the diversity, there is a rough common denominator: FIPPs.[126] The scope, understanding, and application of FIPPs are controversial and dynamic, especially in light of current proposals to add new principles such as accountability and transparency, the right to be forgotten, data breach notification, and more.[127] However, our purpose here is not to evaluate the law and the proposed reforms, but to reverse engineer it in the search for its technological mindset. We focus on several key features: the initial trigger that operates the law (identifiability); the aggregation and integration of data, which will be relevant for our discussion in Part IV; the principle of data minimization; the data's life cycle; and the notion of the database.

**(1) Identifiability**

In the United States, the trigger for the operation of the informational privacy legal machinery is the content of the data: federal law applies only in an enumerated set of cases, mostly based on the kind of data at stake in a particular sector. For example, these laws cover financial data, health data, genetic information, and other specific cases. Two federal laws diverge from the content-based trigger. One is based on the kind of data subjects—children,[128] and another specific law is based on the kind of the data controller—the government.[129] This sector-specific approach assumes that some information is more risky, from a privacy perspective, than other kinds of data. Importantly, data collected in an unregulated (privacy-wise) sector, is up for grabs: in such unregulated sectors, data subjects do not enjoy any (federal) legal protection as far as their informational privacy is concerned.

In some of these American laws, the criterion of identifiability is added on top of the threshold of content. The law would impose duties on data collectors and grant data subjects rights

---

[125] *See* Birnhack, *An Engine of a Global Regime*, *supra* note 13.

[126] For FIPPs, *see supra* note 52, and Robert Gellman, *Fair Information Practices: A Basic History*, version 2.11 (April 2014), available at http://www.bobgellman.com/rg-docs/rg-FIPShistory.pdf.

[127] *See* GDPR, *supra* note 42; FTC REPORT, *supra* note 46; White House Report, *supra* note 53.

[128] Children's Online Privacy Protection Act of 1998 (COPPA), Pub. L. No. 105-277, 112 Stat. 2581 (1998), codified as 15 U.S.C. §§6501-06.

[129] Privacy Act of 1974, codified as 5 U.S.C. §552a (2006).

only if the data includes Personally Identifying Information (PII).[130] There is a lexical priority here: the PII criterion comes in place only after the content criterion. Thus, PII collected in an unregulated sector is not subject to informational privacy law. The FTC adopts a similar view in its proposed rules for consumer privacy, suggesting that they apply to datasets that are not reasonably identifiable.[131] For example, an American consumer's shopping habits are unregulated, even though the consumer is identified: There is simply no federal law that regulates privacy aspects of consumer data.

European data protection law comes into operation only when there is "information relating to an identified or identifiable natural person," no matter whether the data is sensitive such as one's discrete sexual life or mundane such as one's height.[132] The Data Protection Directive's definition explains: "an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity . . ."[133] Note that the definition does not refer to any particular technology. Nevertheless, we argue that the law holds some image of the technologies that operate in this context, and our task is to expose this technological mindset.

The laws that apply an identifiability criterion trust anonymization as a crucial safeguard for data subjects' privacy. This is at least partially a technological assumption: that anonymization is possible. If the data are truly anonymized, the law will not be applied; in the absence of anonymization, the law would apply. In an influential work, Paul Ohm placed this legal assumption of anonymization under scrutiny.[134] Ohm showed that informational privacy laws often opt for anonymization as a "silver bullet solution,"[135] but, based on contemporary research in the field of computer science, he argued that anonymization is largely an obsolete notion: it is possible to de-anonymize data far more easily than lawyers have thus far assumed. Rephrased in the terms applied here, Ohm reverse engineered a seemingly technologically neutral law and highlighted a crucial technological assumption, and moreover, he argued that the technological

---

[130] *See* Paul M. Schwartz & Daniel J. Solove, *The PII Problem: Privacy and a New Concept of Personally Identifiable Information*, 86 N.Y.U. L. REV. 1814 (2011).

[131] FTC REPORT, *supra* note 47, at 22.

[132] Data Protection Directive, art. 2(a).

[133] *Id.* The proposed GDPR adds technological references of location data, online identifier and genetic identity.

[134] *See* Ohm, *Broken Promises*, *supra* note 66, and critique by Wu, *Privacy and Utility*, *supra* note 66.

[135] Ohm, *id.* at 1736.

assumption no longer holds truth. Indeed, in April 2014, the professional working party of the EU, unattractively called Article 29 Working Party, acknowledged that there is "inherent residual risk of re-identification linked to any technical-organizational measure aimed at rendering data 'anonymous.'"[136]

However, the possibility of de-anonymization does not mean that the law should rid itself of a requirement of anonymization: the more difficult and expensive it is, the motivation to attempt de-anonymization is likely to decrease.[137] If accompanied by legal prohibitions, the incentive to engage in de-anonymization is likely to decrease even more. The FTC, for example, recommends that the collecting company would publicly commit not to re-identify data.[138]

PII-American laws and EU law are not oblivious to de-anonymization. These laws signal to data controllers that they have a choice: to anonymize data and avoid most of the hefty legal requirements,[139] or comply with its requirements. Exposing the law's technological mindset indicates that inasmuch as the law assumed that anonymization is possible, the assumption collapses, but the law assists anonymization by other means.

**(2) Aggregation and Integration**

The discussion of identifiability contains yet another of the law's technological assumptions, which we reverse-engineer. The comparison between the EU approach and the American approach is helpful yet again. The American sector-specific approach reflects what we can call an analogue mindset. It assumes that data of different kinds and sources do not mix, or at least not easily so. The concern is the separate bit of data; linking and aggregating data from different sources is not considered a problem.

By contrast, the European disregard to the content of the data reflects a digital mindset. It acknowledges that separate bits of data can be combined together to produce new information: the seemingly innocent pieces of data can be combined to form a whole that is greater than the sum of

---

[136] *See* Article 29 Data Protection Working Party, WP 216 Anonymisation Techniques (2014), at 7, available at http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf.

[137] Indeed, the European Working Party pointed out that data controllers should assess the costs and risks of de-anonymization and that they do so in light of changing technologies. *Id.* at 8-9.

[138] *See* FTC REPORT, *supra* note 47, at 22. A public statement would authorize the FTC to require compliance through its power to enforce the prohibition of deceptive practices. *See* FTC Act, *supra* note 82.

[139] Nevertheless, some requirements may nevertheless apply. *See* WP 216, *supra note* 136, at 11.

its parts. The facts of one's name, age, ethnicity, gender, sexual orientation, profession, biometric data, social relationships, financial status, health, clickstream, consumer behavior, and much more—each on their own might not be considered private or sensitive by some people. But the EU Directive is concerned with the combination of the details. Joining together one's ethnicity with one's profession creates an image, shallow as it may be; adding financial data makes the image a bit richer, until the accumulation of the data creates "our" profile.[140] The profile is created by integrating bits of information, which are then subject to analysis.

EU data protection law acknowledges the possibility of aggregation and integration of data, and reacts. First, the identification-based criterion rather than a content-based criterion means that the law deliberately addresses such situations. Second, some of the Directive's substantive principles refer to the possibility of aggregation and integration, and generally speaking, disapprove thereof. The Directive requires the data collector to inform the data subject of the purpose of collecting the data up front at the time of the collection;[141] receive the subject's consent for the purpose;[142] the purpose should be a legitimate one,[143] and, this stated purpose should be maintained throughout the use of the data.[144] For example, if an insurance company informed a data subject that it will be collecting data from several sources, such as public and private medical service providers, pharmacies, and data publicly available on social networks, and of its purpose (to evaluate the risk), and the subject consented, the data can be used, but only for purposes which are compatible with the initial purpose.[145] If, for example, the insurance company now expands its activities and enters new financial markets, it must not use the data already obtained for the new purpose, unless the subject consents.

---

[140] On profiling, *see* PROFILING THE EUROPEAN CITIZEN: CROSS-DISCIPLINARY PERSPECTIVES (Mireille Hildebrandt & Serge Gutwirth eds., 2008).

[141] Data Protection Directive, art. 10.

[142] Data Protection Directive, art. 2(h) (definition of consent), and art. 7.

[143] Data Protection Directive, art. 6(1)(b).

[144] Data Protection Directive, art. 6(1)(b) ("Member States shall provide that personal data must be . . . collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes.")

[145] For a European interpretation the purpose limitation, *see* Article 29 Data Protection Working Party, WP 203, Opinion 03/2013 on Purpose Limitation (2013), available at http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf. The Working Party provided an expansive interpretation of incompatibility, allowing processing of data for a different purpose as long as the new purpose is compatible with the initial purpose, and offering a "substantive compatibility assessment." *See id.* at 21, 40.

The importance of exposing the law's assumptions about the possibility of integrating data becomes evident once we read the correlative, technological literature, in Part IV.

### (3) Data Minimization

A bedrock principle of data protection law is that of data minimization. The principle means that the data collected should be only that which is required to fulfill the legitimate, stated purpose, to which the data subject had consented. The EU Data Protection Directive requires that the processing of personal data is lawful and fair, and more specifically that personal data must be "adequate, relevant and not excessive in relation to the purposes for which they are collected and/or further processed."[146] The European Data Protection Supervisor explains: "data controllers should collect only the personal data they really need, and should keep it only for as long as they need it."[147] The principle is also found in the 1980 OECD Guidelines, though it is not explicit in the text,[148] and in the 2005 Privacy Framework of the Asian Pacific Economic Cooperation (APEC).[149] The proposed GDPR is even more explicit, and requires that "Personal data must be . . . adequate, relevant, and limited to the minimum necessary in relation to the purposes for which they are processed; they shall only be processed if, and as long as, the purposes could not be fulfilled by processing information that does not involve personal data."[150] In the United States, lacking a general protection of personal data, there is no universal data minimization principle, other than in

---

[146] Data Protection Directive, art. 6(1)(c).

[147] European Data Protection Supervisor, Glossary, available at http://www.edps.europa.eu/EDPSWEB/edps/EDPS/Dataprotection/Glossary/pid/74 ((2013)

[148] 1980 OECD Guidelines, art. 7-9. Article 7 is about collection limitation ("There should be limits to the collection of personal data . . . "), but does not say that the data should be minimal for the purpose; Article 8 ties the data collected to the purpose with a relevance criterion, and requires that the data should be necessary, and article 9 contains the purpose specification principle. The combination of these articles yields the minimization principle. These principles were maintained also in the 2013 OECD Guidelines, *see supra* note 57.

[149] Asian Pacific Economic Cooperation (APEC), Privacy Framework of 2005, art. 18 ("The collection of personal information should be limited to information that is relevant to the purposes of collection . . ."), available at http://www.apec.org/Groups/Committee-on-Trade-and-Investment/~/media/Files/Groups/ECSG/05_ecsg_privacyframewk.ashx.

[150] GDPR, art. 5(c).

the context of public law, namely a state agency collecting data,[151] and in some of the laws that regulate informational privacy.[152]

Phrased in technological terms, the data minimization principle focuses on a well-defined transaction: an action that is carried out by the technological system, with a specific set of goals, which should be clearly communicated to the data subject and his or her consent should be obtained. This requirement means that the data controller should know in advance the purpose of the collection, and carefully examine which data is needed to fulfill that purpose. Any excess data should not be collected. The data minimization principle limits the controllers' ability to integrate data collected for one purpose with data collected for another purpose. As we shall see in Part IV, this mindset stands in conflict with developers' privacy mindset.

**(4) Data's Lifecycle**

No law requires that the data is treated in a particular sequence, and current law does not contain direct reference to the lifecycle of the data.[153] However, reading informational privacy laws clarifies that in regulating the uses of personal data (when regulated) the law assumes a linear lifecycle of this data: what happens to it in different stages, who has control over the data at each stage, and what is the data's flow. Such a sequence is evident in the 1995 EU Data Protection Directive, especially in its definition of "processing." [numbers added to facilitate the discussion]:

> "any operation or set of operations which is performed upon personal data, whether or not by automatic means, such as [1] collection, [2] recording, [3] organization, [4] storage, [5] adaptation or alteration, [6] retrieval, [7] consultation, [8] use, [9] disclosure by transmission, [10] dissemination or [11]

---

[151] *See* Privacy Act 1974, 5 U.S.C. §552a(e)(1) ("Each agency that maintains a system of records shall- . . . maintain in its records only such information about an individual as is relevant and necessary to accomplish a purpose of the agency required to be accomplished by statute or by executive order of the President."); U.S. Department of Homeland Security, Privacy Policy Guidance Memorandum 2008-01, 4 (2008), available at http://www.dhs.gov/xlibrary/assets/privacy/privacy_policyguide_2008-01.pdf ("DHS should only collect PII that is directly relevant and necessary to accomplish the specified purpose(s) and only retain PII for as long as is necessary to fulfill the specified purpose(s).")

[152] *See* e.g., Financial Services Modernization Act, 1999, better known as the Gramm-Leach-Bliley Act, 15 U.S.C. §6802(c) ("Limits on reuse of information.")

[153] The proposed GDPR did not refer to data's lifecycle, but LIBE's amendments, as approved by the EU Parliament in March 2014, explicitly refer to "the entire life cycle management of personal data" in the context of PbD. *See supra* note 44. The FTC 2012 Report proposed, in the context of PbD, that "Companies should maintain comprehensive data management procedures throughout the life cycle of their products and services." *See* FTC REPORT, *supra* note 46, at 30-32. The White House 2012 Report does not use this term.

> otherwise making available, [12] alignment or combination, [13] blocking, [14] erasure or [15] destruction."[154]

The definition begins with a broad, inclusive statement ("any operation") and is then accompanied by a list of activities. The list is illustrative ("such as") and hence, legally, the definition will cover new situations quite easily: either they would fall within a specific example or they are within the more general "use."

Reverse engineering of the law is interested in the law's underlying technological assumptions. Hence, we should read the list in a different way. The illustrative list is organized in a particular manner. It is quite apparent that the organizing theme is a chronological sequence. Accordingly, steps 1-2 (collection, recording) describe *input*; steps 3-5 (organization, storage, adaptation) refer to the *management* of the database; steps 6-8 (retrieval, consultation, use) are *internal usage*; steps 9-12 (disclosure, dissemination, making available, alignment or combination) are *output*. Step 13 (blocking) probably refers to external access to the data and if so, it is an aspect of *output*.[155] The last two steps (erasure and destruction) are post-mortem *clean-up*: what happens with the data once it is no longer in use. This reading indicates that the Directive reflects a progressive assumption about personal data. It conceives the data similarly as human beings: it is born, grows up, becomes productive and ultimately, it dies. Read thus, the sequence of the illustrative list assumes a temporal linearity.

The linearity further assumes that there are a few players involved; each appears in a different segment of the structure. The Directive casts a few such players: data subject,[156] data

---

[154] Data Protection Directive, art. 2(b). Compare it to the taxonomy offered by Daniel Solove, which is divided into four clusters: collection, processing, dissemination and invasion. Each cluster is then sub-divided into further kinds of activities. *See* DANIEL J. SOLOVE, UNDERSTANDING PRIVACY 103 (2008). GDPR, art. 4(3), adds to this list "structuring" as the fourth situation in the list and deletes "blocking."

[155] The proposed GDPR omits "blocking" from the list. *See* GDPR, at 41.

[156] Recall that the Directive defines data subject as "an identified or identifiable natural person." Data Protection Directive, art. 2(a).

controller,[157] data processor,[158] a third party,[159] and a recipient.[160] The players in the initial input segment are the data subject and the collector, which is considered by the Directive to be the controller; the chief player in the management and internal usage segments is the data controller, perhaps with the assistance of the data processor; the players in the output segment are the data controller and the recipient of the data. The final segment (clean-up) is in the hands of the data controller.

Each segment has one or more players, and accordingly there are interactions between players. When the data controller meets the data subject, in the input segment, the former should inform the latter of the intended uses of the data. When the data is internally managed, the data controller is on its own, and is subject to various duties with a watchful eye of enforcement agencies and the subject's (rather weak) control that can be exercised via a right to access the data and require it is amended. When the controller meets the recipient (output), this interaction is regulated so to protect the data subject's rights.[161]

The linear data collection and processing mindset and its segmentation fit many technologies with which we are familiar today and the business models that utilize these technologies. We provide data to various service providers (schools, banks, health providers, communication providers, websites, etc.) who then process it in various internal and external ways, until they lose interest in the data. To anticipate the discussion in Part IV, data warehousing, and more so big data, defy many of these socio-technological assumptions.

---

[157] A data controller is "the natural or legal person, public authority, agency or any other body which alone or jointly with others determines the purposes and means of the processing of personal data." Data Protection Directive, art. 2(d). GDPR, art. 4(5) adds "conditions" between the "purposes" and "means."

[158] A data processor is "a natural or legal person, public authority, agency or any other body which processes personal data on behalf of the controller." Data Protection Directive, art. 2(e). GDPR, art. 4(6) maintains this definition.

[159] A third party is "any natural or legal person, public authority, agency or any other body other than the data subject, the controller, the processor and the persons who, under the direct authority of the controller or the processor, are authorized to process the data." Data Protection Directive, art. 2(f). The proposed GDPR omits this definition.

[160] A recipient is "a natural or legal person, public authority, agency or any other body to whom data are disclosed, whether a third party or not" and excludes authorized authorities. Data Protection Directive, art. 2(g). GDPR, art. 4(7) omits the reference to a third party and the exclusion of authorities.

[161] For a full analysis, *see* Birnhack, *Reverse Engineering*, *supra* note 113, at 81-86.

**(5) The Centrality of Databases**

A central stage in the above lifecycle is that of the database: this is where the data is stored, where it is processed, until it is transferred to others. In the United States, the 1973 Ware Report, which led to the enactment of the 1974 Privacy Act, listed its first principle: "there should be no secret personal data keeping systems."[162] European thought was also concerned with the databases. The concentration of data per se raised privacy concerns. The EU Directive treats the database as a fundamental building block of its legal structure. The Directive defines a "data filing system" as follows: "'personal data filing system' ('filing system') shall mean any structured set of personal data which are accessible according to specific criteria, whether centralized, decentralized or dispersed on a functional or geographical basis . . ."[163] The definition reveals the Directive's line of thought: the destination of the personal data is to be included in a filing system, the database. The Directive uses a technology-neutral language, but the sequence described above indicates that it is geared towards the database.

The Directive's definition of the filing system refers to a structured database, leaving unstructured databases and distributed databases that are able to gather data from many sources upon demand, outside its scope. This treatment reflects the concern: structured databases enable tagging and profiling people according to pre-determined specific criteria. This may lead to what Oscar Gandy called the panoptic sort.[164] Classifying people according to criteria set by the data controller sorts them into categories, and not according to their individuality. The data subjects' human face is lost in a pre-structured database. The Directive reflects a concern as to the dehumanization of subjects. The focus on structured database does not mean that unstructured databases were deemed unimportant. Unstructured databases seemed, at least in the early 1990s when the Directive was debated, to be useless and hence less dangerous. Today, with big data, unstructured datasets are subject to data analytics.

Note yet another element in the above definition: a database, per the Directive, may be centralized or decentralized, namely split into several sub-sets. This approach means that a data controller could not evade the Directive just by splitting its database into two or more sub-sets. However, if each database is maintained and used for its own, separate purpose, and the subject

---

[162] *See supra* note 9.

[163] Data Protection Directive, art. 2(c). GDPR, art. 4(4) maintains this definition.

[164] OSCAR H. GANDY, THE PANOPTIC SORT: A POLITICAL ECONOMY OF PERSONAL INFORMATION (1993).

was informed of the purpose and consented to it, this should not be considered a decentralized database, but rather, two separate databases. This is what the Directive calls functional separation.[165] The separate databases cannot be merged.

*

We saw that American informational privacy law pays more attention to the content of the data; it is not concerned with the option of aggregating and integrating personal data, thus reflecting a technological analogue mindset, with the inadvertent result of being favorable towards data warehousing and big data. American law does not contain a universal principle of data minimization, and unless prohibited in a sector-specific law, it allows, by default, the reusing and repurposing of the data.

European data protection law reflects certain technological assumptions about its subject matter personal data. The law assumes that anonymization of the data is possible, and leaves the choice to the data controller; it treats the data along a line of aggregation/integration; it limits the collection of data to the minimum needed for a legitimate purpose and prohibits reusing the data for other, incompatible purposes; it reveals a linear way of thought as to the processing of data, and it is geared towards the inclusion of the data in a structured database.

Almost none of these features are explicit in the statutory text. Reverse engineering the law enabled us to expose these assumptions and reconstruct its technological mindset. We can now turn to the technological field, and ask the parallel question: Does technology's privacy mindset fit the law?

## IV. READING TECHNOLOGY

The law attempts to regulate the flow of personal data, whereas information systems attempt to facilitate it. This broad juxtaposition carries a grain of truth, but it also caricatures both law and technology. As for the law, as the titles of the OECD Guidelines and the EU Data Protection Directive indicate,[166] regulation is meant to facilitate the transfer and trans-border flows of personal data, albeit in a way that protects informational privacy. As for technology, it does not

[165] *See* WP 203, *supra* note 145, at 30.

[166] *See supra* notes 12 and 11, respectively.

have a pre-determined fixed nature that enables data flows: information technologies are often designed to do so, but they can also be designed otherwise.

Having read the law in the previous Part so to decipher the law's technological mindset, this Part undertakes the parallel task, of reading technology so to expose and decipher technology's privacy mindset. As we explained earlier, we use the term *mindset* to refer to the overall doctrine that emerges from the texts, which has its own objectives, language, and characteristics. Technology's privacy mindset encapsulates the technological perception of privacy: the way the players in the technological field understand privacy and conceive it, the designs they consider desirable and legitimate, and the possible patterns that technology downplays and dismisses.

We search for technology's privacy mindset, by reading leading books in the field of data warehousing and data science.[167] A few methodological comments are in place, explaining firstly, what led us to focus on these fields, secondly, why we focus on books, and thirdly, why the particular books.

Engineering discourse has its own mindset: the set of patterns and constraints that determine how technology can and should be designed. Amongst engineering communities, knowledge about technical and methodological solutions is captured in the term *design patterns*: general reusable solutions to common situations a term borrowed from architecture,[168] which gained popularity in software engineering in the 2000s. Patterns capture the architecture of the systems, for example, the way software components communicate and the way data is defined and exchanged in the system. Most of the literature on design patterns is rather general, outlining patterns to a wide range of software engineering problems, ranging from developing software for a robotic vacuum cleaner to developing an enterprise information system. Given that vacuum cleaners do not usually raise much privacy concerns, we examine information systems, and more specifically, patterns of data warehouse design and data science that are related to privacy.

Data warehouses are massive databases, built using a range of technologies, geared towards managing and processing large quantities of data in organizational, business and government

[167] Elsewhere, we examine technology's privacy mindset by interviewing developers. *See* Hadar et al, *supra* note 8.

[168] *See* CHRISTOPHER ALEXANDER, SARA ISHIKAWA, & MURRAY SILVERSTEIN, A PATTERN LANGUAGE: TOWNS, BUILDINGS, CONSTRUCTION (1977).

domains.[169] Data warehouses are installed in virtually every large enterprise,[170] and are part of products offered by vendors of enterprise information systems, such as SAP, Oracle, or Microsoft. Data warehouses were initiated in the 1990s, and serve as data repositories that fuel business intelligence, data mining, and other analytical frameworks. They rely on the data accumulated in operational systems, financial transactions systems, logging Website clicks, and other ongoing operations. Much of the data is *personal*, either in the European meaning of identifiable data, or the American meaning of specific sectors.[171] Retrospectively, they can be seen as a precursor to current big data technologies, used for similar analytical purposes with larger volumes of data. Thus, data warehousing is especially apt for a privacy-oriented discourse analysis: it is the technology that handles the collection of personal data and enables its processing. Data warehousing fits the paradigmatic case of informational privacy: the corporate threat to privacy. Accordingly, we leave aside the technological literature about other technologies, which pose different threats to privacy, such as social networks or governmental surveillance technologies. Data science, discussed in section C below, is the next step of data warehousing. It applies mining algorithms to vast datasets, to be found in data warehouses and, even more so, in big data, searching for patterns within the data, and then making predictions based on these patterns.

Developers turn to many sources for guidance. These may be professional training, external sources such as books, internal resources within the organization, or informal sources, such as various online forums, open source communities,[172] exchange of knowledge between teams,[173] or colleagues. The choice among the wide array of potential resources depends on the developers' background, the organization, and much more. For example, Balebako et al found that app developers do not have formal privacy training and often search online for answers, or

---

[169] The term *Data Warehouse* is sometimes used to define a particular technology, based on a Rational Database Management software (RDBMS), in contrast to Big Data technologies that handle data in volumes that cannot be handled by traditional databases.

[170] *See* e.g., Mark A. Beyer & Roxane Edjlali, *Magic Quadrant for Data Warehouse Database Management Systems* (Gartner Report, March 2014), available at https://www.gartner.com/doc/2678018.

[171] *See infra* Part III.B(1).

[172] *See* e.g., Andrea Hemetsberger & Christian Reinhardt, *Learning and Knowledge-building in Open-source Communities: A Social-experiential Approach*, 37 MANAGEMENT LEARNING 187 (2006).

[173] *See* e.g., Nikhil Mehta, Dianne Hall, Terry Byrd, *Information Technology and Knowledge in Software Development Teams: The Role of Project Uncertainty*, 51 INFO. MANAGEMENT 417 (2014).

consult friends.[174] The fields of data warehousing and data science are usually a matter of much larger operations than the development of mobile apps, and are relevant to large enterprises. These organizations typically employ developers with formal education who design large databases and data mining algorithms. Professional books are an important source of information, and typically serve as primary reading material in formal education. But we turn to the books for yet another reason: they not only constitute the knowledge in the field, they reflect the state of the art, and more importantly, they encapsulate technology's privacy mindset. This is the Holy Grail we look for in the current study.

Accordingly, the literature that describes the methodology for designing and implementing data warehouses, discussed in sections A and B, provides us with evidence about the design of large-scale data processing systems and allows us to analyze its privacy mindset. We focus on two prominent engineering books. These are Kimball and Ross, *The Data Warehouse Toolkit* (2013), and Inmon, *Building the Data Warehouse* (2005).[175] The two books enjoy a similar stature in enterprise engineering practice, as evident from many citations.[176] Textbooks in various fields, including business intelligence,[177] and information technology,[178] cite these books as the main sources for organizational data warehouse architectures. Literature reviews from 2005,[179] and 2012,[180] position the two books as those that defined and established the field. A managerial review of data warehouse technologies marked Kimball's work as the main authority for developers that design analytical systems.[181] The books offer two technical alternatives for

---

[174] *See* Rebecca Balebako, Abigail Marh, Jialiu Lon, Jason Hong, Lorrie Faith Cranor, *The Privacy and Security Behaviors of Smartphone App Developers*, USEC (2014).

[175] *See supra* notes 4 and 5, respectively.

[176] According to Google Scholar, Kimball and Ross's book is cited 3424 times, and Inmon's book was cited 4831 times. The citation counts includes all editions (last checked, July 12, 2014). Kimball and Ross themselves claim that the book has achieved prominent place. *See* DW TOOLKIT, at 37.

[177] *See* e.g., JIAWEI HAN & MICHELINE KAMBER, DATA MINING: CONCEPTS AND TECHNIQUES 37 (3rd ed. 2011).

[178] PAULRAJ PONNIAH, DATA WAREHOUSING: FUNDAMENTALS FOR IT PROFESSIONALS 18 (2010) ("[Kimball] is among the top authorities in the field of data warehousing and decision support systems.")

[179] Tho Man Nguyen, A Min Tjoa, & Juan Trujillo, *Data Warehousing and Knowledge Discovery: A Chronological View of Research Challenges*, *in* DATA WAREHOUSING AND KNOWLEDGE DISCOVERY 530 (A Min Tjoa & Juan Trujillio eds., 2005).

[180] Ania Cravero, & Samuel Sepúlveda*, A Chronological Study of Paradigms for Data Warehouse Design*, 32(2) INGENIERÍA E INVESTIGACIÓN 58 (2012).

[181] Catherine Ma, David C, Chou, & David C. Yen, *Data Warehousing, Technology Assessment and Management*, 100(3) INDUSTRIAL MANAGEMENT & DATA SYSTEMS 125 (2000).

building data warehouses, and for many years were considered to have rival views.[182] Section C discusses the emerging field of data science, which provides the principles of data mining and data analytics for big data. We read Provost & Fawcett, *Data Science for Business*.[183] The three books were all published in the United States, but enjoy global popularity. It would be interesting to review European books on similar topics and compare the American privacy mindset to the European privacy mindset. We leave this task for another day.

We should emphasize that our task here is not to provide book reviews, but rather to read the books through privacy lenses, searching for direct references to privacy in the text, indirect and subtext references, and its privacy omissions. We do not claim that the books have sole responsibility for constructing technology's privacy mindset. Rather, these books serve as mirror that consolidates the technological mindset, and given their central place in the technological discourse, they have quite likely also contributed to this mindset. We do not fault the authors for not discussing privacy in any satisfactory way: their books are not about privacy. But in order to assess the feasibility of PbD, we search for technology's privacy mindset.

### A. Kimball & Ross, *The Data Warehouse Toolkit*

The first book placed under our privacy-oriented inspection was first published in 1996 under a different sub-title, and is now in its third edition (2013).[184] This is a comprehensive 600 page book, aiming to be, as its 2002 title suggested, a complete guide to data warehousing, i.e., the design of scalable databases that can be mined. *DW Toolkit*'s audience is Information Technology (IT) designers, who develop increasingly complex data warehouses for businesses. *DW Toolkit* provides a fascinating read to sociologists of technology, and for our purposes, to privacy scholars.

---

[182] *See* e.g., Nenad Jukic, *Modeling Strategies and Alternatives for Data Warehousing Projects*, 49(4) COMM. OF THE ACM 83 (April 2006); Mary Breslin, *Data Warehousing Battle of the Giants: Comparing the Basics of the Kimball and Inmon Models*, Winter 2004 BUSINESS INTELLIGENCE JOURNAL 6 (2004).

[183] *See supra* note 6.

[184] RALPH KIMBALL, THE DATA WAREHOUSE TOOLKIT: PRACTICAL TECHNIQUES FOR BUILDING DIMENSIONAL DATA WAREHOUSES (1996). A second edition followed: RALPH KIMBALL & MARGY ROSS, THE DATA WAREHOUSE TOOLKIT: THE COMPLETE GUIDE TO DIMENSIONAL MODELING (2d ed., 2002) (hereafter: DW TOOLKIT 2002 ed.)

**(1) Overview and Intended Audience**

*DW Toolkit* begins with presenting its overall scheme for designing large databases by applying dimensional modeling, innovative when first introduced, and a standard approach today. The 2013 edition broadened the scope to address not only data warehousing, but also its intended use – business analytics. Following an exposition, *DW Toolkit* offers a series of case studies of different kinds, beginning with retail, inventory, procurement, order management, and accounting, continuing with customer relationship management (CRM), human resources management, and then a series of case studies for specific industries: financial services, telecommunications, transportation, education, healthcare, electronic commerce, and insurance. The case studies are followed by overall discussions of the data warehousing and business intelligence (BI) lifecycle, dimensional modeling processes, subsystems and techniques, system designs, and finally, new in the 2013 edition, big data.

Importantly, *DW Toolkit* portrays its readers, IT designers, as serving the businesses for which they build the systems: "DW/BI [data warehouse / business intelligence] systems must be driven from the needs of business users,"[185] and "first and foremost, the DW/BI system must consider the needs of the business."[186] Note that the "user" is the business, rather than end-users or data subjects, to apply privacy parlance. The focus on business is evident in the qualifications that the IT designer reader should have: "With a DW/BI initiative, you have one foot in your information technology (IT) comfort zone while your other foot is on the unfamiliar turf of business users."[187] The designer is instructed to "listen carefully to the business to identify the organization's business processes."[188] The business focus translates into "two primary design drivers," which are ease of use (also referred to as simplicity) and performance.[189]

---

[185] DW TOOLKIT, at xxxiv.

[186] *Id.* at 1. *See* also at 444 ("the business needs are the DW/BI system users' information requirements.")

[187] *Id.* at 4-5.

[188] *Id.* at 70. Later on, the book suggests that this listening should be done with a filter: "The primary focus is uncovering the architectural implications associated with the business's needs. Listen closely for timing, availability, and performance requirements." *Id.* at 417.

[189] *See* e.g., *id.* at 104 (in the context of retail), at 144 (in the context of procurement, "The goal is to reduce complexity by presenting the data in the most effective form for business users.")

With such a strong business-serving approach, it is no surprise that data subjects are almost entirely absent from *DW Toolkit*. Subjects make few guest appearances as *customers*,[190] and *consumers*,[191] but only rarely as data subjects, a term which is not used at all in the book. The declared goals of the system to be designed do not leave much room for other considerations such as privacy. If privacy is to be designed into the system, it has to struggle to enter the toolkit from the outside, and even if successful, it would have a secondary status at most, perceived as a constraint on the design rather than an integral part thereof.

**(2) Privacy: Direct References**

The 2002 edition of *DW Toolkit* devoted a section in its final chapter to privacy, under the heading "political forces demanding security and affecting privacy."[192] This designation located privacy as an external, political issue, rather than an internal and inherent part of the engineering process, driving specific features of the design. However, the discussion implied that privacy was relevant for all fields discussed at the time. The authors mentioned two specific laws, in the field of healthcare (HIPAA),[193] and the protection of children (COPPA).[194] The 2002 edition also cited two privacy books. One was David Brin's dystopian *The Transparent Society*,[195] and the other was Simson Garfinkel's *Database Nation*,[196] from which the authors drew several important privacy principles, though they emphasized notice, data security of various kinds, the data subject's rights to access her data and require its correction, and a mechanism to expunge incorrect, inadmissible or outdated data.[197] These principles are part of FIPPs, but some other privacy principles were omitted: consent (or choice) being the most glaring omission.

---

[190] *See* chapter 3 on retail, chapter 8, on CRM, chapter 10 on financial services, chapter 16 on insurance.

[191] *See* chapter 10 on financial services, chapter 14 on healthcare.

[192] DW TOOLKIT, 2002 ed., at 375.

[193] Health Insurance Portability and Accountability Act (HIPAA) 1996, and its implementing regulations, codified at 45 C.F.R. 164, and its citation in the 2002 ed., at 377. HIPAA was discussed in the final chapter, rather than in the discussion of healthcare.

[194] *See* COPPA, *supra* note 128, and its citation in the 2002 ed., at 377. The reference to COPPA and HIPAA fit Balebako et al's findings in their study of app developers: the developers they interviewed were unaware of any privacy laws with the exception of these two laws. *See* Balebako, *Privacy and Security*, *supra* note 174, at 3.

[195] *See* DAVID BRIN, THE TRANSPARENT SOCIETY: WILL TECHNOLOGY FORCE US TO CHOOSE BETWEEN PRIVACY AND FREEDOM? (1998), discussed in DW TOOLKIT 2002 ed., at 377.

[196] *See* SIMSON GARFINKEL, DATABASE NATION (2000), cited in DW TOOLKIT 2002 ed., at 378.

[197] DW TOOLKIT 2002 ed., at 378-79.

This discussion no longer appears as such in the 2013 edition. The book now devotes two pages to privacy, classified as an issue of "data governance." The current discussion is confined to a new, last chapter on big data, and it is unclear whether the authors think it should apply to the previous chapters as well: there seems to be ambiguous language that "Data governance for big data should be an extension of the approach used to govern all the enterprise data."[198]

The concept of *data governance* enables *DW Toolkit* to apply a division of labor and allocate responsibility: "As core dimensions participating in multiple dimensional models are defined by folks with data governance responsibilities and built by the DW/BI."[199] Moreover, whereas the book refers to a Chief Information Officer (CIO), it does not refer to a Chief Privacy Officer (CPO). The CIO's task does not cover that of a CPO. On the contrary; the CIO is responsible "to break down the historical data silos to achieve information nirvana."'[200]

In other words, governance is not part of the designers' realm: "At a minimum," the authors explain, "data governance embraces privacy, security, compliance, data quality, metadata management, master data management, and the business glossary that exposes definitions and context to the business community."[201] Privacy is first in this list, and explained as "the most important governance perspective."[202] With the references to Brin and Garfinkel no longer included, the 2013 edition does not list any specific privacy principle.

Other than this discussion, privacy is mentioned only in passing, in a few places in this comprehensive book. One mention is found in the course of discussing healthcare: "The patient dimension has historically been challenging, at least in the United States, because of the lack of a reliable national identity number and/or consistent patient identifier across facilities and physicians. To further complicate matters, the *Health Insurance Portability and Accountability Act* (*HIPAA*) includes strict privacy and security requirements to protect the confidential nature of patient information."[203] The hostile tone is evident. Privacy is presented here as a complication and a strict requirement. The chapter does not contain any detailed guidelines as to how to

---

[198] DW TOOLKIT, at 541.

[199] *Id.* at 127.

[200] *Id.* at 377.

[201] *Id.* at 541.

[202] *Id*.

[203] *Id.* at 341-42.

implement HIPAA's requirements. A second mention of privacy appears as a note to designers, concluding a general discussion about compliance: "You should expand the compliance checklist to encompass known security and privacy requirements."[204] A third and last reference of privacy follows as a note: "You should list the data sources and intermediate data steps that will be archived, together with retention policies, and compliance, security, and privacy constraints."[205] What is a *privacy requirement* or a *privacy constraint*? This is not explained. The omission of privacy is particularly apparent—at least to the privacy-oriented reader—when discussing contexts which are regulated in the United States, such as educational data or financial data.[206] Privacy is not mentioned in these discussions.

An interim conclusion is that *DW Toolkit* treats privacy as a secondary, non-design driver, under the responsibility of a data governance department rather than the system's designers. Privacy is not explained. Data subjects are by and large absent from the discussion. The focus is entirely on the business and its needs.

The few direct references to privacy and the privacy omissions are only part of the picture. A more challenging reading is that which reads between the lines, searching for references to elements that are regulated under privacy law. There are numerous subtle references to personal data, and these are not always in line with the conventional privacy toolkit, namely FIPPs. The most important points are identification, the aggregation and integration of data from different sources, meant to be used for different purposes, and data security.

**(3) Identifiability and Anonymity**

One important privacy-friendly element, at least at first sight, in the current edition of *DW Toolkit*, is its application of a single trigger for privacy law, that of identifiability:

> "If you analyze data sets that include identifying information about individuals or organizations, privacy is the most important governance perspective . . . Egregious episodes of compromising the privacy of individuals or groups can damage your reputation, diminish marketplace trust, expose you to civil lawsuits, and get you in trouble with the law. At the least, for most forms of

[204] *Id.* at 446 (in Chapter 19, discussing ETL Subsystems and Techniques).

[205] *Id.* at 448.

[206] *See* Family Educational Rights and Privacy Act, 1974, codified as 20 U.S.C. §1232g (educational data); and Financial Services Modernization Act, 1999, better known as the Gramm-Leach-Bliley Act, codified as 15 U.S.C. §§6801-6809, §§6821-6827 (financial data).

> analysis, personal details must be masked, and data aggregated enough to not allow identification of individuals."[207]

But with a specific reference to Hadoop,[208] that's it. At first sight, this view seems to be in-line with American-PII laws as well as the EU Data Protection Directive's threshold.[209] However, the book does not explain the risk of de-anonymization; it does not refer to technologies that enable re-identification of seemingly anonymous data; and the guidance provided to the designer, other than raising his or her awareness, remains rather vague and general: mask personal details. This is quite a thin approach to privacy, attributing it great importance in the text, but in fact, rendering it almost non-existent.

Moreover, the explicit reference to identifiability does not mesh well with the overall tone of the book. Throughout *DW Toolkit*, designers are instructed to model the system they build with separate lines and dimensions referring to each individual. This means that the data warehouse should be structured in an identifying manner. The readers are instructed to gather as much data as possible, without differentiating personal data from non-personal data.[210] Only towards the end of the book, designers are almost suddenly instructed to mask identities.

The contrast between the overall "collect personal data" spirit and the "mask the identity" instruction is most evident in the discussion of electronic commerce, where the baseline is that the designer is interested in capturing the data subjects' clickstream. Anonymity is presented there as a problem: "The other big frustration with basic clickstream data is the anonymity of the session."[211] *DW Toolkit* explains that the technological challenge is due to several user behaviors, including the provision of false data, several family members using the same computer, or one user using different computers. But the first reason listed is that "Web visitors want to be anonymous. They may have no reason to trust you, the internet, or their computer with personal

---

[207] DW TOOLKIT, at 541-42. The passage also extends privacy to organizations, a view not commonly accepted in the privacy literature, despite Alan Westin's famous definition, which referred to "The claims of individuals, groups, or institutions." *See* ALAN WESTIN, PRIVACY AND FREEDOM 7 (1967). For a discussion of corporations' privacy rights, *see* LEE BYGRAVE, DATA PROTECTION LAW: APPROACHING ITS RATIONALE, LOGIC AND LIMITS (2002).

[208] Hadoop, or more accurately, Apache Hadoop, is "a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models." *See* http://hadoop.apache.org/.

[209] *See supra*, Part III.B(1).

[210] *See* e.g., DW TOOLKIT, at 33 ("Depending on the industry, the list might include date, customer, product, employee, facility, provider, student, faculty, account, and so on.") This list mixes personal data with business data, and treats them alike.

[211] *Id.* at 354.

identification or credit card information."[212] Anonymity is presented as a "challenging problem."[213] Thus, the formal instruction is in line with privacy, but the overall mindset runs against it.

**(4) Aggregation and Integration**

Perhaps to state the obvious, as one can expect from a book about data warehousing, *DW Toolkit* celebrates the collection of data, its integration, and (internal) sharing. The fundamental concept discussed throughout the book is that of ETL: extract, transform, and load data, which we can translate into privacy parlance as the collection and initial processing of the data. The data is needed so that the business can "better understand customer purchases" (in the context of retail),[214] based on "a multitude of geographic, demographic, behavioral, and other differentiating shopper characteristics."[215] In the context of order management, the interest is that "Many organizations want to supplement these historical performance metrics with facts from other processes to help project what lies ahead."[216] And more generally, in the context of customer relation management, *DW Toolkit* states that CRM is "based on the simple notion that the better you know your customers, the better you can maintain long-lasting, valuable relationships with them . . . To do so, the organization must develop a single, integrated view of each customer."[217] Accordingly, the instruction is that "the CRM mindset is to integrate these customer activities. Key customer metrics and characteristics are collected at each touch point and made available to the others."[218]

This CRM mindset, which we rename more broadly as a *Data Warehousing mindset*, is meant to achieve the ultimate goal of the data warehouse: it "is the foundation that supports the panoramic 360-degree view of your customers."[219] The interest in collecting as much data as

[212] *Id.* at 357.

[213] *Id.* at 356.

[214] *Id.* at 74.

[215] *Id.* at 96.

[216] *Id.* at 198.

[217] *Id.* at 230.

[218] *Id.* at 231.

[219] *Id.* at 232.

possible from different sources applies also to employees,[220] and in the financial services,[221] and in fact, any context discussed in the book.

Sharing data within the organization is celebrated, and compartmentalization of data is met with subtle hostility: "In many cases, the source systems are special purpose applications without any commitment to sharing common data such as product, customer, geography, or calendar with other operational systems in the organization. Of course, a broadly adopted cross-application enterprise resource planning (ERP) system or operational master data management system could help address these shortcomings."[222] Thus, a purpose-specific system is considered a shortcoming of the overall system.

The interest in integrating data is also evident in the tasks allocated to the CIO, which affect the organization's information mode: "The folks in the trenches have pledged intent to share data rather than squirreling it away for a single purpose . . . They're clamoring to get rid of the isolated pockets of data while ensuring they have access to detail and summary data at both the enterprise and line-of-business levels."[223]

However, inasmuch as the data is about human beings, isolation of data might be legally required so as to meet the promise made to data subjects, that their personal data would be used for a specific purpose and for that purpose alone. This is the purpose limitation principle, which instructs that data collected for one purpose are not reused for another purpose.[224] The technological data warehousing mindset which *DW Toolkit* assumes, describes, and constructs, is that of analysis of vast amounts of personal data: collecting as much data as possible and integrating it together into one warehouse. The purpose is not known in advance. Under a data warehousing mindset, notifying the data subject about the intended uses of the data cannot be specific. At most, it can be as general as "for any commercial use." Informed consent is also dubious under such circumstances. This mindset does not fit the notice requirement of FIPPs.[225]

---

[220] *Id.* chapter 9, on human resource management.

[221] *Id.* chapter 10.

[222] *Id.* at 17.

[223] *Id.* at 377.

[224] *See supra* note 145.

[225] Tene & Polonetsky, *Big Data for All*, *supra* note 100, at 259-60 discuss the conflict between big data and the purpose limitation principle, from a legal point of view.

Yet more privacy principles are difficult to comply with under a data warehousing mindset – various limitations on the collection of the data. For example, the principle of data minimization, which instructs the data controller to collect only the minimum data needed for the legitimate purpose,[226] is at odds with the data warehousing mindset.

Finally in this context, one of the quotations above mentioned the organizations' interest in the history of their records and the continuous updating thereof.[227] This feature of a Data Warehousing is repeated several times in the book.[228] The idea of deleting data is acceptable only if the data is incorrect or no longer valid.[229] The *right to be forgotten* suggested in the European GDPR and now renamed, *right to erasure*,[230] have no room under the data warehousing mindset.

**(5) Data Security**

One situation where the design instructions and privacy are closer aligned is data security. *DW Toolkit* begins with stating its baseline: "One of the most important assets of any organization is its information," and the logical application to IT designers is that "An organization's informational crown jewels are stored in the data warehouse."[231] It cautions that the data stored is "potentially harmful . . . in the hands of the wrong people,"[232] with the conclusion that "The DW/BI system must effectively control access to the organization's confidential information."[233] Thus, data security is crucial.[234] It is unclear though, whether the book refers to personal data, to trade secrets, or both.

*DW Toolkit* is not a security guide. It couples security with privacy under the responsibility of the data governance department,[235] and in fact, it places security as an opposite to design: "Security . . . often remains an afterthought and an unwelcome burden to most DW/BI teams. The

---

[226] *See supra* Part III.B(3).

[227] DW TOOLKIT, at 230.

[228] *See* e.g., *id.* at 150, 377.

[229] *See* e.g., *id.* at 465.

[230] *See* GDPR, art. 17; LIBE Amendments, *supra* note 44, at 98.

[231] DW TOOLKIT, at 4.

[232] *Id*.

[233] *Id*.

[234] *See* e.g., *id.* at 492.

[235] *Id.* at 446 (referring the designer to senior management) and at 541.

basic rhythms of the data warehouse are at odds with the security mentality; the data warehouse seeks to publish data widely to decision makers, whereas the security interests assume data should be restricted to those with a need to know."[236] If we follow the logic that privacy and security are on the same side of the equation, and security is at odds with "the basic rhythms" of data warehousing (what we have called here the data warehousing mindset), the conclusion should be that privacy too, is at odds with data warehousing.

But delving more into the *DW Toolkit*'s guidance, security has many facets. Security protects the warehouse from external attacks, but also from internal breaches: "A serious security breach is much more likely to come from within the organization than from someone hacking in from the outside. Although we don't like to think it, the folks on the ETL team present as much a potential threat as any group inside the organization."[237] This warning carries specific recommendations, including the installation of comprehensive "authorized access" to all data and metadata stored in the warehouse, as well as keeping records of access to the warehouse.[238]

Data security is an important element of FIPPs.[239] Securing the data from external unauthorized access prevents its abuse by third parties for purposes of which the data subject was not notified and to which she did not consent. Data security also assists in achieving a related informational principle, that of confidentiality. Confidentiality requires that the personal data that the subject provided to the data controller will not be leaked by the controller. Internal security measures, such as a system that requires specific permissions to access certain areas, means that the data is compartmentalized into different sections. Each staff member has authorized access only to some compartments within the data warehouse, thus reducing the risk of deliberate or inadvertent breaches.

*

To summarize the reverse engineering of *DW Toolkit*: we found only few explicit references to privacy, which treat privacy as an external, political constraint on technological design; the book ignores privacy also when it is most apt, when personal data is clearly at stake. The book advocates anonymization so to comply with privacy laws, but does not provide any guidance, and this general

---

[236] *Id.* at 446.

[237] *Id.* at 492.

[238] *Id*.

[239] *See* e.g., 1980 OECD Guidelines, art. 11; Data Protection Directive, Art. 17.

instruction contradicts the book's spirit to collect as much personal data as possible, without separating it from non-personal data. More subtly, many of the design instructions contained in the book that form its data warehousing mindset (e.g., data aggregation and indefinite retention), directly contradict privacy law's mindset, with the exception of data security. However, informational privacy is much more than security.

### B. Inmon, *Building the Data Warehouse*

The second book we closely read was William Inmon's fourth edition of his important work, *Building DW*, published in 2005.[240] Although there are newer books, Inmon's discussion is considered fundamental and highly influential in the fields of analytical systems.[241] Inmon posits his book as offering a different perspective of data warehousing than that of Kimball and Ross, the latter advocating a multidimensional model, and his book advocating a relational model.[242]

#### (1) Overview and Intended Audience

*Building DW* provides a thorough discussion of data warehousing. It begins with discussing the data warehousing environment and its design, moving to fundamental principles of data warehousing, such as granularity and various technological requirements, and then a number of special situations, such as a distributed data warehouse, the business' use of the data warehouse, merging of external data with the data warehouse, migration of the data warehouse, data that originates from the internet, unstructured data, as well as "the really large data warehouse" (chapter 12). *Building DW* continues with "advanced topics" and further business-related aspects, such as costs, and devotes a chapter to "corporate information compliance" (chapter 17), which mentions HIPAA, but deals almost exclusively with financial regulations.

Throughout its 574 pages, *Building DW* does not discuss privacy directly even once. The only indirect reference is to HIPAA. Unlike *DW Toolkit*, there is no discussion of social aspects of data warehousing. By and large, *Building DW* is a technical guide. Its audience is "architects

---

[240] *See supra* note 5.

[241] *See* e.g., PONNIAH, DATA WAREHOUSING, *supra* note 178.

[242] BUILDING DW, at 356.

and system designers,"[243] and they are to serve the business for which they design the data warehouse. Here too, the "user" is the business's data analyst:[244] defined in the glossary as the "person or process issuing commands and messages to the information system." Data subjects are not referred to as such, and make some guest appearances as *customers*,[245] or *internet users*.[246] In any case, they are treated as the object producing the data and the business's target, rather than independent agents.

Key concepts of the privacy discourse, such as identifiability or anonymization are also absent from the discussion. The book's preface explains the concept of data integration: "It was into this mindset that data warehouse was born."[247] In order to figure out the book's privacy mindset, we point to its internal logic, doctrine, or if you wish, to its ideology.

**(2) Aggregation and Integration**

From a technological point of view, the basic principles of data warehousing according to *Building DW* are granularity of data and its partitioning. The former is "[t]he single most important aspect of the design of a data warehouse,"[248] and it means "the level of detail or summarization of the units of data in the data warehouse."[249] The latter means the breakup of data into separate units that can be handled independently. The purpose of the data warehouse is integration of the granular data. The integrated data forms the data warehouse.[250] Integration is the single most important purpose of data warehousing: it is what distinguishes it from its predecessor form of managing data, the master files. These building blocks of data warehousing, per *Building DW*, have direct implications for informational privacy.

---

[243] *Id*. at xxv. The preface to the second edition (included in the 4th ed.) mentioned the "manager and the developer" as its audience, and the third edition mentioned (also included in the 4th ed.) "developers, managers, designers, data administrators, database administrators, and others who are building systems in a modern data processing environment." The managers were omitted from the 4th ed.

[244] BUILDING DW, at 20.

[245] *See* e.g., *id*. at 121 (customer's habits), 297 (profile contains a thumbnail sketch of the customer), 307 (CRM freely gathers demographic data), 333 (historical data plays an important role in understanding the customer).

[246] *Id*. at 290 ("In a word, the clickstream data is the key to understanding the stream of consciousness of the Internet user.")

[247] *Id.* at xxiv.

[248] *Id.* at 41.

[249] *Id.*

[250] *See id*. at 495 (definition of 'Data Warehouse.')

The ideal data warehouse, according to *Building DW*, contains data from a variety of sources ("Data is fed from multiple, disparate sources into the data warehouse,")[251] and should be able to receive and integrate external data.[252] Accordingly, *Building DW* instructs the designer how to integrate data from different sources, by finding intersections. The example is integrating structured data from CRM about customers such as the customer's age, gender, education, and address, with unstructured data such as communications: emails, letters etc.[253] Another example is data about employees: "In the unstructured environment, you have the name, Social Security number, and employee ID. In the structured environment, you have the name, address, telephone number, and employee ID."[254] The ID provides the intersection.

In privacy terms, integrating data from various sources is not prohibited *per se*. However, data subjects who provide data in one context would typically expect that the data is used only within that context. This is the purpose limitation principle: data collected for one purpose should not be used for another purpose, unless the subject has consented to this. Under American law, integrating data from different sources might frustrate one's reasonable expectations about the use of her data.[255] An explicit notice, which enables an informed and free choice, could solve these problems, at least from a legal point of view. However, *Building DW* does not instruct designers to notify users, to gather their consent, or to query the sources of the data for their original purposes.

Indeed, the ideal data warehouse is meant to enable many uses, including unforeseen uses. This view is explicit in *Building DW*: "Data in the data warehouse is able to be used for many different purposes, including sitting and waiting for future requirements which are unknown today,"[256] or "By storing as much data as possible, you can do any kind of analysis that might happen along. Because the nature of DSS [decision support system, i.e., the business use – MB,

---

[251] *Id.* at 30.

[252] *Id.* at 50 ("External data and other data can be freely mixed with data warehouse data in the course of doing exploration and mining.") and chapter 8, which is devoted to integrating external data within the data warehouse.

[253] *Id.* at 307.

[254] *Id.* at 312.

[255] The reasonable expectations test in American privacy law developed in the context of law enforcement, *see* Katz v. United States, 389 U.S. 347, 361 (1967) (Harlan, J., concurring).

[256] BUILDING DW, at 29.

ET, IH] is delving into the unknown, who knows what detail you will need?"[257] The unknown uses are known in advance. This mindset leads to a reversal of the regular mode of designing a data warehouse. Instead of beginning with the requirements, in a top-down manner, the data warehouse is built in reverse. Hence, as *Building DW* argues, the term design is inaccurate, "because it suggests that elements can be planned out in advance. The requirements for the data warehouse cannot be known until it is partially populated and in use."[258] Accordingly, the instruction is to begin with the data: "Once the data is in hand, it is integrated and then tested to see what bias there is to the data, if any. Programs are then written against the data. The results of the programs are analyzed, and finally the requirements of the system are understood."[259] Thus, the result is "a classic data-driven development life cycle."[260]

These instructions stand in direct contrast to FIPPs, which require that the data subject is notified about the uses of her personal data before the fact, and that the data is not further used for purposes to which she has not consented. Moreover, in a data-driven design, transparency is difficult, perhaps even impossible. The data controllers cannot say what the data's uses are. They might not know themselves.[261]

The commandment of integration raises the challenge of size. Too big a data warehouse, Inmon warns, might be expensive, limit the effective use of the data as important trends can hide behind endless records,[262] and not foster reuse of data.[263] Historical data is one of the causes of the expansion of the data warehouse. Here, *Building DW* points to an interesting rivalry between the technologists (not necessarily the data warehousing designers) and the businesses that they are supposed to serve. The quantity of the data affects performance (or at least this was the state of art in 2005), and this leads to a conflict: "So, naturally, the systems programmer and the applications developer removed historical data as quickly as possible to get good response time.

---

[257] *Id.* at 253.

[258] *Id.* at 71.

[259] *Id.* at 21.

[260] *Id.* at 22.

[261] *See* the legal analysis of Tene & Polonetsky, *Big Data for All*, *supra* note 100.

[262] BUILDING DW, at 254 ("Given very large amounts of data to be processed, important trends and patterns can hide behind the mask of endless records of detailed data.")

[263] *Id.*

But historical data plays an important role in understanding the customer."[264] *Building DW* chooses the business' side, and repeatedly emphasizes the importance of historical data. For example, in listing technological challenges, the book states that "A second major obstacle is that there is not enough historical data stored in the applications to meet the needs of the DSS request,"[265] and "Furthermore, the more historical detailed data you can get, the better, because you never can tell how far back you need to go to do a given DSS analysis."[266] The book provides an example of the value of historical data in the context of CRM, and concludes: "when a corporation understands the history of a customer, the corporation is in a position to be proactive in offering products and services."[267]

Collecting vast amounts of data over time has at least three implications for informational privacy. First, maintaining data over time might violate the data subject's reasonable expectation (in American parlance) or the purpose limitation principle (FIPPs parlance). Second, preserving data overtime contradicts the right to be forgotten.[268] Third, old data might lose its accuracy or relevance over time. Here, the interests of the data warehouse and the data subjects, as reflected in FIPPs, converge. *Building DW* addresses the issue of incorrect data,[269] advocates integrity of the data, defined as "the property of a database that ensures that the data contained in the database is as accurate and consistent as possible,"[270] and acknowledges that "Every piece of information — external or otherwise — has a useful lifetime."[271]

**(3) Additional Design Principles**

There are few other comments in *Building DW* that provide some more clues to its privacy mindset. Generally, a centralized data warehouse is needed, but the author is willing to accept in some cases that a distributed data warehouse, for example, may occur when the business is geographically

---

[264] *Id.* at 333.

[265] *Id.* at 13.

[266] *Id.* at 253.

[267] *Id.* at 426.

[268] *See* GDPR, art. 17.

[269] BUILDING DW, at 67.

[270] *Id.* at 498 (definition of 'Integrity').

[271] *Id.* at 267.

distributed. In such a case, *Building DW* suggests that there are a local data warehouse and a global data warehouse.[272] Inmon's use of the local/global distinction does not necessarily mean a data warehouse located abroad, but the international global aspect is important. The legal implication is that a distributed structure enables global businesses to comply with local data protection laws. For example, a multinational cooperation that operates both in the United States and in Germany, will be able to design its system with a local data warehouse, each complying with the local laws, to the extent that they apply. Under European law, however, one database split into two or more databases might nevertheless be treated as one. The functional separation would be a key to the decision whether this is one or more databases.[273]

Another privacy-related clue in the book is about organizational aspects. As mentioned above, *Building DW* is addressed to designers, and they in turn are to serve the business. This implies that the designers have an executive role, but not affect the requirements (though, recall, a data warehouse is said to be data-driven, rather than requirement-driven.) In comparing commercial data warehouse to a governmental data warehouse, the author comments that "The political issues of data sharing are still up to the politicians."[274] In other words, the further uses of the data is portrayed as an external, non-technological decision, to be made by politicians as far as governmental databases are at stake, and the logical conclusion is that in corporate data warehouses, similar decisions are to be made by the business, rather than by the designers.

Finally, the book observes that "In the commercial world, security for data warehouses is lax (and this is probably an understatement). Very little emphasis is put on the security of a commercial warehouse. The commercial impetus is to get the warehouse up and running and to start to use the warehouse. Most organizations think of warehouse security as an afterthought."[275] This rather surprising comment, we may speculate, was true at the time of the book's publication, or perhaps, it assumed that this is an internal system and secured by the enterprise's firewall. In any case, for our purposes here, this statement indicates that technology's privacy mindset lacks awareness of yet another informational privacy principle, that of data security, inasmuch as the data warehouse contains personal data.

---

[272] *Id.* at 194.

[273] *See supra* note 163.

[274] BUILDING DW, at 405.

[275] *Id.* at 405.

*

The overall reading of Inmon's *Building DW* suggests that informational privacy is simply not within the scope of data warehouse designers' attention or responsibility. They are not guided or informed about privacy matters. To the contrary. The data warehousing mindset that emerges from this book advocates the collection of more data, from diverse sources, for longer periods of time, integrating the data so to serve future and yet-unknown purposes, which, inevitably, the data subject is unaware of. This mindset frustrates FIPPs such as notice, consent, data minimization, and purpose limitation, and would frustrate the right to be forgotten, if it enters the hall of FIPPs.

### C. Provost & Fawcett, Data Science for Business

Data science, dealing with big data, is a new technological paradigm in data analytics, rather than just "more of the same."[276] It has evolved based on theories and techniques from the fields of data warehousing, data mining, statistics, and other fields, in order to utilize the growing accessibility to data, and so it is highly applicable (although not restricted) to big data. As the field of data science is still relatively new, it is too early to point to any single book as a canonical text. We chose to focus on *Data Science for Business*. The book offers an overall scheme of data science, in which big data is one component.[277] *DS Business* follows a style similar to the data warehousing books we discussed earlier, and addresses the design of technological solutions to business problems. Its focus goes beyond the technological implementation, and emphasizes the business use of the technology. The book is relevant for our study, as it relates directly to data warehousing, tying data mining to the integration process of data from multiple sources in data warehouses, which allows us to compare the privacy mindset in the data warehousing books with the *DS Business*'s mindset. Finally, the authors are renowned experts in the field, and at least one of them has done some work on privacy.[278] Again, our purpose is to reverse engineer the book so to expose its underlying privacy mindset.

---

[276] *See* Michael Birnhack, *S-M-L-XL Data: Big Data as a New Informational Privacy Paradigm*, *in* BIG DATA AND PRIVACY: MAKING ENDS MEET 7 (Future of Privacy Forum & Center for Internet & Society, Stanford Law School, 2013).

[277] DS BUSINESS, at 8-9, 17.

[278] *See* Professor Foster Provost, NYU Stern, http://people.stern.nyu.edu/fprovost/.

### (1) Overview and Intended Audience

*DS Business* discusses methods for making sense of data: "extracting useful knowledge from data,"[279] which is the science of data. The primary goal of data science is to solve business problems.[280] Vast amounts of data are collected, matched with other sources, and analyzed; patterns are observed, and predictions are made. Personal data about human beings is central in this field, intended for numerous uses, such as targeted marketing and online customized advertising and recommendations.[281]

The book addresses three audiences: business people who manage data science projects, developers who implement data science solutions, and data scientists.[282] These audiences are similar the two books on data warehousing we discussed earlier. The book's goal is to "align the understanding of the business, technical/development, and data science teams,"[283] and induce "a close interaction between the data scientists and the business people."[284] Accordingly, *DS Business* offers "the most fundamental concepts of data science,"[285] and explains that in addition to these principles, there is room for "intuition, creativity, common sense, and domain knowledge."[286] *DS Business* discusses topics such as predictive modeling, overfitting, finding similarity in data, clustering, various ways to evaluate the models, and much more. Our task here is to figure out if and how privacy fits in in the book's 384 pages: Is privacy part of the fundamental concepts of data science?

### (2) Privacy: Direct References

Privacy appears early on in *DS Business*, but only so to be excluded from its scope. In discussing the now well-known case of Target's prediction about its customers' pregnancy,[287] a footnote

---

[279] DS BUSINESS, at 2.

[280] *Id*. at 28, 31.

[281] *Id*. at 1.

[282] *Id*. at xi.

[283] *Id*.

[284] *Id*. at 13.

[285] *Id*. at xii.

[286] *Id*. at 2.

[287] Target applied data analytics to its customers' purchasing history, identified certain changes that indicated pregnancy, and accordingly, was able to predict pregnancy. Target then tailored advertisements to the presumably

points to the privacy implications of this targeting, stating that "Concerns of ethics and privacy are interesting and very important, but we leave their discussion for another time and place."[288] This bracketing of privacy characterizes the book throughout, though in most cases, privacy is not mentioned at all.

Privacy reappears 216 pages later, as an explanation for not using a certain dataset in earlier examples, "because these attribute names and values have been anonymized extensively to preserve customer privacy. This leaves very little meaning in the attributes and their values, which would have interfered with our discussions."[289] Thus, the authors themselves respect customers' privacy, but present it as an impediment to data analytics. Indeed, anonymization is an obvious privacy preserving strategy, but is only one step in preserving privacy. Measures should be taken to assure that de-anonymization is difficult (we avoid saying that it can be impossible).[290] This aspect is not discussed in the book.

Another reference to privacy is again in passing, in discussing the implications of the shift to mobile devices. The book explains that advertisers may see the location of users, based on their mobile devices, "depending on my privacy settings," thus acknowledging the user's potential power to control the diffusion of her data.[291] This remains a descriptive statement.

Finally, *DS Business* devotes a few paragraphs in the book's Conclusion to *Privacy, Ethics, and Mining Data About Individuals*.[292] The book states that the ethical dimension should not be ignored,[293] and highlights the importance of decisions made based on "detailed data on all of us."[294] The conflict is a direct one: "the more fine-grained data you collect on individuals, the better you can predict things about them that are important for business decision-making."[295] The authors

---

pregnant customers. *See* Neil M. Richards, *The Dangers of Surveillance*, 126 HARV. L. REV. 1934, 1939, 1955 (2013) (discussing the case, and pointing to the option that the targeted customer might not even know she was pregnant). We would add yet another unpleasant situation: imagine the pregnant woman had a miscarriage, but is still tagged as pregnant.

[288] DS BUSINESS, at 7, n.2.

[289] *Id*. at 223.

[290] *See* Ohm, *Broken Promises*, *supra* note 66.

[291] *Id*. at 334.

[292] *Id*. at 341-42.

[293] *Id*. at 341.

[294] *Id*.

[295] *Id*. at 341-42.

raise the option of PbD (not referring to it explicitly), but then point to "Possibly, the biggest impediment to the reasoned consideration of privacy-friendly data science designs," which is defining what privacy is.[296] The authors find support in two notable authorities in privacy studies, Daniel Solove[297] and Helen Nissenbaum.[298] They quote Solove's opening statement, that "Privacy is in disarray," and refer to the length of Nissenbaum's book. The difficulty—defining privacy—is not answered in the book, other than the general suggestion to the data scientist and business stakeholder, that they "should care about privacy concerns, and [ ] will need to invest serious time in thinking carefully about them."[299] However, no further instructions are given; privacy is noted, but only to be distanced from the engineers' desktop, without indicating who should take care of it.[300] The book refers to an online appendix,[301] however, we could not locate it.

**(3) Bracketing Privacy**

*DS Business* discusses all kinds of data: data about objects or businesses as well as data about people. The former cases do not raise privacy concerns, whereas the latter should. But the book treats all kinds of data in the same way, and instructs that to achieve better results, more detailed data is better. For example, the authors explain that "Sociodemographic data provide a substantial ability to model the sort of consumers that are more likely to purchase one product or another. However, sociodemographic data only go so far; after a certain volume of data, no additional advantage is conferred. In contrast, detailed data on customers' individual (anonymized) transactions improve performance substantially over just using sociodemographic data."[302] Facebook, Google, Amazon, and other large service providers serve as repeat examples in the

---

[296] *Id*. at 342.

[297] The taxonomy now appears in SOLOVE, UNDERSTANDING PRIVACY, *supra* note 154, at 103. The taxonomy is an attempt to describe privacy in a socio-legal way, rather than choosing its best meaning and defending it.

[298] NISSENBAUM, PRIVACY IN CONTEXT, *supra* note 124. Similar to Solove's taxonomy, Nissenbaum's approach, while it offers an excellent working framework to identify privacy violations, lacks a theory of privacy. For this criticism, *see* Michael Birnhack, *A Quest for A Theory of Privacy: Context and Control: Review of Helen Nissenbaum's Privacy in Context,* 51 JURIMETRICS: J. L. SCI. & TEC. 447 (2011).

[299] DS BUSINESS, at 342.

[300] This mindset fits our findings in a related study, in which we interviewed developers. *See* Hadar et al, *supra* note 8.

[301] DS BUSINESS, at 342.

[302] DS BUSINESS, at 11.

book, but the privacy implications of the vast amount of personal data are not discussed.[303] The examples that relate to human beings clarify that personal data is not treated in any special way: "In our churn example, a customer would be an entity of interest, and each customer might be described by a large number of attributes, such as usage, customer service history, and many other factors."[304] Or, in illustrating a cellular-phone churn-prediction, Table 3-2 lists users' attributes, such as the customer's level of education and annual income.[305] The book takes it for granted that such data is available to the service provider, without pausing to query the privacy implications of collecting and using such data, which is unrelated to the immediate telecommunication service.

More generally, when describing specific examples and methods, *DS Business* is indifferent to the privacy dimension of personal data. In discussing the evaluation of the data analysis, the book asks and instructs that "often stakeholders are looking to see whether the model is going to do more good than harm."[306] The criteria for such evaluations are the validity and reliability of the data mining results, and that the model satisfies the business goals.[307] Ethical or legal considerations, including privacy, are not part of this evaluation. The subjects' point of view is not taken into consideration.

Practices that would immediately alert the privacy-minded observer do not ring a bell here. For example, in listing common types of data mining tasks,[308] *DS Business* mentions profiling,[309] and provides typical cell phone usage as an illustration. The purpose is to find anomalies in the usage. Profiling people and communication data are constant issues in the privacy realm,[310] but *DS Business* ignores this aspect.

---

[303] *See* e.g., *id*. at 12 (regarding Facebook).

[304] *Id*. at 15.

[305] *Id*. at 74.

[306] *Id*. at 31.

[307] *Id*.

[308] *Id*. at 19.

[309] *Id*. at 22. Profiling is discussed later on in the book in greater detail, *id*. at 296-301.

[310] On profiling, *see* e.g., PROFILING THE EUROPEAN CITIZEN, *supra* note 140. In the United States, communications data is regulated under one of the components of the Electronic Communications Privacy Act, known as the Pen Registers and Trap and Trace Devices Act, codified as 18 U.S.C. §§3121-3127.

A common data mining task is that of "data reduction," which is meant to focus the large dataset.[311] *DS Business* explains the benefits of reduction (datasets are more manageable and can enable better insights),[312] providing an example of movie-viewing preferences: "smaller dataset reveal[s] the consumer taste preferences that are latent in the viewing data."[313] Once again, the fact that this data is often considered personal data does not raise any special attention. Later on, the book returns to movie viewing preferences, and discusses the Netflix Prize, awarded for the best prediction of consumers' movie ratings.[314] *DS Business* comments that the case is famous in data science circles,[315] but, as with the Target case, omits another well-known aspect of the Netflix Prize: the privacy outcry that followed. The company released a dataset so that the competitors can figure out the prediction method, but data scientists matched the data with the publicly accessible IMDb database, which enabled them to expose viewers' identities. A class action followed, and was settled in 2010.[316] *DS Business* refers readers to the Wikipedia entry on the Netflix Prize, which does report the privacy concerns, but this discussion was left outside of the data science realm.[317]

Another example for the disregard to privacy is found in the discussion of targeted marketing,[318] where *DS Business* discusses ways to improve evaluations, and does so without paying attention to the privacy implications, or put in economic parlance, it ignores the externalities of the analytical models. Indeed, *DS Business* explicitly considers data as an asset:[319] data is treated exclusively from the business' point of view, as a property owned by the business. The data subjects have no stake in this asset. This approach is evident in a telling comparison that the book makes, regarding the targeting of online consumers with advertisements: "As consumers, we have become used to getting a vast amount of information and services on the Web seemingly

---

[311] DS BUSINESS, at 22.

[312] *Id*. at 302.

[313] *Id*. at 22-23.

[314] *Id*. at 303-06.

[315] *Id*. at 303.

[316] Doe v. Netflix No. C09-05903 (N.D. Ca., 2010). For discussion, *see* Jane Yakowitz, *Tragedy of the Data Commons*, 25 HARV. J.L. & TECH. 1, 24-27 (2011).

[317] *See* https://en.wikipedia.org/wiki/Netflix_Prize.

[318] DS BUSINESS, at 195.

[319] *Id*. at 207.

for free. Of course, the 'for free' part is very often due to the existence or promise of revenue from online advertising, similar to how broadcast television is 'free'."[320] While both broadcast and many web services are based on an advertisement model, there is an important difference between 'free' broadcast and 'free' web. The broadcast model is interested in the overall number of eyeballs; consumers pay with their attention.[321] The web advertising model is interested in targeting specific users, who pay not only by paying attention, but by providing their personal data.[322] *DS Business* ignores this important difference.

**(4) Prediction**

A central data mining principle is that of prediction. *DS Business* devotes chapter 3 to predictive (as opposed to descriptive) models of data.[323] The purpose of a predictive model is to "estimate[] the unknown values of interest,"[324] such as predicting which customers will default on their credit.[325] This is done by analyzing data that is already in the dataset, and deriving conclusions therefrom. The result is a "general rule," statistically speaking.[326] When the dataset is about human behavior, it may include personal data such as age, income etc.,[327] or as we saw earlier, Target's pregnant customers.

An important element in making predictive models is the "supervised segmentation" of the data,[328] so that large datasets can become more workable, and as a result, more meaningful to its business users.[329] This process requires classifying the data, identifying "informative attributes," and creating tree-structured models that describe the data. Generally speaking, segmentation

---

[320] *Id*. at 233.

[321] *See* Niva Elkin-Koren, *Let the Crawlers Crawl: On Virtual Gatekeepers and the Right to Exclude Indexing*, 26 U. DAYTON L. REV. 179, 183-84 (2001) (comparing television viewers to search engine users).

[322] Tene & Polonetsky, *Big Data for All*, *supra* note 100, at 255 write that "online interactions are barter-like transactions where individuals exchange personal data for free services."

[323] DS BUSINESS, at 43, 46.

[324] *Id*. at 45.

[325] *Id*. at 45.

[326] *Id*. at 47.

[327] *Id*. at 46.

[328] *Id*. at 48.

[329] The interest in useful patterns implies also a set of strategies to avoid chance occurrences that do not generalize. This is the issue of overfitting, discussed in chapter 5 of the book.

requires human decisions as to the structure of the data. An important feature of this process is that tree-structured models may not be the most accurate model.[330]

What are the privacy implications of such an analytical process? Predictive modeling is based on personal data, and thus triggers data protection law during the collection and use of the data. Following the sequence of FIPPs, which follows a linear timeline as to data, we can note that the analysis of datasets does not ask what kind of data is in the dataset, how it was collected, was the data collected for a specific purpose, or did the data subjects attach any strings to its use. The predictive model takes for granted that historical data is already available and accessible, and does not ask how the data was collected. This assumption in itself does not run afoul of privacy principles. It may be the case that the dataset is FIPPs-compliant, and includes only personal data that the subjects consented to its collection and use in this manner. However, consent to a specific purpose is tricky. The exploration of a dataset might be a shot in the dark, leading to *un*supervised segmentation of the data: "we would like to explore our data, possibly with only a vague notion of the exact problem we are solving."[331] The book acknowledges that "Both data scientists themselves and the people who work with them often avoid—perhaps without even realizing it—connecting the results of mining data back to the goal of the undertaking."[332] Thus, if the business does not know what it is looking for in the dataset, it would be difficult to inform the subject and ask for her consent, other than asking for general consent to data analysis. Of course, the deficiency here is not only technological, but legal too. It is hard to imagine how a consent for data mining would look like, other than being a *carte blanche* waiver.

In other words, the creation of datasets and their analysis do not violate privacy rights *per se*. However, while the very purpose of predicting future behavior based on past behavior is not necessarily a matter of privacy, it may harm other interests and rights. These are dignitary harms to a person, namely, that the subjects are not asked to make their own choices about themselves, but are externally characterized and profiled, with someone else making decisions about their prospective behavior. Target's pregnant customer, targeted with tailored advertisements, perhaps without her prior informed consent, is an illustration of the harm to dignity. When the targeting reveals data about the person to others, for example, to the parents of the customer, the harm might

---

[330] DS BUSINESS, at 79.

[331] *Id*. at 182.

[332] *Id*. at 187.

be real. Data protection law (in its European version, and especially its view of privacy as the right to informational self-determination),[333] aims at regulating the collection and use of personal data, *inter alia*, to prevent such harms.

The dignitary harm, which may result from data analytics is not always easy to grasp, and we shall leave the task of developing a full argument for another day.[334] The book's comments about the level of accuracy of the data assist us in illustrating this kind of harm. For the business interested in general patterns of behavior, statistical inaccuracy might not matter much, but for the person who was erred upon, it might mean a great deal: he or she might be denied an opportunity or an entitlement, due to such an inaccuracy. More so, the subject might not even know that she was classified, how she was profiled, what was the prediction made, and how it was applied to her. For the erred-upon subject, the inaccuracy might result in a Kafkaesque situation: that she is manipulated by unseen powers without an explanation. Moreover, the fact that segmentation is derived from large groups of subjects, means that the individual data subject is disregarded as an independent autonomous agent, and is treated in an essentialist way, as belonging to a larger group. Current data protection law tries to empower the subject, so to be able to resist to such harms. Specifically, the data protection principle known as data quality requires *inter alia* that the data is kept accurate.[335] The inherent inaccuracy in predictive modeling, limited as it may be, indicates a discrepancy between law and technology design.

*

To summarize, *DS Business* offers a comprehensive discussion of the emerging business-related aspects of data science. Privacy is not part of this rich picture, although several of the data science practices directly implicate privacy: the data analyzed is often personal data, data science is interested in analyzing as detailed data as possible (and then reduce it to manageable datasets), and the data is used in ways which are not always compatible with privacy principles. *DS Business* presents data analytics as a business and design enterprise combined. The decisions that designers

---

[333] *See* e.g., Antoinette Rouvroy & Yves Poullet, *The Right to Information Self-Determination and the Value of Self-Development: Reassessing the Importance of Privacy for Democracy*, *in* REINVENTING DATA PROTECTION?, 45 (Serge Gutwirth, Yves Poullet, Paul de Hert**,** Cécile de Terwangne, Sjaak Nouwt**,** eds. 2009).

[334] European data protection law is based on the notion of human dignity. *See* Whitman, *The Two Western Cultures*, *supra* note 2.

[335] *See* e.g., Data Protection Directive, art. 6(1)(d).

are required to make along the way are not only technical; they are human decisions, based on business goals, but rarely do they take the subjects' privacy into consideration.

In one of its last chapters, the book addresses the business strategy, and suggests that "the management must create a culture where data science, and data scientists, will thrive."[336] However, privacy is by and large left outside this culture. The data science mindset leaves little, if any, room for privacy.

## V. CONCLUSION: SIGNS OF CHANGE?

Privacy by Design has gained support in regulatory circles over the past few years. The European Union even considers making it a binding legal requirement. This article explored how privacy law views technology and how technology views privacy. Reverse engineering each field so to expose its underlying assumptions and its overall mindset, we found deep, ideological differences between the law's technological mindset and technology's privacy mindset.

One possible conclusion would be pessimistic: PbD is doomed to fail, and we should search for other regulatory modalities to address privacy concerns (or join Mark Zuckerberg in declaring that privacy is no longer a social norm).[337] We opt for the optimist view. Lawyers and engineers can change their view of privacy, or, to borrow from one of the anonymous reviewers of this article, they need to learn how to think of each other in a privacy context in which both engage together. In fact, there might be signs of new and fresh winds blowing in the air. These new winds also point to possible ways for reconciling privacy law and engineering practices. By way of conclusion, we briefly comment on several such avenues, and suggest them as future research. On the engineering side, these potential avenues include professional literature, professional education, organizational privacy climate, and design practices. On the legal side, these avenues include education and incentive-based regulation.

*Professional Literature*. We supplemented our reading of the above-discussed books with another interesting example of the way in which the emerging big data literature treats privacy: a practitioner's viewpoint, offered by Arvind Sathi in 2012.[338] This book offers a genuine attempt

---

[336] DS BUSINESS, at 313.

[337] *See* Bobbie Johnson, *Privacy No Longer a Social Norm, Says Facebook Founder*, THE GUARDIAN (January 11, 2010), available at http://www.theguardian.com/technology/2010/jan/11/facebook-privacy.

[338] ARVIND SATHI, BIG DATA ANALYTICS: DISRUPTIVE TECHNOLOGIES FOR CHANGING THE GAME 1 (2012).

to address privacy issues. Privacy is mentioned explicitly, although at the outset it is designated as a foreign concept within big data. Sathi writes: "The big elephant in the room is data privacy. I confess I have not taken a position on data privacy, nor have I predicted how the world will deal with it. It is an evolving topic, with many complications, geographical differences, and unknown consequences."[339] Later on, there is a direct reference to data subjects (referred to as customers) and to the purpose limitation (not referred to as such in the book): "consumer data can be protected and used only as permitted by the customer."[340] An implicit reference to PbD follows: "As expected, there are many avenues for abuse of customer data, and data privacy must be engrained in the architecture for an effective protection of customer data."[341] Sathi does not avoid the elephant, and points to several policy solutions, such as auditing and the FTC's investigations.[342] But he also points to technological solutions, such as data obfuscation processes,[343] and data masking algorithms.[344] Sathi is aware of privacy concerns, refers to its basic trigger of identifiability and to the purpose limitation principle, and most interestingly, is open to the idea of PbD. This book is an important step in the right direction.

Academic works suggested engineering frameworks for embedding privacy in the design of information systems,[345] and in ubiquitous computing.[346] In *The Privacy Engineer's Manifesto*, published in 2014, the authors offer comprehensive guidance "from policy to code."[347] Their discussion includes detailed translations of legal rules into engineering requirements.[348] Professional literature that acknowledges the importance and relevance of privacy to the design process is much needed. We hope that future editions of the books we read here, on data

---

[339] SATHI, BIG DATA, *id.* at 5-6. *See* also at 73 ("Privacy is a difficult topic that should be handled with care.")

[340] SATHI, *id.* at 24, and *see* also at 46: "If properly managed, the data privacy framework provides gated access to marketers based on permission granted by the consumer and can significantly boost consumer confidence and ability to finance data monetization."

[341] *Id.* at 24.

[342] *Id.* at 44.

[343] *Id*.

[344] *Id.* at 45.

[345] *See* Spiekermann & Cranor, *Engineering Privacy*, *supra* note 80.

[346] Langheinrich, *Privacy by Design*, *supra* note 34.

[347] MICHAELLE FINNERAN DENNEDY, JONATHAN FOX & THOMAS R. FINNERAN, THE PRIVACY ENGINEER'S MANIFESTO: GETTING FROM POLICY TO CODE TO QA TO VALUE (2014).

[348] *Id.* at 99-103.

warehousing and data science will address privacy in their text and no less importantly, in their subtext.

*Education*. Privacy should enter engineers' formal curricular. There are already first courses and study programs devoted to teaching privacy to engineers.[349] These are first indications of a change. Privacy should enter engineers' classrooms.

*Organizations*. Integrating a privacy-oriented approach within an organization is an on-going process, so to make privacy part of the organizational climate. In a related study, based on interviewing engineers, we found that privacy is mostly absent from the typical data driven organization.[350] A promising driver for such a change is the emergence of the new profession of privacy specialists, especially in the United States and the new positions of CPOs in major, multinational companies.[351] Another organizational role that can undertake the mediating role between privacy and technology is that of a knowledge broker within the organization.[352]

*Design processes*. Privacy should become part of the technological specifications and system requirements. Admittedly, this is easier said than done. Introducing privacy into professional literature, education, and organizational practices is worthwhile exploring, but these are long-term changes. In the meantime, until the engineers internalize the importance of privacy to their design, organizations should explore other means *to do* privacy. Put bluntly, there should be a privacy expert in the designers' room.

On the *legal side*, the law should better understand the technological process, its internal trajectories, and sensitivities. A blunt top-down requirement to design privacy into systems, in the spirit of the European GDPR, is unlikely to achieve much. In other words, the law should engage with technology, and lawmakers should converse with designers. Instead of a binding, perhaps somewhat arrogant intervention, the law should explore more subtle ways to encourage PbD. Positive incentives for PbD practices may be an interesting avenue to explore, for example, by providing a safe harbor legal immunity for *bona fide* PbD efforts within an organization.

---

[349] The first is probably Carnegie Mellon University's Master of Science in Information Technology – Privacy Engineering program. *See* http://privacy.cs.cmu.edu/.

[350] *See* Hadar et al, *supra* note 8.

[351] For the rise of the new profession, *see* Bamberger & Mulligan, *Privacy on the Books*, *supra* note 101.

[352] *See* Olga Volkoff, Michael B. Elmes, Diane M. Strongmay, *Enterprise Systems, Knowledge Transfer and Power Users*, 13(4) J. STRATEGIC INFO. SYS. 279 (2004).

None of these options is easy to achieve. There is much more work to bring law and technology closer, but if there is a will, the way is worth exploring.