IFIP AICT 547

Eleni Kosta
Jo Pierson
Daniel Slamanig
Simone Fischer-Hübner
Stephan Krenn
(Eds.)

Privacy and Identity Management

Fairness, Accountability, and Transparency in the Age of Big Data

13th IFIP WG 9.2, 9.6/11.7, 11.6/SIG 9.2.2
International Summer School
Vienna, Austria, August 20–24, 2018
Revised Selected Papers

Springer

TUTORIAL

IFIP Advances in Information and Communication Technology 547

IFIP – The International Federation for Information Processing

IFIP was founded in 1960 under the auspices of UNESCO, following the first World Computer Congress held in Paris the previous year. A federation for societies working in information processing, IFIP's aim is two-fold: to support information processing in the countries of its members and to encourage technology transfer to developing nations. As its mission statement clearly states:

> *IFIP is the global non-profit federation of societies of ICT professionals that aims at achieving a worldwide professional and socially responsible development and application of information and communication technologies.*

IFIP is a non-profit-making organization, run almost solely by 2500 volunteers. It operates through a number of technical committees and working groups, which organize events and publications. IFIP's events range from large international open conferences to working conferences and local seminars.

The flagship event is the IFIP World Computer Congress, at which both invited and contributed papers are presented. Contributed papers are rigorously refereed and the rejection rate is high.

As with the Congress, participation in the open conferences is open to all and papers may be invited or submitted. Again, submitted papers are stringently refereed.

The working conferences are structured differently. They are usually run by a working group and attendance is generally smaller and occasionally by invitation only. Their purpose is to create an atmosphere conducive to innovation and development. Refereeing is also rigorous and papers are subjected to extensive group discussion.

Publications arising from IFIP events vary. The papers presented at the IFIP World Computer Congress and at open conferences are published as conference proceedings, while the results of the working conferences are often published as collections of selected and edited papers.

IFIP distinguishes three types of institutional membership: Country Representative Members, Members at Large, and Associate Members. The type of organization that can apply for membership is a wide variety and includes national or international societies of individual computer scientists/ICT professionals, associations or federations of such societies, government institutions/government related organizations, national or international research institutes or consortia, universities, academies of sciences, companies, national or international associations or federations of companies.

More information about this series at http://www.springer.com/series/6102

Eleni Kosta · Jo Pierson ·
Daniel Slamanig · Simone Fischer-Hübner ·
Stephan Krenn (Eds.)

Privacy and Identity Management

Fairness, Accountability, and Transparency in the Age of Big Data

13th IFIP WG 9.2, 9.6/11.7, 11.6/SIG 9.2.2
International Summer School
Vienna, Austria, August 20–24, 2018
Revised Selected Papers

Springer

Editors
Eleni Kosta
TILT
Tilburg University
Tilburg, The Netherlands

Jo Pierson
Vrije Universiteit Brussel
Brussels, Belgium

Daniel Slamanig
AIT Austrian Institute of Technology
Vienna, Austria

Simone Fischer-Hübner
Karlstad University
Karlstad, Sweden

Stephan Krenn
AIT Austrian Institute of Technology
Vienna, Austria

ISSN 1868-4238 ISSN 1868-422X (electronic)
IFIP Advances in Information and Communication Technology
ISBN 978-3-030-16743-1 ISBN 978-3-030-16744-8 (eBook)
https://doi.org/10.1007/978-3-030-16744-8

Library of Congress Control Number: 2019935815

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This volume contains the proceedings of the 13th IFIP Summer School on Privacy and Identity Management – "Fairness, Accountability and Transparency in the Age of Big Data"—which took place during August 20–24, 2018, in Vienna, Austria.

The 2018 IFIP Summer School was a joint effort among IFIP Working Groups 9.2, 9.6/11.7, 11.6, Special Interest Group 9.2.2 in co-operation with the International Association for Cryptologic Research (IACR) and several European and national projects: the EU H2020 projects CREDENTIAL, PRISMACLOUD, LIGHTest, SECREDAS, VIRT-EU and the German Privacy Forum (Forum Privatheit). It was hosted and also supported by the Austrian Institute of Technology (AIT).

This IFIP Summer School brought together more than 50 junior and senior researchers and practitioners from different parts of the world from many disciplines, including many young entrants to the field. They came to share their ideas, build up a collegial relationship with others, gain experience in giving presentations, and have the opportunity to publish a paper through these proceedings.

One of the goals of the IFIP Summer School is to encourage the publication of thorough research papers by students and emerging scholars. To this end, it had a three-phase review process for submitted papers. In the first phase, authors were invited to submit short abstracts of their work. Abstracts within the scope of the call were selected for presentation at the school and the authors were encouraged to submit full papers of their work. All papers appeared in the unreviewed online pre-proceedings on the school's website. After the school, the authors received two to three reviews by members of the Program Committee and were given time to revise and resubmit their papers for inclusion in these proceedings. In total, the school received 27 short paper submissions. Out of these submissions, ten were finally accepted, including the papers by Sascha van Schendel on "Risk Profiling by Law Enforcement Agencies in the Big Data Era: Is There a Need for Transparency?" and Yefim Shulman and Joachim Meyer on "Is Privacy Controllable?," which were judged to be the summer school's best student papers.

In addition to the submitted papers, this volume also includes reviewed papers summarizing the results of workshops and tutorials that were held at the summer school as well as papers contributed by several of the invited speakers.

We are grateful to all contributors of the summer school and especially to the Program Committee for reviewing the abstracts and papers, and advising the authors on their revisions. Our thanks too to all supporting projects, and especially to the AIT for their support of the activities of the school.

February 2019

Eleni Kosta
Daniel Slamanig
Jo Pierson
Simone Fischer-Hübner
Stephan Krenn

Organization

Program Chairs

Eleni Kosta	TILT-Tilburg University, The Netherlands
Jo Pierson	Vrije Universiteit Brussel imec-SMIT, Belgium
Daniel Slamanig	AIT Austrian Institute of Technology, Austria

General Chairs

Simone Fischer-Hübner	Karlstad University, Sweden
Stephan Krenn	AIT Austrian Institute of Technology, Austria

Program Committee

Felix Bieker	Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein, Germany
Michael Birnhack	Tel Aviv University, Israel
Jan Camenisch	Dfinity, Switzerland
José M. Del Álamo	Universidad Politécnica de Madrid, Spain
David Derler	Dfinity, Switzerland
Gerard Draper Gil	European Commission, Joint Research Centre (JRC), Italy
Zekeriya Erkin	Delft University of Technology, The Netherlands
Hannes Federrath	University of Hamburg, Germany
Simone Fischer-Hübner	Karlstad University, Sweden
Pedro Freitas	Law School of University of Minho, Portugal
Lothar Fritsch	Karlstad University, Sweden
Carlisle George	Middlesex University, UK
Thomas Gross	University of Newcastle upon Tyne, UK
Antti Hakkala	University of Turku, Finland
Marit Hansen	Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein, Germany
Marko Hölbl	University of Maribor, Slovenia
Stefan Katzenbeisser	TU Darmstadt, Germany
Stephan Krenn	AIT Austrian Institute of Technology, Austria
Alessandro Mantelero	Politecnico di Torino, Italy
Leonardo Martucci	Karlstad University, Sweden
Joachim Meyer	Tel Aviv University, Israel
E. Moyakine	University of Groningen, The Netherlands
Steven J. Murdoch	University College London, UK
Igor Nai Fovino	European Commission, Joint Research Centre (JRC), Italy

Ricardo Neisse	European Commission, Joint Research Centre (JRC), Italy
Sebastian Pape	Goethe University Frankfurt, Germany
Norberto Patrignani	Politecnico of Torino, Italy
Robin Pierce	TILT-Tilburg Law School, The Netherlands
Tobias Pulls	Karlstad University, Sweden
Charles Raab	The University of Edinburgh, UK
Kai Rannenberg	Goethe University Frankfurt, Germany
Arnold Roosendaal	Privacy Company, The Netherlands
Ignacio Sanchez	European Commission, Joint Research Centre (JRC), Italy
Viola Schiaffonati	Politecnico di Milano, Italy
Jetzabel Serna-Olvera	Goethe University Frankfurt, Germany
Christoph Striecks	AIT Austrian Institute of Technology, Austria
Tjerk Timan	TNO, The Netherlands
Rosamunde van Brakel	Vrije Universiteit Brussel, Belgium
Frederik Zuiderveen Borgesius	University of Amsterdam, Institute for Information Law, The Netherlands
Harald Zwingelberg	Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein, Germany
Rose-Mharie Åhlfeldt	University of Skövde, Sweden
Melek Önen	EURECOM, France

Additional Reviewer

Rasmus Dahlberg

Contents

Keynotes and Invited Papers

Workshop and Tutorial Papers

Selected Papers

Keynotes and Invited Papers

A Causal Bayesian Networks Viewpoint on Fairness

Silvia Chiappa(✉) and William S. Isaac

DeepMind, London, UK
{csilvia,williamis}@google.com

Abstract. We offer a graphical interpretation of unfairness in a dataset as the presence of an unfair causal effect of the sensitive attribute in the causal Bayesian network representing the data-generation mechanism. We use this viewpoint to revisit the recent debate surrounding the COMPAS pretrial risk assessment tool and, more generally, to point out that fairness evaluation on a model requires careful considerations on the patterns of unfairness underlying the training data. We show that causal Bayesian networks provide us with a powerful tool to measure unfairness in a dataset and to design fair models in complex unfairness scenarios.

1 Introduction

Machine learning is increasingly used in a wide range of decision-making scenarios that have serious implications for individuals and society, including financial lending [10,35], hiring [8,27], online advertising [26,40], pretrial and immigration detention [5,42], child maltreatment screening [13,46], health care [18,31], and social services [1,22]. Whilst this has the potential to overcome undesirable aspects of human decision-making, there is concern that biases in the data and model inaccuracies can lead to decisions that treat historically discriminated groups unfavourably. The research community has therefore started to investigate how to ensure that learned models do not take decisions that are *unfair* with respect to *sensitive attributes* (*e.g.* race or gender).

This effort has led to the emergence of a rich set of fairness definitions [12,15,20,23,37] providing researchers and practitioners with criteria to evaluate existing systems or to design new ones. Many such definitions have been found to be mathematically incompatible [7,12,14,15,29], and this has been viewed as representing an unavoidable trade-off establishing fundamental limits on fair machine learning, or as an indication that certain definitions do not map on to social or legal understandings of fairness [16].

Most fairness definitions focus on the relationship between the model output and the sensitive attribute. However, deciding which relationship is appropriate for the model under consideration requires careful considerations about the patterns of unfairness underlying the training data. Therefore, the choice of a fairness definition always needs to consider the dataset used to train the

Published by Springer Nature Switzerland AG 2019
E. Kosta et al. (Eds.): Privacy and Identity 2018, IFIP AICT 547, pp. 3–20, 2019.
https://doi.org/10.1007/978-3-030-16744-8_1

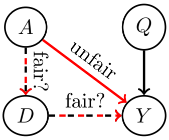
model. In this manuscript, we use the framework causal Bayesian network draw attention to this point, by visually describing unfairness in a dataset as the presence of an unfair *causal effect* of the sensitive attribute in the data-generation mechanism. We then use this viewpoint to raise concern on the fairness debate surrounding the COMPAS pretrial risk assessment tool. Finally, we show that causal Bayesian networks offer a powerful tool for representing, reasoning about, and dealing with complex unfairness scenarios.

2 A Graphical View of (Un)fairness

Consider a dataset $\mathit{\Delta} = \{a^n, x^n, y^n\}_{n=1}^N$, corresponding to N individuals, where a^n indicates a sensitive attribute, and x^n a set of observations that can be used (together with a^n) to form a prediction $\hat{y}^n$ of outcome y^n. We assume a binary setting $a^n, y^n, \hat{y}^n \in \{0, 1\}$ (unless otherwise specified), and indicate with $A, \mathcal{X}$, Y, and $\hat{Y}$ the (set of) random variables[1] corresponding to a^n, x^n, y^n, and $\hat{y}^n$.

In this section we show at a high-level that a correct use of fairness definitions concerned with statistical properties of $\hat{Y}$ with respect to A requires an understanding of the patterns of unfairness underlying $\mathit{\Delta}$, and therefore of the relationships among A, $\mathcal{X}$ and Y. More specifically we show that:

(i) Using the framework of causal Bayesian networks (CBNs), unfairness in $\mathit{\Delta}$ can be viewed as the presence of an unfair causal path from A to $\mathcal{X}$ or Y.
(ii) In order to determine which properties $\hat{Y}$ should possess to be fair, it is necessary to question and understand unfairness in $\mathit{\Delta}$.



Assume a dataset $\mathit{\Delta} = \{a^n, x^n = \{q^n, d^n\}, y^n\}_{n=1}^N$ corresponding to a college admission scenario in which applicants are admitted based on qualifications Q, choice of department D, and gender A; and in which female applicants apply more often to certain departments. This scenario can be represented by the CBN on the left (see Appendix A for an overview of BNs, and Sect. 3 for a detailed treatment of CBNs). The causal path $A \rightarrow Y$ represents direct influence of gender A on admission Y, capturing the fact that two individuals with the same qualifications and applying to the same department can be treated differently depending on their gender. The indirect causal path $A \rightarrow D \rightarrow Y$ represents influence of A on Y through D, capturing the fact that female applicants more often apply to certain departments. Whilst the direct influence $A \rightarrow Y$ is certainly an unfair one, the paths $A \rightarrow D$ and $D \rightarrow Y$, and therefore $A \rightarrow D \rightarrow Y$, could either be considered as fair or as unfair. For example, rejecting women more often due to department choice could be considered fair with respect to college responsibility. However, this could be considered unfair with respect to societal responsibility if the departmental differences were a result of systemic historical or cultural factors (*e.g.* if female applicants apply to specific

[1] Throughout the paper, we use capital and small letters for random variables and their values, and calligraphic capital letters for sets of variables.

departments at lower rates because of overt or covert societal discouragement). Finally, if the college were to lower the admission rates for departments chosen more often by women, then the path $D \rightarrow Y$ would be unfair.

Deciding whether a path is fair or unfair requires careful ethical and sociological considerations and/or might not be possible from a dataset alone. Nevertheless, this example illustrates that we can view unfairness in a dataset as the presence of an unfair causal path from the sensitive attribute A to $\mathcal{X}$ or Y.

Different (un)fair path labeling requires $\hat{Y}$ to have different characteristics in order to be fair. In the case in which the causal paths from A to Y are all unfair (*e.g.* if $A \rightarrow D \rightarrow Y$ is considered unfair), a $\hat{Y}$ that is statistically independent of A (denoted with $\hat{Y} \perp\!\!\!\perp A$) would not contain any of the unfair influence of A on Y. In such a case, $\hat{Y}$ is said to satisfy *demographic parity.*

Demographic Parity (DP). $\hat{Y}$ satisfies demographic parity if $\hat{Y} \perp\!\!\!\perp A$, *i.e.* $p(\hat{Y} = 1|A = 0) = p(\hat{Y} = 1|A = 1)$, where *e.g.* $p(\hat{Y} = 1|A = 0)$ can be estimated as

$$p(\hat{Y} = 1|A = 0) \approx \frac{1}{N_0} \sum_{n=1}^{N} \mathbb{1}_{\hat{y}^n=1,a^n=0},$$

with $\mathbb{1}_{\hat{y}^n=1,a^n=0} = 1$ if $\hat{y}^n = 1$ and $a^n = 0$ (and zero otherwise), and where N_0 is the number of individuals with $a^n = 0$. Notice that many classifiers, rather than a binary prediction $\hat{y}^n \in \{0, 1\}$, output a degree of belief that the individual belongs to class 1, r^n, also called *score.* This could correspond to the probability of class 1, $r^n = p(y^n = 1|a^n, x^n)$, as in the case of logistic regression. To obtain the prediction $\hat{y}^n \in \{0, 1\}$ from r^n, it is common to use a threshold θ, *i.e.* $\hat{y}^n = \mathbb{1}_{r^n>\theta}$. In this case, we can rewrite the estimate for $p(\hat{Y} = 1|A = 0)$ as

$$p(\hat{Y} = 1|A = 0) \approx \frac{1}{N_0} \sum_{n=1}^{N} \mathbb{1}_{r^n>\theta,a^n=0}.$$

Notice that $R \perp\!\!\!\perp A$ implies $\hat{Y} \perp\!\!\!\perp A$ for all values of θ.

In the case in which the causal paths from A to Y are all fair (*e.g.* if $A \rightarrow Y$ is absent and $A \rightarrow D \rightarrow Y$ is considered fair), a $\hat{Y}$ such that $\hat{Y} \perp\!\!\!\perp A|Y$ or $Y \perp\!\!\!\perp A|\hat{Y}$ would be allowed to contain such a fair influence, but the (dis)agreement between Y and $\hat{Y}$ would not be allowed to depend on A. In these cases, $\hat{Y}$ is said to satisfy *equal false positive/false negative rates* and *calibration* respectively.

Equal False Positive and Negative Rates (EFPRs/EFNRs). $\hat{Y}$ satisfies EFPRs and EFNRs if $\hat{Y} \perp\!\!\!\perp A|Y$, *i.e.* (EFPRs) $p(\hat{Y} = 1|Y = 0, A = 0) = p(\hat{Y} = 1|Y = 0, A = 1)$ and (EFNRs) $p(\hat{Y} = 0|Y = 1, A = 0) = p(\hat{Y} = 0|Y = 1, A = 1)$.

Calibration. $\hat{Y}$ satisfies calibration if $Y \perp\!\!\!\perp A|\hat{Y}$. In the case of score output R, this condition is often instead called *predictive parity* at threshold θ, $p(Y = 1|R > \theta, A = 0) = p(Y = 1|R > \theta, A = 1)$, and calibration defined as requiring $Y \perp\!\!\!\perp A|R$.

In the case in which at least one causal path from A to Y is unfair (*e.g.* if $A \rightarrow Y$ is present), EFPRs/EFNRs and calibration are inappropriate criteria, as they would not require the unfair influence of A on Y to be absent from $\hat{Y}$ (*e.g.* a perfect model ($\hat{Y} = Y$) would automatically satisfy EFPRs/EFNRs and calibration, but would contain the unfair influence). This observation is particularly relevant to the recent debate surrounding the *correctional offender management profiling for alternative sanctions* (COMPAS) pretrial risk assessment tool. We revisit this debate in the next section.

2.1 The COMPAS Debate

Over the past few years, numerous state and local governments around the United States have sought to reform their pretrial court systems with the aim of reducing unprecedented levels of incarceration, and specifically the population of low-income defendants and racial minorities in America's prisons and jails [2,24,30]. As part of this effort, quantitative tools for determining a person's likelihood for reoffending or failure to appear, called *risk assessment instruments* (RAIs), were introduced to replace previous systems driven largely by opaque discretionary decisions and money bail [6,25]. However, the expansion of pretrial RAIs has unearthed new concerns of racial discrimination which would nullify the purported benefits of these systems and adversely impact defendants' civil liberties.

An intense ongoing debate, in which the research community has also been heavily involved, was triggered by an exposé from investigative journalists at ProPublica [5] on the COMPAS pretrial RAI developed by Equivant (formerly Northpointe) and deployed in Broward County in Florida. The COMPAS *general recidivism risk scale* (GRRS) and *violent recidivism risk scale* (VRRS), the focus of ProPublica's investigation, sought to leverage machine learning techniques to improve the predictive accuracy of recidivism compared to older RAIs such as the *level of service inventory-revised* [3] which were primarily based on theories and techniques from a sub-field of psychology known as the psychology of criminal conduct [4,9][2].

ProPublica's criticism of COMPAS centered on two concerns. First, the authors argued that the distribution of the risk score $R \in \{1, \ldots, 10\}$ exhibited discriminatory patterns, as black defendants displayed a fairly uniform distribution across each value, while white defendants exhibited a right skewed

[2] While the exact methodology underlying GRRS and VRRS is proprietary, publicly available reports suggest that the process begins with a defendant being administered a 137 point assessment during intake. This is used to create a series of dynamic risk factor scales such as the *criminal involvement scale and history of violence scale*. In addition, COMPAS also includes static attributes such as the defendant's age and prior police contact (number of prior arrests). The raw COMPAS scores are transformed into decile values by ranking and calibration with a normative group to ensure an equal proportion of scores within each scale value. Lastly, to aid practitioner interpretation, the scores are grouped into three risk categories. The scale values are displayed to court officials as either Low (1–4), Medium (5–7), and High (8–10) risk.

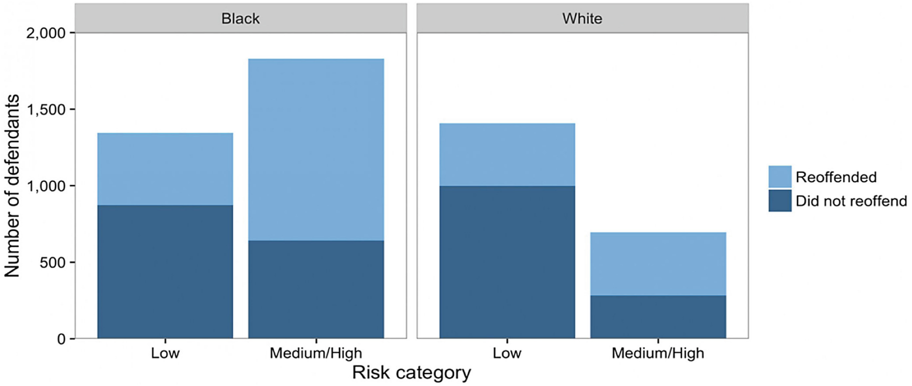


Fig. 1. Number of black and white defendants in each of two aggregate risk categories [14]. The overall recidivism rate for black defendants is higher than for white defendants (52% vs. 39%), *i.e.* $Y \not\perp\!\!\!\perp A$. Within each risk category, the proportion of defendants who reoffend is approximately the same regardless of race, *i.e.* $Y \perp\!\!\!\perp A | \hat{Y}$. Black defendants are more likely to be classified as medium or high risk (58% vs. 33%) *i.e.* $\hat{Y} \not\perp\!\!\!\perp A$. Among individuals who did not reoffend, black defendants are more likely to be classified as medium or high risk than white defendants (44.9% to 23.5%). Among individuals who did reoffend, white defendant are more likely to be classified as low risk than black defendants (47.7% vs 28%), *i.e.* $\hat{Y} \not\perp\!\!\!\perp A | Y$.

distribution, suggesting that the COMPAS recidivism risk scores disproportionately rated white defendants as lower risk than black defendants. Second, the authors claimed that the GRRS and VRRS did not satisfy EFPRs and EFNRs, as FPRs = 44.9% and FNRs = 28.0% for black defendants, whilst FPRs = 23.5% and FNRs = 47.7% for white defendants (see Fig. 1). This evidence led ProPublica to conclude that COMPAS had a disparate impact on black defendants, leading to public outcry over potential biases in RAIs and machine learning writ large.

In response, Equivant published a technical report [19] refuting the claims of bias made by ProPublica and concluded that COMPAS is sufficiently calibrated, in the sense that it satisfies predictive parity at key thresholds. Subsequent analyses [12, 15, 29] confirmed Equivant's claims of calibration, but also demonstrated the incompatibility of EFPRs/EFNRs and calibration due to differences in base rates across groups ($Y \not\perp\!\!\!\perp A$) (see Appendix B). Moreover, the studies suggested that attempting to satisfy these competing forms of fairness force unavoidable trade-offs between criminal justice reformers' purported goals of racial equity and public safety.

As explained in Sect. 2, $R \perp\!\!\!\perp A$ is an appropriate fairness criterion when influence from A is considered unfair, whilst EFPRs/EFNRs and calibration, by requiring the rate of (dis)agreement between Y and $\hat{Y}$ to be the same for black and white defendants (and therefore by not being concerned with dependence of Y on A), are appropriate when influence from A is considered fair. Therefore, if dependence of Y on A includes influence of A in Y through an unfair causal path, both EFPRs/EFNRs and calibration would be inadequate, and the fact that they cannot be satisfied at the same time irrelevant.

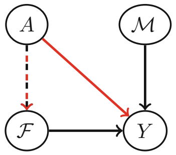


Fig. 2. Possible CBN underlying the dataset used for COMPAS.

As previous research has shown [28,34,43], modern policing tactics center around targeting a small number of neighborhoods—often disproportionately populated by non-white and low income residents—with recurring patrols and stops. This uneven distribution of police attention, as well as other factors such as funding for pretrial services [30,45], means that differences in base rates between racial groups are not reflective of ground truth rates. We can rephrase these findings as indicating the presence of a direct path $A \rightarrow Y$ (through unobserved neighborhood) in the CBN representing the data-generation mechanism (Fig. 2). Such tactics also imply an influence of A on Y through the set of variables $\mathcal{F}$ containing number of prior arrests. In addition, the influence of A on Y through $A \rightarrow Y$ and $A \rightarrow \mathcal{F} \rightarrow Y$ could be more prominent or contain more unfairness due to racial discrimination.

These observations indicate that EFPRs/EFNRs and calibration are inappropriate criteria for this case, and more generally that the current fairness debate surrounding COMPAS gives insufficient consideration to the patterns of unfairness underlying the data. Our analysis formalizes the concerns raised by social scientists and legal scholars on mismeasurement and unrepresentative data in the US criminal justice system. Multiple studies [21,33,36,45] have argued that the core premise of RAIs, to assess the likelihood a defendant reoffends, is impossible to measure and that the empirical proxy used (*e.g.* arrest or conviction) introduces embedded biases and norms which render existing fairness tests unreliable.

This section used the CBN framework to describe at a high-level different patterns of unfairness that can underlie a dataset and to point out issues with current deployment of fairness definitions. In the remainder of the manuscript, we use this framework more extensively to further advance our analysis on fairness. Before doing that, we give some background on CBNs [17,38,39,41,44], assuming that all variables except A are continuous.

3 Causal Bayesian Networks

A *Bayesian network* is a *directed acyclic graph* where nodes and edges represent random variables and statistical dependencies. Each node X_i in the graph is associated with the conditional distribution $p(X_i|\text{pa}(X_i))$, where $\text{pa}(X_i)$ is the set of *parents* of X_i. The joint distribution of all nodes, $p(X_1, \dots, X_I)$, is given by the product of all conditional distributions, *i.e.* $p(X_1, \dots, X_I) = \prod_{i=1}^{I} p(X_i|\text{pa}(X_i))$ (see Appendix A for more details on Bayesian networks).

When equipped with causal semantic, namely when representing the data-generation mechanism, Bayesian networks can be used to visually express causal relationships. More specifically, CBNs enable us to give a graphical definition of causes and causal effects: if there exists a *directed path* from A to Y, then A is a *potential cause* of Y. Directed paths are also called *causal paths*.

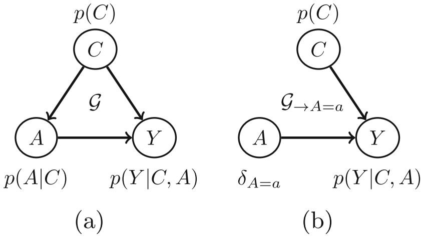


Fig. 3. (a): CBN with a confounder C for the effect of A on Y. (b): Modified CBN resulting from intervening on A.

The causal effect of A on Y can be seen as the information traveling from A to Y through causal paths, or as the conditional distribution of Y given A restricted to causal paths. This implies that, to compute the causal effect, we need to disregard the information that travels along non-causal paths, which occurs if such paths are *open*. Since paths with an arrow emerging from A are either causal or closed (*blocked*) by a *collider*, the problematic paths are only those with an arrow pointing into A, called *back-door paths*, which are open if they do not contain a collider.

An example of an open back-door path is given by $A \leftarrow C \rightarrow Y$ in the CBN $\mathcal{G}$ of Fig. 3(a): the variable C is said to be a *confounder* for the effect of A on Y, as it confounds the causal effect with non-causal information. To understand this, assume that A represents hours of exercise in a week, Y cardiac health, and C age: observing cardiac health conditioning on exercise level from $p(Y|A)$ does not enable us to understand the effect of exercise on cardiac health, since $p(Y|A)$ includes the dependence between A and Y induced by age.

Each parent-child relationship in a CBN represents an autonomous mechanism, and therefore it is conceivable to change one such a relationship without changing the others. This enables us to express the causal effect of $A = a$ on Y as the conditional distribution $p_{\rightarrow A=a}(Y|A=a)$ on the modified CBN $\mathcal{G}_{\rightarrow A=a}$ of Fig. 3(b), resulting from replacing $p(A|C)$ with a Dirac delta distribution $\delta_{A=a}$ (thereby removing the link from C to A) and leaving the remaining conditional distributions $p(Y|A,C)$ and $p(C)$ unaltered – this process is called *intervention* on A. The distribution $p_{\rightarrow A=a}(Y|A=a)$ can be estimated as $p_{\rightarrow A=a}(Y|A = a) = \int_C p_{\rightarrow A=a}(Y|A=a,C)p_{\rightarrow A=a}(C|A=a) = \int_C p(Y|A=a,C)p(C)$. This is a special case of the following back-door adjustment formula.

Back-Door Adjustment. If a set of variables $\mathcal{C}$ satisfies the back-door criterion relative to $\{A, Y\}$, the causal effect of A on Y is given by $p_{\rightarrow A}(Y|A) = \int_{\mathcal{C}} p(Y|A,\mathcal{C})p(\mathcal{C})$. $\mathcal{C}$ satisfies the back-door criterion if (a) no node in $\mathcal{C}$ is a *descendant* of A and (b) $\mathcal{C}$ blocks every back-door path from A to Y.

The equality $p_{\rightarrow A=a}(Y|A=a,\mathcal{C}) = p(Y|A=a,\mathcal{C})$ follows from the fact that $\mathcal{G}_{A\rightarrow}$, obtained by removing from $\mathcal{G}$ all links emerging from A, retains all (and only) the back-door paths from A to Y. As $\mathcal{C}$ blocks all such paths, $Y \perp\!\!\!\perp A|\mathcal{C}$ in $\mathcal{G}_{A\rightarrow}$. This means that there is no non-causal information traveling from A to Y when conditioning on $\mathcal{C}$ and therefore conditioning on A coincides with intervening.

Conditioning on C to block an open back-door path may open a closed path on which C is a collider. For example, in the CBN of Fig. 4(a), conditioning on C closes the paths $A \leftarrow C \leftarrow X \rightarrow Y$ and $A \leftarrow C \rightarrow Y$, but opens the path $A \leftarrow E \rightarrow C \leftarrow X \rightarrow Y$ (additional conditioning on X would close $A \leftarrow E \rightarrow C \leftarrow X \rightarrow Y$).

The back-door criterion can also be derived from the rules of do-calculus [38,39], which indicate whether and how $p_{\rightarrow A}(Y|A)$ can be estimated using observations from $\mathcal{G}$: for many graph structures with unobserved confounders the only way to compute causal effects is by collecting observations directly from $\mathcal{G}_{\rightarrow A}$ – in this case the effect is said to be *non-identifiable*.

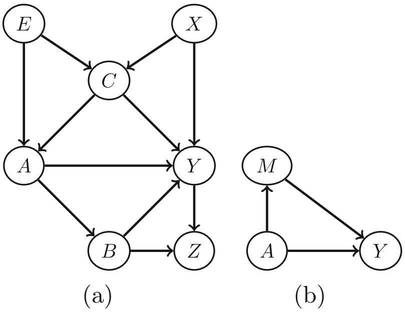


Fig. 4. (a): CBN in which conditioning on C closes the paths $A \leftarrow C \leftarrow X \rightarrow Y$ and $A \leftarrow C \rightarrow Y$ but opens the path $A \leftarrow E \rightarrow C \leftarrow X \rightarrow Y$. (b): CBN with one direct and one indirect causal path from A to Y.

Potential Outcome Viewpoint. Let $Y_{A=a}$ be the random variable with distribution $p(Y_{A=a}) = p_{\rightarrow A=a}(Y|A=a)$. $Y_{A=a}$ is called *potential outcome* and, when not ambiguous, we will refer to it with the shorthand Y_a. The relation between Y_a and all the variables in $\mathcal{G}$ other than Y can be expressed by the graph obtained by removing from $\mathcal{G}$ all the links emerging from A, and by replacing Y with Y_a. If Y_a is independent on A in this graph, then[3] $p(Y_a) = p(Y_a|A=a) = p(Y|A=a)$. If Y_a is independent of A in this graph when conditioning on $\mathcal{C}$, then

$$p(Y_a) = \int_{\mathcal{C}} p(Y_a|\mathcal{C})p(\mathcal{C}) = \int_{\mathcal{C}} p(Y_a|A=a,\mathcal{C})p(\mathcal{C}) = \int_{\mathcal{C}} p(Y|A=a,\mathcal{C})p(\mathcal{C}),$$

i.e. we retrieve the back-door adjustment formula.

In the remainder of the section we show that, by performing different interventions on A along different causal paths, it is possible to isolate the contribution of the causal effect of A on Y along a group of paths.

Direct and Indirect Effect

Consider the CBN of Fig. 4(b), containing the direct path $A \rightarrow Y$ and one indirect causal path through the variable M. Let $Y_a(M_{\bar{a}})$ be the random variable with distribution equal to the conditional distribution of Y given A restricted to causal paths, with $A = a$ along $A \rightarrow Y$ and $A = \bar{a}$ along $A \rightarrow M \rightarrow Y$. The *average direct effect* (ADE) of $A = a$ with respect to $A = \bar{a}$, defined as

$$\text{ADE}_{\bar{a}a} = \langle Y_a(M_{\bar{a}})\rangle_{p(Y_a(M_{\bar{a}}))} - \langle Y_{\bar{a}}\rangle_{p(Y_{\bar{a}})},$$

[3] The equality $p(Y_a|A=a) = p(Y|A=a)$ is called *consistency*.

where *e.g.* $\langle Y_a\rangle_{p(Y_a)} = \int_{Y_a} Y_a p(Y_a)$, measures the difference in flow of causal information from A to Y between the case in which $A = a$ along $A \rightarrow Y$ and $A = \bar{a}$ along $A \rightarrow M \rightarrow Y$ and the case in which $A = \bar{a}$ along both paths.

Analogously, the *average indirect effect* (AIE) of $A = a$ with respect to $A = \bar{a}$, is defined as $\text{AIE}_{\bar{a}a} = \langle Y_{\bar{a}}(M_a)\rangle_{p(Y_{\bar{a}}(M_a))} - \langle Y_{\bar{a}}\rangle_{p(Y_{\bar{a}})}$.

The difference $\text{ADE}_{\bar{a}a} - \text{AIE}_{a\bar{a}}$ gives the *average total effect* (ATE) $\text{ATE}_{\bar{a}a} = \langle Y_a\rangle_{p(Y_a)} - \langle Y_{\bar{a}}\rangle_{p(Y_{\bar{a}})}$[4].

Path-Specific Effect

To estimate the effect along a specific group of causal paths, we can generalize the formulas for the ADE and AIE by replacing the variable in the first term with the one resulting from performing the intervention $A = a$ along the group of interest and $A = \bar{a}$ along the remaining causal paths. For example, consider the CBN of Fig. 5 (top) and assume that we are interested in isolating the effect of A on Y along the direct path $A \rightarrow Y$ and the paths passing through M, $A \rightarrow M \rightarrow, \ldots, \rightarrow Y$, namely along the red links. The *path-specific effect* (PSE) of $A = a$ with respect to $A = \bar{a}$ for this group of paths is defined as

$$\text{PSE}_{\bar{a}a} = \langle Y_a(M_a, L_{\bar{a}}(M_a))\rangle - \langle Y_{\bar{a}}\rangle,$$

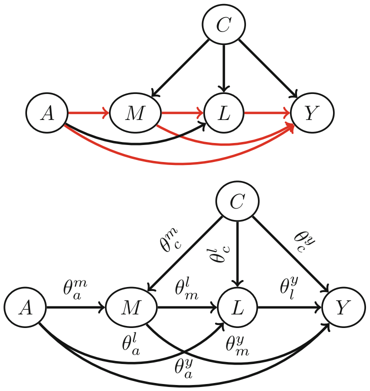


Fig. 5. Top: CBN with the direct path from A to Y and the indirect paths passing through M highlighted in red. Bottom: CBN corresponding to (1). (Color figure online)

where $p(Y_a(M_a, L_{\bar{a}}(M_a)))$ is given by

$$\int_{C,M,L} p(Y|A = a, C, M, L)p(L|A = \bar{a}, C, M)p(M|A = a, C)p(C).$$

In the simple case in which the CBN corresponds to a linear model, *e.g.*

$$\begin{aligned}
&A \sim \text{Bern}(\pi),\ C = \epsilon_c,\\
&M = \theta^m + \theta^m_a A + \theta^m_c C + \epsilon_m,\ L = \theta^l + \theta^l_a A + \theta^l_c C + \theta^l_m M + \epsilon_l,\\
&Y = \theta^y + \theta^y_a A + \theta^y_c C + \theta^y_m M + \theta^y_l L + \epsilon_y,
\end{aligned} \tag{1}$$

[4] Often the AIE of $A = a$ with respect to $A = \bar{a}$ is defined as $\text{AIE}^a_{\bar{a}a} = \langle Y_a\rangle_{p(Y_a)} - \langle Y_a(M_{\bar{a}})\rangle_{p(Y_a(M_{\bar{a}}))} = -\text{AIE}_{a\bar{a}}$, which differs in setting A to a rather than to $\bar{a}$ along $A \rightarrow Y$. In the linear case, the two definitions coincide (see Eqs. (2) and (3)). Similarly the ADE can be defined as $\text{ADE}^a_{\bar{a}a} = \langle Y_a\rangle_{p(Y_a)} - \langle Y_{\bar{a}}(M_a)\rangle_{p(Y_{\bar{a}}(M_a))} = -\text{ADE}_{a\bar{a}}$.

where ϵ_c, ϵ_m, ϵ_l and ϵ_y are unobserved independent zero-mean Gaussian variables, we can compute $\langle Y_{\bar{a}} \rangle$ by expressing Y as a function of $A = \bar{a}$ and the Gaussian variables, by recursive substitutions in C, M and L, *i.e.*

$$\begin{aligned} Y_{\bar{a}} &= \theta^y + \theta^y_a \bar{a} + \theta^y_c \epsilon_c + \theta^y_m(\theta^m + \theta^m_a \bar{a} + \theta^m_c \epsilon_c + \epsilon_m) \\ &\quad + \theta^y_l(\theta^l + \theta^l_a \bar{a} + \theta^l_c \epsilon_c + \theta^l_m(\theta^m + \theta^m_a \bar{a} + \theta^m_c \epsilon_c + \epsilon_m) + \epsilon_l) + \epsilon_y, \end{aligned}$$

and then take the mean, obtaining $\langle Y_{\bar{a}} \rangle = \theta^y + \theta^y_a \bar{a} + \theta^y_m(\theta^m + \theta^m_a \bar{a}) + \theta^y_l(\theta^l + \theta^l_a \bar{a} + \theta^l_m(\theta^m + \theta^m_a \bar{a}))$. Analogously

$$\langle Y_a(M_a, L_{\bar{a}}(M_a)) \rangle = \theta^y + \theta^y_a a + \theta^y_m(\theta^m + \theta^m_a a) + \theta^y_l(\theta^l + \theta^l_a \bar{a} + \theta^l_m(\theta^m + \theta^m_a a)).$$

For $a = 1$ and $\bar{a} = 0$, this gives

$$\text{PSE}_{\bar{a}a} = \theta^y_a(a - \bar{a}) + \theta^y_m \theta^m_a (a - \bar{a}) + \theta^y_l \theta^l_m \theta^m_a (a - \bar{a}) = \theta^y_a + \theta^y_m \theta^m_a + \theta^y_l \theta^l_m \theta^m_a.$$

The same conclusion could have been obtained by looking at the graph annotated with path coefficients (Fig. 5 (bottom)). The PSE is obtained by summing over the three causal paths of interest ($A \rightarrow Y$, $A \rightarrow M \rightarrow Y$, and $A \rightarrow M \rightarrow L \rightarrow Y$) the product of all coefficients in each path.

Notice that $\text{AIE}_{\bar{a}a}$, given by

$$\begin{aligned} \text{AIE}_{\bar{a}a} &= \langle Y_{\bar{a}}(M_a, L_a(M_a)) \rangle - \langle Y_{\bar{a}} \rangle \\ &= \theta^y + \theta^y_a \bar{a} + \theta^y_m(\theta^m + \theta^m_a a) + \theta^y_l(\theta^l + \theta^l_a a + \theta^l_m(\theta^m + \theta^m_a a)) \\ &\quad - \theta^y + \theta^y_a \bar{a} + \theta^y_m(\theta^m + \theta^m_a \bar{a}) + \theta^y_l(\theta^l + \theta^l_a \bar{a} + \theta^l_m(\theta^m + \theta^m_a \bar{a})) \\ &= \theta^y_m \theta^m_a (a - \bar{a}) + \theta^y_l(\theta^l_a(a - \bar{a}) + \theta^l_m \theta^m_a (a - \bar{a})), \end{aligned} \tag{2}$$

coincides with $\text{AIE}^a_{\bar{a}a}$, given by

$$\begin{aligned} \text{AIE}^a_{\bar{a}a} &= \langle Y_a \rangle - \langle Y_a(M_{\bar{a}}, L_{\bar{a}}(M_{\bar{a}})) \rangle \\ &= \theta^y + \theta^y_a a + \theta^y_m(\theta^m + \theta^m_a a) + \theta^y_l(\theta^l + \theta^l_a a + \theta^l_m(\theta^m + \theta^m_a a)) \\ &\quad - \theta^y + \theta^y_a a + \theta^y_m(\theta^m + \theta^m_a \bar{a}) + \theta^y_l(\theta^l + \theta^l_a \bar{a} + \theta^l_m(\theta^m + \theta^m_a \bar{a})). \end{aligned} \tag{3}$$

Effect of Treatment on Treated. Consider the conditional distribution $p(Y_a | A = \bar{a})$. This distribution measures the information travelling from A to Y along all open paths, when A is set to a along causal paths and to $\bar{a}$ along non-causal paths. The *effect of treatment on treated* (ETT) of $A = a$ with respect to $A = \bar{a}$ is defined as $\text{ETT}_{\bar{a}a} = \langle Y_a \rangle_{p(Y_a|A=\bar{a})} - \langle Y_{\bar{a}} \rangle_{p(Y_{\bar{a}}|A=\bar{a})} = \langle Y_a \rangle_{p(Y_a|A=\bar{a})} - \langle Y \rangle_{p(Y|A=\bar{a})}$. As the PSE, the ETT measures difference in flow of information from A to Y when A takes different values along different paths. However, the PSE considers only causal paths and different values for A along different causal paths, whilst the ETT considers all open paths and different values for A along causal and non-causal paths respectively. Similarly to $\text{ATE}_{\bar{a}a}$, $\text{ETT}_{\bar{a}a}$ for the CBN of Fig. 4(b) can be expressed as

$$\text{ETT}_{\bar{a}a} = \underbrace{\langle Y_a(M_{\bar{a}}) \rangle - \langle Y_{\bar{a}} \rangle}_{\text{ADE}_{\bar{a}a|\bar{a}}} - (\underbrace{\langle Y_a(M_{\bar{a}}) \rangle - \langle Y_a \rangle}_{\text{AIE}_{a\bar{a}|\bar{a}}}).$$

Notice that, if we define difference in flow of non-causal (along the open back-door paths) information from A to Y when $A = a$ with respect to when $A = \bar{a}$ as $\text{NCI}_{\bar{a}a} = \langle Y_{\bar{a}} \rangle_{p(Y_{\bar{a}}|A=a)} - \langle Y \rangle_{p(Y|A=\bar{a})}$, we obtain

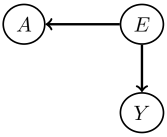
$$\begin{aligned}\langle Y \rangle_{p(Y|A=a)} - \langle Y \rangle_{p(Y|A=\bar{a})} &= \langle Y_{\bar{a}} \rangle_{p(Y_{\bar{a}}|A=a)} - \langle Y \rangle_{p(Y|A=\bar{a})} \\ &\quad - \left(\langle Y_{\bar{a}} \rangle_{p(Y_{\bar{a}}|A=a)} - \langle Y \rangle_{p(Y|A=a)}\right) \\ &= \text{NCI}_{\bar{a}a} - \text{ETT}_{a\bar{a}} = \text{NCI}_{\bar{a}a} - \text{ADE}_{a\bar{a}|a} + \text{AIE}_{\bar{a}a|a}.\end{aligned}$$

4 Fairness Considerations Using CBNs

Equipped with the background on CBNs from Sect. 3, in this section we further investigate unfairness in a dataset $\varDelta = \{a^n, x^n, y^n\}_{n=1}^N$, discuss issues that might arise when building a decision system from it, and show how to measure and deal with unfairness in complex scenarios, revisiting and extending material from [11,32,47].

4.1 Back-Door Paths from A to Y

In Sect. 2 we have introduced a graphical interpretation of unfairness in a dataset $\varDelta$ as the presence of an unfair causal path from A to $\mathcal{X}$ or Y. More specifically, we have shown through a college admission example that unfairness can be due to an unfair link emerging (a) from A or (b) from a subsequent variable in a causal path from A to Y (*e.g.* $D \rightarrow Y$ in the example). Our discussion did not mention paths from A to Y with an arrow pointing into A, namely back-door paths. This is because such paths are not problematic.



To understand this, consider the hiring scenario described by the CBN on the left, where A represents religious belief and E educational background of the applicant, which influences religious participation ($E \rightarrow A$). Whilst $Y \not\perp\!\!\!\perp A$ due to the open back-door path from A to Y, the hiring decision Y is only based on E.

4.2 Opening Closed Unfair Paths from A to Y

In Sect. 2, we have seen that, in order to reason about fairness of $\hat{Y}$, it is necessary to question and understand unfairness in $\varDelta$. In this section, we warn that another crucial element needs to be considered in the fairness discussion around $\hat{Y}$, namely

(i) The subset of variables used to form $\hat{Y}$ could project into $\hat{Y}$ unfair patterns in $\mathcal{X}$ that do not concern Y.

This could happen, for example, if a closed unfair path from A to Y is opened when conditioning on the variables used to form $\hat{Y}$.

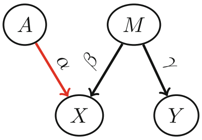


Fig. 6. CBN underlying a music degree scenario.

As an example, assume the CBN in Fig. 6 representing the data-generation mechanism underlying a music degree scenario, where A corresponds to gender, M to music aptitude (unobserved, *i.e.* $M \notin \Delta$), X to the score obtained from an ability test taken at the beginning of the degree, and Y to the score obtained from an ability test taken at the end of the degree. Individuals with higher music aptitude M are more likely to obtain higher initial and final scores ($M \rightarrow X$, $M \rightarrow Y$). Due to discrimination occurring at the initial testing, women are assigned a lower initial score than men for the same aptitude level ($A \rightarrow X$). The only path from A to Y, $A \rightarrow X \leftarrow M \rightarrow Y$, is closed as X is a collider on this path. Therefore the unfair influence of A on X does not reach Y ($Y \perp\!\!\!\perp A$). Nevertheless, as $Y \not\!\perp\!\!\!\perp A|X$, a prediction $\hat{Y}$ based on the initial score X only would contain the unfair influence of A on X. For example, assume the following linear model: $Y = \gamma M, \quad X = \alpha A + \beta M$, with $\langle A^2 \rangle_{p(A)} = 1$ and $\langle M^2 \rangle_{p(M)} = 1$. A linear predictor of the form $\hat{Y} = \theta_X X$ minimizing $\langle (Y - \hat{Y})^2 \rangle_{p(A)p(M)}$ would have parameters $\theta_X = \gamma\beta/(\alpha^2 + \beta^2)$, giving $\hat{Y} = \gamma\beta(\alpha A + \beta M)/(\alpha^2 + \beta^2)$, *i.e.* $\hat{Y} \not\!\perp\!\!\!\perp A$. Therefore, this predictor would be using the sensitive attribute to form a decision, although implicitly rather than explicitly. Instead, a predictor explicitly using the sensitive attribute, $\hat{Y} = \theta_X X + \theta_A A$, would have parameters

$$\begin{pmatrix} \theta_X \\ \theta_A \end{pmatrix} = \begin{pmatrix} \alpha^2 + \beta^2 & \alpha \\ \alpha & 1 \end{pmatrix}^{-1} \begin{pmatrix} \gamma\beta \\ 0 \end{pmatrix} = \begin{pmatrix} \gamma/\beta \\ -\alpha\gamma/\beta \end{pmatrix},$$

i.e. $\hat{Y} = \gamma M$. Therefore, this predictor would be fair. From the CBN we can see that the explicit use of A can be of help in retrieving M. Indeed, since $M \not\!\perp\!\!\!\perp A|X$, using A in addition to X can give information about M. In general (*e.g.* in a non-linear setting) it is not guaranteed that using A would ensure $\hat{Y} \perp\!\!\!\perp A$. Nevertheless, this example shows how explicit use of the sensitive attribute in a model can ensure fairness rather than lead to unfairness.

This observation is relevant to one of the simplest fairness definitions, motivated by legal requirements, called *fairness through unawareness*, which states that $\hat{Y}$ is fair as long as it does not make explicit use of the sensitive attribute A. Whilst this fairness criterion is often indicated as problematic because some of the variables used to form $\hat{Y}$ could be a proxy for A (such as neighborhood for race), the example above shows a more subtle issue with it.

4.3 Path-Specific Population-Level Unfairness

In this section, we show that the path-specific effect introduced in Sect. 3 can be used to quantify unfairness in Δ in complex scenarios.

Consider the college admission example discussed in Sect. 2 (Fig. 7). In the case in which the path $A \rightarrow D$, and therefore $A \rightarrow D \rightarrow Y$, is considered unfair, unfairness overall population can be quantified with $\langle Y \rangle_{p(Y|a)} - \langle Y \rangle_{p(Y|\bar{a})}$

(coinciding with $\text{ATE}_{\bar{a}a} = \langle Y_a \rangle_{p(Y_a)} - \langle Y_{\bar{a}} \rangle_{p(Y_{\bar{a}})}$) where, for example, $A = a$ and $A = \bar{a}$ indicate female and male applicants respectively.

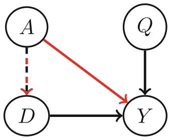


Fig. 7. CBN underlying a college admission scenario.

In the more complex case in which the path $A \rightarrow D \rightarrow Y$ is considered fair, unfairness can instead be quantified with the path-specific effect along the direct path $A \rightarrow Y$, $\text{PSE}_{\bar{a}a}$, given by

$$\langle Y_a(D_{\bar{a}}) \rangle_{p(Y_a(D_{\bar{a}}))} - \langle Y_{\bar{a}} \rangle_{p(Y_{\bar{a}})}.$$

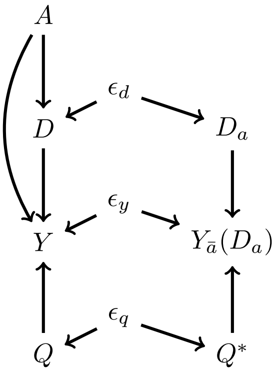
Notice that computing $p(Y_a(D_{\bar{a}}))$ requires knowledge of the CBN. If the CBN structure is not known or estimating its conditional distributions is challenging, the resulting estimate could be imprecise.

4.4 Path-Specific Individual-Level Unfairness

In the college admission example of Fig. 7 in which the path $A \rightarrow D \rightarrow Y$ is considered fair, rather than measuring unfairness overall population, we might want to know *e.g.* whether a rejected female applicant $\{a^n = a = 1, q^n, d^n, y^n = 0\}$ was treated unfairly. We can answer this question by estimating whether the applicant would have been admitted had she been male ($A = \bar{a} = 0$) along the direct path $A \rightarrow Y$ from $p(Y_{\bar{a}}(D_a)|A = a, Q = q^n, D = d^n)$ (notice that the outcome in the actual world, y^n, corresponds to $p(Y_a(D_a)|A = a, Q = q^n, D = d^n) = \mathbb{1}_{Y_a(D_a)=y^n}$).

To understand how this can be achieved, consider the following linear model associated to a CBN with the same structure as the one in Fig. 7

$$A \sim \text{Bern}(\pi), Q = \theta^q + \epsilon_q, D = \theta^d + \theta^d_a A + \epsilon_d, Y = \theta^y + \theta^y_a A + \theta^y_q Q + \theta^y_d D + \epsilon_y.$$



The relationships between A, Q, D, Y and $Y_{\bar{a}}(D_a)$ in this model can be inferred from the *twin Bayesian network* [38] on the left resulting from the intervention $A = a$ along $A \rightarrow D$ and $A = \bar{a}$ along $A \rightarrow Y$: in addition to A, Q, D, Y, the network contains the variables Q^*, D_a and $Y_{\bar{a}}(D_a)$ corresponding to the counterfactual world in which $A = \bar{a}$ along $A \rightarrow Y$. The two groups of variables are connected through $\epsilon_d, \epsilon_q, \epsilon_y$, indicating that the factual and counterfactual worlds share the same unobserved randomness. From this network, we can deduce that $Y_{\bar{a}}(D_a) \perp\!\!\!\perp \{A, Q, D\}|\{\epsilon_q, \epsilon_d\}$[5], and therefore that we can express $p(Y_{\bar{a}}(D_a)|A = a, Q = q^n, D = d^n)$ as

[5] Notice that $Y_{\bar{a}}(D_a) \perp\!\!\!\perp A$, but $Y_{\bar{a}}(D_a) \not\perp\!\!\!\perp A|D$.

$$p(Y_{\bar{a}}(D_a)|a, q^n, d^n) = \int_{\epsilon_q, \epsilon_d} p(Y_{\bar{a}}(D_a)|\epsilon_q, \epsilon_d, \cancel{a}, \cancel{q^n}, \cancel{d^n}) p(\epsilon_q, \epsilon_d|a, q^n, d^n). \tag{4}$$

As $\epsilon_q^n = q^n - \theta^q$, $\epsilon_d^n = d^n - \theta^d - \theta_a^d$, we obtain[6] $\langle Y_{\bar{a}}(D_a)\rangle_{p(Y_{\bar{a}}(D_a)|A=a,Q=q^n,D=d^n)} = \theta^y + \theta_q^y q^n + \theta_d^y d^n$.

Equation (4) suggests that, in more complex scenarios (*e.g.* in which the variables are non-linearly related), we can obtain a Monte-Carlo estimate of $p(Y_{\bar{a}}(D_a)|a, q^n, d^n)$ by sampling ϵ_q and ϵ_d from $p(\epsilon_q, \epsilon_d|a, q^n, d^n)$.

In [11], we used this approach to introduce a prediction system such that the two distributions $p(\hat{Y}_{\bar{a}}(D_a)|A = a, Q = q^n, D = d^n)$ and $p(\hat{Y}_a(D_a)|A = a, Q = q^n, D = d^n)$ coincide – we called this property *path-specific counterfactual fairness*.

5 Conclusions

We used causal Bayesian networks to provide a graphical interpretation of unfairness in a dataset as the presence of an unfair causal effect of a sensitive attribute. We used this viewpoint to revisit the recent debate surrounding the COMPAS pretrial risk assessment tool and, more generally, to point out that fairness evaluation on a model requires careful considerations on the patterns of unfairness underlying the training data. We then showed that causal Bayesian networks provide us with a powerful tool to measure unfairness in a dataset and to design fair models in complex unfairness scenarios.

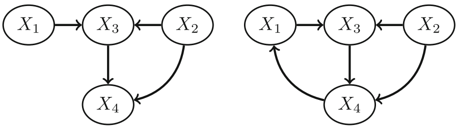
Our discussion did not cover difficulties in making reasonable assumptions on the structure of the causal Bayesian network underlying a dataset, nor on the estimations of the associated conditional distributions or of other quantities of interest. These are obstacles that need to be carefully considered to avoid improper usage of this framework.

Acknowledgements. The authors would like to thank Ray Jiang, Christina Heinze-Deml, Tom Stepleton, Tom Everitt, and Shira Mitchell for useful discussions.

Appendix A Bayesian Networks

A *graph* is a collection of nodes and links connecting pairs of nodes. The links may be directed or undirected, giving rise to *directed* or *undirected graphs* respectively. A *path* from node X_i to node X_j is a sequence of linked nodes starting at X_i and ending at X_j. A *directed path* is a path whose links are directed and pointing from preceding towards following nodes in the sequence.

[6] Notice that $\langle Y_{\bar{a}}(D_a)\rangle_{p(Y_{\bar{a}}(D_a)|A=a,Q=q^n,D=d^n)} = \langle Y\rangle_{p(Y|A=a,Q=q^n,D=d^n)} - \text{PSE}_{\bar{a}a}^{\cancel{d}}$. Indeed $\langle Y\rangle_{p(Y|A=a,Q=q^n,D=d^n)} = \theta^y + \theta_a^y + \theta_q^y q^n + \theta_d^y d^n$ and $\text{PSE}_{\bar{a}a} = \theta_a^y$. This equivalence does not hold in the non-linear setting.



(a) (b)

Fig. 8. Directed (a) acyclic and (b) cyclic graph.

A *directed acyclic graph* (DAG) is a directed graph with no directed paths starting and ending at the same node. For example, the directed graph in Fig. 8(a) is acyclic. The addition of a link from X_4 to X_1 makes the graph cyclic (Fig. 8(b)). A node X_i with a directed link to X_j is called *parent* of X_j. In this case, X_j is called *child* of X_i.

A node is a *collider* on a path if it has (at least) two parents on that path. Notice that a node can be a collider on a path and a non-collider on another path. For example, in Fig. 8(a) X_3 is a collider on the path $X_1 \rightarrow X_3 \leftarrow X_2$ and a non-collider on the path $X_2 \rightarrow X_3 \rightarrow X_4$.

A node X_i is an *ancestor* of a node X_j if there exists a directed path from X_i to X_j. In this case, X_j is a *descendant* of X_i.

A *Bayesian network* is a DAG in which nodes represent random variables and links express statistical relationships between the variables. Each node X_i in the graph is associated with the conditional distribution $p(X_i|\text{pa}(X_i))$, where $\text{pa}(X_i)$ is the set of parents of X_i. The joint distribution of all nodes, $p(X_1,\ldots,X_I)$, is given by the product of all conditional distributions, *i.e.* $p(X_1,\ldots,X_I) = \prod_{i=1}^{I} p(X_i|\text{pa}(X_i))$.

In a Bayesian network, the sets of variables $\mathcal{X}$ and $\mathcal{Y}$ are statistically independent given $\mathcal{Z}$ ($\mathcal{X} \perp\!\!\!\perp \mathcal{Y} \,|\, \mathcal{Z}$) if all paths from any element of $\mathcal{X}$ to any element of $\mathcal{Y}$ are *closed* (or *blocked*). A path is closed if at least one of the following conditions is satisfied:

(a) There is a non-collider on the path which belongs to the conditioning set $\mathcal{Z}$.
(b) There is a collider on the path such that neither the collider nor any of its descendants belong to the conditioning set $\mathcal{Z}$.

Appendix B EFPRs/EFNRs and Calibration

Assume that EFPRs/EFNRs are satisfied, *i.e.* $p(\hat{Y}=1|A=0,Y=1) = p(\hat{Y}=1|A=1,Y=1) \equiv p_{\hat{Y}_1|Y_1}$ and $p(\hat{Y}=1|A=0,Y=0) = p(\hat{Y}=1|A=1,Y=0) \equiv p_{\hat{Y}_1|Y_0}$. From

$$p(Y=1|A=0,\hat{Y}=1) = \frac{p_{\hat{Y}_1|Y_1}\overbrace{p(Y=1|A=0)}^{p_{Y_1|A_0}}}{p_{\hat{Y}_1|Y_1}p_{Y_1|A_0} + p_{\hat{Y}_1|Y_0}(1-p_{Y_1|A_0})},$$

$$p(Y=1|A=1,\hat{Y}=1) = \frac{p_{\hat{Y}_1|Y_1}p_{Y_1|A_1}}{p_{\hat{Y}_1|Y_1}p_{Y_1|A_1} + p_{\hat{Y}_1|Y_0}(1-p_{Y_1|A_1})},$$

we see that, to also satisfy $p(Y=1|A=0,\hat{Y}=1) = p(Y=1|A=1,\hat{Y}=1)$, we need $\big(\cancel{p_{\hat{Y}_1|Y_1}p_{Y_1|A_1}} + \cancel{p_{\hat{Y}_1|Y_0}}(1-\cancel{p_{Y_1|A_1}})\big)p_{Y_1|A_0} = \big(\cancel{p_{\hat{Y}_1|Y_1}p_{Y_1|A_0}} + \cancel{p_{\hat{Y}_1|Y_0}}(1-\cancel{p_{Y_1|A_0}})\big)p_{Y_1|A_1}$, *i.e.* $p_{Y_1|A_0} = p_{Y_1|A_1}$.

References

1. AI Now Institute. Litigating Algorithms: Challenging Government Use of Algorithmic Decision Systems (2018)
2. Alexander, M.: The New Jim Crow: Mass Incarceration in the Age of Colorblindness. The New Press, New York (2012)
3. Andrews, D.A., Bonta, J.: Level of Service Inventory - Revised. Multi-Health Systems, Toronto (2000)
4. Andrews, D.A., Bonta, J., Wormith, J.S.: The recent past and near future of risk and/or need assessment. Crime Delinq. **52**(1), 7–27 (2006)
5. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine Bias: there's software used across the country to predict future criminals. And it's biased against blacks, May 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
6. Arnold, D., Dobbie, W., Yang, C.S.: Racial bias in bail decisions. Q. J. Econ. **133**, 1885–1932 (2018)
7. Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A.: Fairness in criminal justice AQ3 risk assessments: the state of the art. Sociol. Methods Res. (2018)
8. Bogen, M., Rieke, A.: Help wanted: an examination of hiring algorithms, equity, and bias. Technical report, Upturn (2018)
9. Brennan, T., Dieterich, W., Ehret, B.: Evaluating the predictive validity of the COMPAS risk and needs assessment system. Crim. Justice Behav. **36**(1), 21–40 (2009)
10. Byanjankar, A., Heikkilä, M., Mezei, J.: Predicting credit risk in peer-to-peer lending: a neural network approach. In: IEEE Symposium Series on Computational Intelligence, pp. 719–725 (2015)
11. Chiappa, S.: Path-specific counterfactual fairness. In: Thirty-Third AAAI Conference on Artificial Intelligence (2019)
12. Chouldechova, A.: Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. Big Data **5**(2), 153–163 (2017)
13. Chouldechova, A., Putnam-Hornstein, E., Benavides-Prado, D., Fialko, O., Vaithianathan, R.: A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. Proc. Mach. Learn. Res. **81**, 134–148 (2018)
14. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A.: A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear, October 2016. https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/?utm_term=.8c6e8c1cfbdf
15. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A.: Algorithmic decision making and the cost of fairness. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 797–806 (2017)
16. Corbett-Davies, S., Goel, S.: The measure and mismeasure of fairness: a critical review of fair machine learning. CoRR, abs/1808.00023 (2018)
17. Dawid, P.: Fundamentals of statistical causality. Technical report, University College London (2007)
18. De Fauw, J., et al.: Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat. Med. **24**(9), 1342–1350 (2018)
19. Dieterich, W., Mendoza, C., Brennan, T.: COMPAS risk scales: demonstrating accuracy equity and predictive parity (2016)

20. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, pp. 214–226 (2012)
21. Eckhouse, L., Lum, K., Conti-Cook, C., Ciccolini, J.: Layers of bias: a unified approach for understanding problems with risk assessment. Crim. Justice Behav. **46**, 185–209 (2018)
22. Eubanks, V.: Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. St. Martin's Press, New York (2018)
23. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 259–268 (2015)
24. Flores, A.W., Bechtel, K., Lowenkamp, C.T.: False positives, false negatives, and false analyses: a rejoinder to "Machine Bias: there's software used across the country to predict future criminals. And it's biased against blacks". Fed. Probat. **80**(2), 38–46 (2016)
25. Harvard Law School. Note: Bail reform and risk assessment: The cautionary tale of federal sentencing. Harvard Law Rev. **131**(4), 1125–1146 (2018)
26. He, X., et al.: Practical lessons from predicting clicks on ads at Facebook. In: Proceedings of the Eighth International Workshop on Data Mining for Online Advertising, pp. 1–9 (2014)
27. Hoffman, M., Kahn, L.B., Li, D.: Discretion in hiring. Q. J. Econ. **133**(2), 765–800 (2018)
28. Isaac, W.S.: Hope, hype, and fear: the promise and potential pitfalls of artificial intelligence in criminal justice. Ohio State J. Crim. Law **15**(2), 543–558 (2017)
29. Kleinberg, J., Mullainathan, S., Raghavan, M.: Inherent trade-offs in the fair determination of risk scores. In: 8th Innovations in Theoretical Computer Science Conference, pp. 43:1–43:23 (2016)
30. Koepke, J.L., Robinson, D.G.: Danger ahead: risk assessment and the future of bail reform. Wash. Law Rev. **93**, 1725–1807 (2017)
31. Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V., Fotiadis, D.I.: Machine learning applications in cancer prognosis and prediction. Comput. Struct. Biotechnol. J. **13**, 8–17 (2015)
32. Kusner, M.J., Loftus, J.R., Russell, C., Silva, R.: Counterfactual fairness. In: Advances in Neural Information Processing Systems 30, pp. 4069–4079 (2017)
33. Lum, K.: Limitations of mitigating judicial bias with machine learning. Nat. Hum. Behav. **1**(7), 1 (2017)
34. Lum, K., Isaac, W.: To predict and serve? Significance **13**(5), 14–19 (2016)
35. Malekipirbazari, M., Aksakalli, V.: Risk assessment in social lending via random forests. Expert Syst. Appl. **42**(10), 4621–4631 (2015)
36. Mayson, S.G.: Bias in, bias out. Yale Law Sch. J. **128** (2019)
37. Mitchell, S., Potash, E., Barocas, S.: Prediction-based decisions and fairness: a catalogue of choices, assumptions, and definitions (2018)
38. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge (2000)
39. Pearl, J., Glymour, M., Jewell, N.P.: Causal Inference in Statistics: A Primer. Wiley, Hoboken (2016)
40. Perlich, C., Dalessandro, B., Raeder, T., Stitelman, O., Provost, F.: Machine learning for targeted display advertising: transfer learning in action. Mach. Learn. **95**(1), 103–127 (2014)

41. Peters, J., Janzing, D., Schölkopf, B.: Elements of Causal Inference: Foundations and Learning Algorithms. MIT Press, Cambridge (2017)
42. Rosenberg, M., Levinson, R.: Trump's catch-and-detain policy snares many who call the U.S. home, June 2018. https://www.reuters.com/investigates/special-report/usa-immigration-court
43. Selbst, A.D.: Disparate impact in big data policing. Georgia Law Rev. **52**, 109–195 (2017)
44. Spirtes, P., et al.: Causation, Prediction, and Search. MIT Press, Cambridge (2000)
45. Stevenson, M.T.: Assessing risk assessment in action. Minnesota Law Rev. **103**, 303 (2017)
46. Vaithianathan, R., Maloney, T., Putnam-Hornstein, E., Jiang, N.: Children in the public benefit system at risk of maltreatment: identification via predictive modeling. Am. J. Prev. Med. **45**(3), 354–359 (2013)
47. Zhang, J., Bareinboim, E.: Fairness in decision-making - the causal explanation formula. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence (2018)

Sharing Is Caring, a Boundary Object Approach to Mapping and Discussing Personal Data Processing

Rob Heyman(✉)

imec-SMIT, Vrije Universiteit Brussel, Pleinlaan 9, 1050 Brussels, Belgium
rob.heyman@vub.be

Abstract. This work answers the following question, how to gather and act on personal data in smart city projects using boundary objects? Smart city projects require new mapping methods so they can share and discuss work collectively. Working collectively is necessary because smart city projects are difficult to map in one singular view for personal data because different smart city stakeholders have a part of the required information. Summarising data processing operations is most often taken for granted and under-defined in Data Protection Impact Assessment methods.

This paper is a plea for the use of boundary objects for GDPR compliance and research in smart cities. Therefore, this article is a comparison of the original context boundary objects with the context of smart cities to illustrate the need for a similar approach. The main results of this paper point to a new approach to enable collaborative GDPR compliance where specialist knowledge trickles down to developers and other actors not educated to comply with GDPR requirements.

Keywords: GDPR · Boundary object · Smart city

1 Introduction

This research paper is an answer to a pragmatic question: how to gather and act on personal data in smart city projects using boundary objects? In smart city projects, there is a need to map data processing operations due to risks to privacy and the threat of GDPR[1] fines. Collecting this information is challenging because people with this task rarely have access to the information or have the expertise to understand all technicalities.

Also, the nature of smart city projects brings its own challenges. Smart city projects comprise new innovations where multiple departments or organizations have to work together on a new technology from concept, prototype to launch. This means that:

[1] The General Data Protection Regulation replaced the Data Protection Directive in May 2018. An important change in this reform is the principle of accountability. It requires data controllers to be able to demonstrate compliance with GDPR principles. For this requirement smart cities require full awareness about their data processing activities and this is perceived as new by many data controllers.

Published by Springer Nature Switzerland AG 2019
E. Kosta et al. (Eds.): Privacy and Identity 2018, IFIP AICT 547, pp. 21–31, 2019.
https://doi.org/10.1007/978-3-030-16744-8_2

different smart city stakeholders hold a part of the required information, the project changes over time which results in personal data flows that are vague and underdefined. Because of this complexity, we could say smart city projects exist at the boundaries between different social worlds[2] where GDPR compliance adds extra complexity beyond the complexity of getting a working proof of concept.

Problem Statement: How to improve the lack of information about personal data processing in smart city projects through boundary objects? This essay compares boundary objects as an approach that may improve personal data processing documentation required for further steps in GDPR compliance. To answer this question, we look at factors that render this a difficult undertaking and compare these to the data collection challenges that prompted the use of boundary objects as a concept and tool.

Relevance: From a societal perspective this is important as more and more cities invest in smart city projects, which means they monitor public space more than before. To prevent a trade-off between more smart city benefits versus privacy, a complete mapping of the data collection process may help in creating the circumstances to have smart city benefits *and* privacy. From a scientific perspective, there is little research on how to collect and distribute information about an innovation in development and discuss it collectively. The literature is clear on the requirements to mitigate privacy challenges during development but very unclear on how to implement these in the development process. Smart cities, as an object of study, are interesting because it is still an underdefined concept that will most likely crystallize soon. It means that many actors have different views on what a smart city should be. This means two things. That boundary objects are very helpful to bridge different viewpoints and that stakeholders can work together via boundary objects to come to an acceptable definition of smart cities for all.

We believe GDPR compliance can only achieve Privacy-by-design if an organization distributes legal requirements on the ground. It will not work if GDPR compliance remains the isolated responsibility of one person. To break GDPR compliance out of its isolation, we need to cross disciplinary boundaries. Last, improving transparency tools like boundary objects for the GDPR may also increase relevance of this regulation as more stakeholders become empowered to take part in smart cities.

Results: We found that ensuring GDPR compliance in smart city projects is a context with a high degree of interpretive flexibility; things have different meanings to different people. This is challenging because there is a collective goal of delivering a working proof of concept which is GDPR compliant. Just like museums for vertebrate zoology[3], there is a need to create a shared language to talk about specific information between different backgrounds and disciplines.

I structure this paper as follows. We first summarize the history of boundary objects, their main components and their theoretical background. We then continue by situating smart cities, smart city projects and the challenges for GDPR compliance.

[2] Cfr. infra.

[3] The case Star and Griesemer used to develop the concept of boundary objects.

Next, we compare how the problems solved with boundary objects in museums are like those of smart city projects. We then conclude with advantages of using boundary objects in smart city projects.

2 Boundary Objects History and Architecture

For this part we aim to provide the context in which Star and Griesemer developed boundary objects, the perspectives they are part of and the architecture or basic building blocks they comprise. We require this to compare this context to that of smart city projects.

Boundary objects as a concept resulted from Star and Griesemer's research into cooperative work in the absence of consensus [1]. This occurs if different social worlds have to arrive at a coherent (research) result. For example, during the developments of Berkeley's Museum of Vertebrate Zoology, 1907–39 [2]. The cause and context for this research was the authors' search for situations in need of cooperative work where little to no consensus exists.

Star and Griesemer describe the beginning of the 20th century in West America as a situation where zoologic museums were evolving from amateur collections to academic curated collections. Here, amateur enthusiasts curated museums for fellow enthusiasts. In Stars case, these amateurs were interested in Californian ornithology. What changed, was biological science's need for large information collections. This meant that amateurs from different hobbies increased their interaction with curators and scientists to create scientific knowledge together.

Grinnel, an ecology biologist and director of Berkeley's Museum of Vertebrate Zoology is instrumental to facilitate this change from amateur museum to a research center: "Joseph Grinnell was the first director of the Museum of Vertebrate Zoology. He worked on problems of speciation, migration and the role of the environment in Darwinian evolution. Grinnell's research required the labours of (among others) university administrators, professors, research scientists, curators, amateur collectors, private sponsors and patrons, occasional field hands, government officials and members of scientific clubs" [2].

Grinnell changed the practices of amateur bird collectors and museum curators to fit his needs as a biologist. His aim was to map how environment plays a role in Darwin's evolution theory to support Darwin's work with proof. To do so, Grinnell collected bird sightings, bird cadavers and references to where these were found. This collection needed to be so large Grinnell could not do it alone.

The group of amateurs already gathered information and birds to contribute to the museum but in a non-standardised, incoherent way. Thus the practices of this group did not fit the rigorous reporting Grinnell required. To use their information for scientific progress, Grinnel had to overcome two barriers. First, the work Grinnell expected them to do had to fit in their current activities. Second, he had to discipline amateurs to understand what he expected from them without overburdening their free time.

Grinnell taught his different information providers how to do fieldwork with the bare minimum of required biological knowledge and methods to standardise this. Next, he encouraged the use of boundary objects; specifically structured field notes containing information fields of interest to all involved parties but still rigorous enough for scientists to use as input for their research.

For example, when amateur and professional trappers capture animals, they needed to understand what parts of the animal had to remain intact, how to preserve them and last, what information trappers needed to add about the environment they captured it in. Grinnell facilitated the first two through teaching and the latter with field notes.

In Star's words, the information and the actors that contributed needed discipline. As we will see, this challenge is very similar to our smart city environment where non-experts collect information and where this information needs discipline as well. This brings us to the following section on how to carry this out.

2.1 Boundary Object Architecture

Boundary objects and standardization facilitate information sharing where actors try to perform cooperative work in the absence of consensus. Interpretive flexibility of a subject causes the latter. Different people see different meanings in the same phenomenon. This poses challenges for collective work because the chances to align everyone spontaneously are non-existent. Therefore, mutual understanding increases through standardisation and boundary objects.

Standardization takes place before the use of boundary objects, because it is work that brings different social worlds together: "By emphasizing how, and not what or why, methods standardization both makes information compatible and allows for a longer 'reach' across divergent worlds" [2].

A boundary object is a material or organisational structure to allow groups without consensus to work together. In that regard the object functions as an information carrier to facilitate work between groups that wish to cooperate despite their lack of consensus. It comprises two elements: object and boundary [1]. Object refers to something people act toward and with while boundary refers to a line between two worlds or social groups.

The object requires the following [1]:

- It has to exist between two or more social worlds
- It is vague for common purposes but particular social groups can specify information further because it remains open enough
- If no consensus exists, it can be a back and forth object until they reach consensus

For example, index cards to classify collections of bird samples. These exist between the person who brought in the bird cadaver, the curator and the scientist. It satisfies the first condition. It is open because it contains more information fields than any of the involved parties require. Each actor is interested in a subset of these information fields. Last, the history of the museum Star describes is one where Grinnell and others had to work on convincing others and part of that convincing comprised changing reporting methods to better fit in their free time.

This is where Star and Griesemer divert from Actor-Network theory on which they rely. There is a soft kind of steering that strives for coexistence instead of domination: "Grinnell's methods emphasis thus translated the concerns of his allies in such a way that their pleasure was not impaired - the basic activities of going on camping trips, adding to personal hobby collections and preserving California remained virtually untouched. With respect to the collectors, Grinnell created a mesh through which their products must pass if they want money or scientific recognition, but not so narrow a mesh that the products of their labour cannot be easily used" [2].

2.2 Origin of the Approach

In this section we situate Star and Griesemer's work within the respective disciplines they borrow from, symbolic interactionism and ANT. This is done to show the subtle difference between boundary objects and ANT. This difference will be used in the conclusion to refer to two styles of collaboration.

Symbolic interactionism "is part of a tradition of thought which comes from symbolic interactionism which seeks to qualify articulation mechanisms surrounding the perspectives of authors belonging to heterogeneous social worlds" [3]. Symbolic interactionism is "a micro-sociology tradition which rejects both sociological and biological determinism and privileges explanation on the basis of dynamic interactions which are observable between individuals. It underlines the fact that the sense of phenomena results from interpretations made by actors in context. These interpretations are to do with interpretive frameworks which move away from interactions between actors (verbal and non-verbal symbolic interactions)" [3].

Social worlds are defined as "activity groups which have neither clear boundaries nor formal and stable organisations. They develop through the relationship between social interactions which move away from the primary activity and the definition of pattern and reality. The notion derives from the symbolic interactionism tradition" [3].

ANT is adapted to fit with symbolic interactionism and the aim of the authors, to enable consensus between social worlds where there is none. In order to do that they render ANT more ecological. This means that they approach the interessement phase of Callon [4] in a manner where each actor is equally capable of defining translations. In order to understand this we need to define translations, interessement and obligatory passage points.

Translation [5] refers to the situation where one actor A succeeds in convincing another actor B to change their behaviour, usually in favour of the first actor (A). This is most often initiated by creating interessement [4]. This means that actor A needs to redefine his problem in such a way that it also becomes actor B's problem. In obligatory passage points (OPP) actor A has created a situation where B has to pass through because there is no other way and in so doing they change their behaviour to align with the prescribed behaviour A expects from B.

What the authors criticize ANT for, is that the latter's analysis aims at mapping obligatory points of passage. According to Star and Griesemer this means that networks are always reduced to the dominance of one actor.

In my opinion, Star and Griesemer reduce ANT to this top down thinking where one actor is dominant, and this makes sense since ANT is full of examples of OPPs that reinforce this observation. Nevertheless, the method ANT prescribes is one where each actor should be described in the same way without any a priori distinction [4, 6]. What is more, this approach leaves room for resistance in the concept of anti-scripts [7] and circumscription [8], two concepts that refer to resistance of actors to proposed translations.

This theoretical discussion set aside, what can be said is that Star and Griesemer are using ANT in a different context. The context is what they refer to as N-translation [2]. All involved actors are suggesting translations for other actors and because this is a context of collaboration, the question changes from domination to cooperation. How is it possible to have many translations while keeping consensus?

This means that the boundary object approach is more concerned about solutions than contestations. ANT is more suited to unearth hidden conflicts and contestations in dominant configuration. This approach focuses on the critical aspects of social constructivism. Boundary objects as an approach, use the fact that social constructivism exists to enable cooperation between actors; different understandings of the same phenomenon exist and that this can be overcome if all actors agree on standard definitions and approaches that can be used by all because they fit in their practices and social worlds.

3 Smart City Definition and Project Context

The aim of this part is to summarize what a smart city might mean and what this means for smart city projects that process personal data. Without reducing smart cities in one definition, we could say that this concept broadly refers to a development where multiple actors decide to work together on a smart city challenge. In terms of cooperation, the ideal refers to working with as much stakeholders as possible: "a smart city should focus on collaborating with diverse stakeholders, using technology as an enabler to achieve better and more efficient services to citizens" [9].

Walravens et al. pursued a more empirical definition in close collaboration with large Flemish cities: "A Smart City is a city in which all relevant city actors from the quadruple helix work together on more efficient and effective solutions, with the goal of tackling urban challenges. This collaboration is characterized by enabling innovative solutions that respect the local context and individuality of a city. Collecting, processing, sharing and opening data with relevant stakeholders contributes to concrete policy making and solutions. The local government can take up different roles, depending on the projects and which stakeholders or technological solutions are involved: local government can initiate, facilitate, direct, stimulate, regulate, experiment, test, validate, implement… The local government performs this function to serve and protect the public interest" [10].

In these definitions stakeholders are defined as members of the quadruple helix which "includes four types of actors (1) policy, (2) citizens, (3) research and (4) private partners" [10]. We can safely assume and testify in what follows that the quadruple helix consists of actors coming from different social worlds. But also that they wish to cooperate but lack consensus.

3.1 Smart City Project Context

It is impossible to talk about THE smart city, so instead we focus on concrete projects. That is the reason why we refer to the smart city project context. We define this context as an interdisciplinary and inter-organisational context which focuses on results first and GDPR compliance a bit later. Since we focus on GDPR compliance, we will follow with a section on privacy in smart city contexts after the more general description which follows now.

Smart city projects tackle urban challenges first and address GDPR challenges created later. Due to the newness and vagueness of smart cities as a concept, many projects are proofs of concept to illustrate possibilities. As a result, projects are evaluated on their performance first and secondly on their privacy protection. On a more practical level it means that practical solutions to implement a project are added first and that these are then evaluated with regard to the GDPR. This is pragmatic as the GDPR alone cannot steer innovation.

As previously mentioned authors point out the need to include the quadruple helix in smart city projects. These cooperations are believed to foster the exchange of ideas and technology [11]. This requirement to involve many different stakeholders results in a bigger variety of disciplinary backgrounds or social worlds. We define three challenges that render consensus difficult, a vertical challenge of working on different levels or social worlds. A horizontal challenge of working with different organisations. And lastly, the GDPR compliance challenge itself.

Project Vertical (Interdisciplinary Challenges)

With vertical I refer to the different layers in a single organization and the challenge of reaching consensus in one organization. In each project, different roles have to contribute to the successful implementation of a project: project management, development, legal, user testing, communication are roles that are often taken up by different people inside an organisation. The first challenge consists of aligning these different layers of an organisation. This is difficult because these layers are tied to different social worlds. For example, project managers may not have the same technical background their development team has.

Project Horizontal (Inter-organisational Challenges)

The quadruple helix includes different organisations and stakeholders' involvement in smart city projects. This adds an additional layer of factors that slow down the collection of relevant information and decision making based on this information. We can discern three factors: More time and effort is required to reach the right people that have information, the GDPR itself becomes an item of interpretation between different organisations and all of the vertical challenges apply to each organisation. For example, a city may propose a joint controllership but their partner may not believe that this is really necessary. The newness and vagueness of the text create discussions on what it really means for smart city projects.

Particular Compliance Context

In this article we limit ourselves to accountability: "as a principle which requires that organisations put in place appropriate technical and organisational measures and be able to demonstrate what they did and its effectiveness when requested" [12]. Here the EDPS refers to the following measures: "adequate documentation on what personal data are processed, how, to what purpose, how long; documented processes and procedures aiming at tackling data protection issues at an early state when building information systems or responding to a data breach; the presence of a Data Protection

Officer that be integrated in the organisation planning and operations etc." [12]. In this paper we focus on adequate documentation to ensure the ensuing steps of accountability.

This means that the GDPR poses a requirement from the legal layer to the other layers in one organization and between multiple organizations: be able to account for all the data processing on personal information. The challenge is very similar to that of the vertebrate zoology museum. Information has to be gathered in a disciplined manner and information itself has to be standardised to become useful. In practice this means that a person charged with documenting how data are processed has to do two things:

- Ask other parties to collect information about data processing operations
- Facilitate decision making about data processing operations

In sum, it could be said that knowledge is gathered without really knowing what needs to be collected. Decisions have to be made without really understanding what is at stake. This is a very similar situation to the vertebrate zoology museum. In what follows we explain why and how the Berkeley's Museum of Vertebrate Zoology is the same situation as our smart city projects.

4 Comparison of Situations

In this section we compare the context of the museum with that of smart city projects. What they have in common, is the need to include other parties to gather information, that all involved actors care for a common goal and that there is a need for standardization and boundary objects.

In both cases, the people who have access to the required information for good biological research or data protection, are not the people responsible for this goal. The people who have access to this information have access for different reasons. It means that they are from a different social world than biologists or GDPR managers. In the case of the smart city, these people are in the business of making sure their part of the smart city project works.

All involved stakeholders hold privacy dearly as it is important for the realization of their own goals. Trappers, amateurs, etc. capture and observe birds and care about understanding the environment of these birds. In both cases, there is an overlap but each actor may differ in terms of engagement or degree of importance. An engineer will not add the same priority to GDPR compliance as a compliance manager.

Those who care most about gathered information are the compliance manager and biologist but these have to rely on other actors to gather the necessary information to reach their common goal. Mapping bird ecologies or compliance with the GDPR. Because of this, those that gather information other than the specialists need to use standardized methods and boundary objects to attain this goal.

In conclusion we can say that both situations are highly similar. It could be argued that documenting data processing operations is easier compared to the museum context because the smart city context works with paid professionals rather than hobbyists[4]. GDPR compliance as a goal may also be more contested than the preservation and collection of wild life. GDPR can be motivated positively, for example, because we value privacy, or negatively, because we try not to get fined or receive negative attention.

4.1 Incoherent GDPR Compliance

This section adds examples of lack of consensus that can either be solved through standardization or boundary objects to further prove the point that this approach would add value to smart city projects. The lack of consensus may lead to a mismatch between legal concepts and other disciplinary concepts, the lacking existing boundary objects to gather information and the need for data visualizations that are understandable for all.

In many cases, the biggest challenge consists of reducing the mismatch between key legal concepts and the definition stemming from other social worlds. For example, knowing what personal data is, is the most important definition because it allows anyone involved in a smart city project to understand that a data collection or processing operation should be documented. A very common mistake consists of interpreting personal data as data the person who sees the data can use to identify an individual. This also means that data that do not allow someone to identify an individual is non-personal. As a result, IP-addresses but also infrared images are too quickly categorised as non-personal or anonymous data.

Another challenge is perhaps the need to fill in a data register. This is a repository to store all GDPR required information in an excel or database format. These registers use GDPR jargon making them difficult to use for non-GDPR users. What is more, a data register is a really difficult format to see the risks posed by the data because it is too abstract to identify possible risks.

Lastly, a lot of time is lost in discussions where legal concepts have to be applied in agreements that follow reality. Here agreements are drawn without a clear idea of what will happen in the project. A mismatch between the reality that is assumed by the legal department and that of the actual personal data flows may follow as a result.

5 Conclusion

In the theoretical discussion about Star and Griesemer’s divergence from ANT, the big difference was a difference of approach to problems. In ANT a problem was solved through careful manipulation of different actors until they acted in one way as planned by the most dominant actor. In this case, all other actors have to pass through an obligatory passage point. In boundary objects, a more ecological approach is put forward. Collaboration can only exist if knowledge is shared through the largest

[4] At least this was the case for the projects this author was involved in. Citizen science and participatory action research are becoming common place in smart city projects which means that hobbyists will become part of the solution. www.curieuzeneuzen.be is an example of an approach including hobbyists.

common denominator and when new actions are aligned with existing practices. For GDPR compliance, we face the same choice. Compliance can be achieved through domination or collaboration.

In the case of domination, this means that a compliance manager needs to take care that GDPR requirements become an obligatory passage point. Regardless of what other actors think or feel, there is no alternative than to comply with the behavior prescribed by the compliance manager. On a more pragmatic level this sort of GDPR compliance will look imposed and top down. A compliance manager will have to impose her or himself in project processes and demand to be reckoned with.

If collaborative GDPR compliance is the goal. This is not achieved by forcing other actors through an obligatory passage point, compliance from the boundary object point of view is achieved by fitting compliance efforts in the actions that already occur in smart city projects. This means that boundary objects and thus compliance documents and methods need to make sense and be useful for other tasks than mere compliance.

The Latter Approach has Advantages for Overall Privacy in Smart Cities

- If the aim is to increase privacy-by-design, then GDPR concepts should be present in the minds of developers and project owners.
- Boundary objects create a back and forth dynamic and the documentation of such a process may be interesting to understand the development and the meaning of a particular smart city. Moreover, such documents would increase accountability as all decisions are documented.
- Boundary objects allow laymen to aid in the easiest tasks of GDPR compliance. By enabling partners to map their own data processing operations, money and effort is saved that would otherwise go to an internal or external specialist. This would not only be costly but also a lost learning opportunity.

References

1. Star, S.L.: This is not a boundary object: reflections on the origin of a concept. Sci. Technol. Hum. Values **35**, 601–617 (2010)
2. Star, S.L., Griesemer, J.R.: Institutional ecology, ‘translations’ and boundary objects: amateurs and professionals in Berkeley’s Museum of Vertebrate Zoology, 1907-39. Soc. Stud. Sci. **19**, 387–420 (1989)
3. Trompette, P., Vinck, D.: Revisiting the notion of boundary object. Rev. Anthropol. Connaiss. **3**(1), 3 (2009)
4. Callon, M.: Some elements of a sociology of translation: domestication of the scallops and the fishermen of St Brieuc Bay. In: Power, Action and Belief: A New Sociology of Knowledge? pp. 196–223. Routledge, London (1986)
5. Callon, M.: Techno-economic networks and irreversibility. In: Law, J. (ed.) A Sociology of Monsters: Essays on Power, Technology, and Domination, pp. 132–161. Routledge, London; New York (1991)
6. Law, J.: After ANT: complexity, naming and topology. Sociol. Rev. **47**, 1–14 (1999)

7. Latour, B.: Where are the missing masses? The sociology of a few mundane artifacts. In: Bijker, W.E., Law, J. (eds.) Shaping Technology/Building Society: Studies in Sociotechnical Change, pp. 225–258. MIT Press, Cambridge (1992)
8. Akrich, M.: The de-scription of technical objects. In: Shaping Technology/Building Society: Studies in Sociotechnical Change, pp. 205–224. MIT Press, Cambridge (1992)
9. Mechant, P., Walravens, N.: E-government and smart cities: theoretical reflections and case studies. Media Commun. **6**, 119 (2018)
10. Walravens, N., Waeben, J., Van Compernolle, M., Colpaert, P.: Co-creating a practical vision on the smart city. In: Proceedings of the 15th Architectural Humanities Research Association International Conference, Eindhoven (2018)
11. Baccarne, B., Mechant, P., Schuurman, D.: Empowered cities? An analysis of the structure and generated value of the smart city ghent. In: Dameri, R.P., Rosenthal-Sabroux, C. (eds.) Smart City. PI, pp. 157–182. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-06160-3_8
12. Accountability. https://edps.europa.eu/data-protection/our-work/subjects/accountability_en

Workshop and Tutorial Papers

Who You Gonna Call When There's Something Wrong in Your Processing? Risk Assessment and Data Breach Notifications in Practice

Susan Gonscherowski and Felix Bieker(✉)

Unabhängiges Landeszentrum für Datenschutz (ULD, Independent Centre for Data and Privacy Protection) Schleswig-Holstein, Kiel, Germany
{sgonscherowski,fbieker}@datenschutzzentrum.de

Abstract. With the assessment of the risk to the rights and freedoms of natural persons the GDPR introduces a novel concept. In a workshop participants were introduced to the notion of risk, based on the framework of the German data protection authorities, focusing on personal data breach notifications. This risk framework was then used by participants to assess case studies on data breaches. Taking the perspective of either a controller or a data protection authority, participants discussed the risks, the information provided and the necessary steps required by the GDPR after a data breach.

Keywords: Data breach · Notification · Supervisory authority · DPA · Risk to rights and freedoms · Risk assessment · General Data Protection Regulation · Data protection · Privacy

1 Introduction

Over the last years, data breaches have occurred with increased frequency and more severe impacts on data subjects [1–8]. Since the General Data Protection Regulation (GDPR) has become applicable, new notification and communication requirements in case of a data breach must be fulfilled. As every breach is different, handling it appropriately may prove difficult especially in terms of consequences and risk assessment.

In order to determine whether the data protection authority has to be notified of a data breach and whether this has to be communicated to the data subjects, the controller has to evaluate the risk to the rights and freedoms of natural persons according to Articles 33 and 34 GDPR. This assessment is also crucial for the implementation of technical and organisational measures according to the provisions on the responsibility of controllers, data protection by design and the security of the processing under Articles 24(1), 25(1) and 32(1) and (2) GDPR as well as for the determination whether a Data Protection Impact Assessment has to be carried out (Article 35 GDPR) and whether the prior consultation mechanism of Article 36 should be triggered. The particular notion of risk in the GDPR is thus essential for the correct interpretation and implementation of the regulation as a whole.

Published by Springer Nature Switzerland AG 2019
E. Kosta et al. (Eds.): Privacy and Identity 2018, IFIP AICT 547, pp. 35–50, 2019.
https://doi.org/10.1007/978-3-030-16744-8_3

In this paper, we introduce the notion of risk in the GDPR as well as the obligations of controllers with regard to data breaches. This will be completed by a summary of the participants' discussions of two fictitious case studies during a workshop at the 2018 IFIP Summer School. Although the provisions on data breaches in Articles 33 and 34 of the GDPR impose some specific obligations on controllers, other aspects, such as the risk methodology, have to be determined. This is also true for determining actions to deal with risks resulting from the incident, e.g. identity theft.

The advantage of an interactive workshop was a more "hands-on" approach. Divided in two groups the participants took the perspectives of a controller or a data protection authority. The first case study involved a public hospital and the disclosure of several categories, including special categories of personal data. The task was to determine whether the submitted breach notification conformed to the requirements of Article 33 GDPR. The notification form is based on actual forms provided by data protection authorities.

The second case involved a private sector controller and in this case the data breach was based on a database mix-up. The second group was asked to determine whether the incident required communication with the data subjects according to Article 34 GDPR.

2 Assessing the Risks to the Rights and Freedoms of Natural Persons

The question, whether notification and communications of data breaches are necessary depends on the assessment of the risk to the rights and freedoms of natural persons according to Articles 33 and 34 GDPR. While the GDPR does not further define this notion of risk, recitals 75 and 76 state that risks of varying likelihood and severity may lead to damage for individuals.

2.1 The Notion of Risk to the Rights and Freedoms of Natural Persons

The specific notion of risk to the rights and freedoms of natural persons introduced in the GDPR can thus best be defined as the product of the likelihood and severity of potential damage for individuals [9]. However, this should not be taken to imply that this is a mathematically precise formula. Rather the assessment, according to recital 76 must be carried out with reference to the nature, scope, context and purpose of the processing and should, as prescribed by the principle of accountability according to Article 5(2) GDPR, be based on verifiable facts.

Recital 75 finds that the damage to individuals, which can be caused by the processing of personal data can be physical, material and non-material, for instance causing bodily harm, financial loss or deprive data subjects of their rights. With reference to Article 8 Charter of Fundamental Rights (CFR) this becomes even clearer: as every processing of personal data constitutes an interference with this fundamental right, any processing operation can potentially cause damage to the rights of individuals [10]. However, this interference can, of course, be justified under the conditions of

Article 52(1) CFR, which contains a clause on justifications for all fundamental rights contained in the CFR.[1] For Article 8 CFR this is the case when the interference caused by the processing of personal data is as minimal as possible, i.e. when the risks to the rights and freedoms of individuals have been mitigated appropriately [11]. However, there are also other rights that must be considered, such as the right to privacy under Article 7 CFR, the freedom of speech and assembly according to Articles 11 et seq. CFR as well as the rights to non-discrimination of Articles 21 and 23 CFR [12].

2.2 Risk Identification

The assessment must take into account negative consequences of the processing operation as planned as well as deviations from the processing, such as access by unauthorized parties, unauthorized or accidental disclosure, linking or destruction of data, failure or unavailability of designated procedures, accidental or intentional alteration of data or the failure to provide information. All of these incidents may be caused by parties internal or external to the controller. Thus, the assessment must include all potential negative consequences of a processing operation for the rights and freedoms of natural persons, their economic, financial and non-material interests, their access to goods and services, their professional and social reputation, health status and any other legitimate interests [9].

Furthermore, the sources of these risks must be identified. Under data protection law, the organization itself is a significant risk source, as it processes the individuals' personal data. Thus, for instance the marketing department or individual employees using data without authorizations pose risks for data subjects. However, risks may also emanate from authorized or unauthorized third parties, such as processors, contractors, manufacturers, hackers or public authorities, especially law enforcement, pursuing vested interests. Further, technical malfunctions and external factors, such as force majeure, have to be taken into account [13].

2.3 Risk Assessment

The likelihood and severity of potential damage must be assessed. However, attempts trying to ascribe a precise numerical value to either of these should be rejected, as they suggest an objectivity that is not attainable. Rather, the likelihood and severity should be classified in categories, which give an estimate and follow from a thoroughly argued justification providing the basis of considerations for the assessment. A classification could use a four-tiered scale, ranging from minor, limited to high and major [9].

[1] In order to be justified, the interference must respect the essence of the law, pursue a legitimate aim and be proportionate. On the level of secondary law, this is implemented by Article 6 GDPR: In order to protect the fundamental rights of individuals and as every processing of personal data interferes at least with Article 8 CFR, the processing of data is only permissible when it is based on a legal basis (as provided in Article 6(a)–(f) or (2) and (3)), which must be proportionate. Further, the controller must implement safeguards in order to ensure a level of security appropriate to the risk for fundamental rights, cf. Article 32(1) GDPR.

The likelihood describes how likely a certain event, which itself might be damage, occurs and how likely it is that this may lead to (further) damage. The motivation and the operational possibilities of an organisation to use data for incompatible purposes should, inter alia, be taken into consideration as criteria for the assessment of the likelihood [9].

The severity of potential damage must, according to recital 76 be determined with respect to the nature, scope, context and purposes of the processing. The Article 29 Working Party has identified criteria to assess whether a high risk is likely to occur for a processing operation [14]. These criteria are derived from provisions where the legislator considered potential damage to be particularly sever, such as where the processing occurs on a large scale (recital 75), affects vulnerable individuals, such as children or employees (recital 75), may prevent data subjects from exercising their rights (recitals 75 and 91), concerns special categories of data (Articles 9 and 10 GDPR) involves automated decision-making and profiling (Article 22 and 35(3)(a) GDPR), or allows for systematic monitoring (Article 35(3)(c) GDPR).

Once the likelihood and severity of potential damage have been assessed, the risk to the rights and freedoms of natural persons has to be evaluated and classified according to the categories of low, medium or high risk. However, the GDPR does not contain any provisions concerning a specific methodology for this evaluation.

The risk of the processing operation follows from the highest risk category of all individual risks. However, in cases where there are many individual risks within a category, this leads to cumulative effects, which require a higher classification of the risk [15].

3 Data Breaches

Data Breaches occur on an almost daily basis. Often, such an event confronts the controller with a multitude of problems. One of the first problems for the controller might be to determine whether or not an incident related to the data processing is a data breach at all. In 2017 approximately 20% of small and medium-sized businesses in Germany indicated they had had no IT-security incidents at all [16]. However, this is not a reason to celebrate the high standard in IT-security and data protection in those companies. It is more likely their detection measures are insufficient and incidents were simply not detected.

If and when a controller becomes aware of a problem, the next question is whether a security breach or a data breach occurred. A personal data breach is defined in Article 4(12) GDPR as “a breach of security leading to the accidental or unlawful destruction, loss or alteration, unauthorised disclosure of, or access to personal data transmitted, stored or otherwise processed”. While a data breach is often understood as an unauthorised disclosure of personal data [1–8], data protection law also classifies loss of integrity and/or availability as a breach. The reference to the three standing requirements for IT-security leads to the conclusion that every data breach is also a security breach, but not every security breach is always a personal data breach [14]. In case of an incident the controller must investigate if personal data is affected by the breach in any way. The awareness that personal data was indeed compromised then triggers a

72 h deadline for further actions. According to the GDPR, a notification of the supervisory authority (Article 33) and possibly the communication of the breach to the data subject (Article 34) may be necessary. At this point the risks posed to individuals by the data breach have to be assessed. This differs from the initial assessment of the risks posed by the processing operation, which is obligatory under the provisions concerning the responsibility of controllers, data protection by design and the security of the processing in Articles 24, 25(1) and 32 GDPR. As described above, the likelihood and severity are the main factors in risk assessment. In case of a personal data breach the likelihood of the risks relating to a breach is 100% leaving the severity as the only variable. The potential impact on the rights and freedoms of data subjects ranges from no risk, to risk up to a high risk.

Article 34 of the GDPR defines three preconditions that must be fulfilled before a controller has the obligation to communicate the incident to concerned individuals. Firstly, the controller must become aware of a breach. Secondly, the risk assessment leads to the conclusion that the breach poses a high risk to the rights and freedoms. Thirdly, the controller is able to identify the individuals affected by the breach.

There are circumstances when these pre-conditions are difficult to meet and therefore data subjects do not receive a message informing them of a breach that may have negative consequences for them. Some circumstances were already foreseen by the legislator. If the controller is not sure whether or not a breach occurred the issue may be investigated further for three days. Therefore, it is crucial that effective detection methods are being used. As mentioned above a considerable number of controllers and processors lack the ability to achieve this. However, the ability to detect a breach is an essential part to ensure the security of processing as demanded in Article 32 GDPR.

It is also possible that the controller responsible for a breach is no longer available, because the company went bankrupt or the service was terminated. For instance, unintentionally or unlawfully disclosed sensitive personal data originating from phishing attacks are freely available on the internet [17]. In these cases there is no controller to become aware of a breach. The legislator did not provide a solution in the GDPR. However the research project EIDI aims to develop a warning system in order to close this gap in responsibility [18].

The risk assessment for a certain data breach represents the risk for that precise moment. However, the external conditions may change. Technological developments can weaken encryption and the combination of data stemming from different breaches might reveal sensitive information e.g. linking clear text passwords for specific accounts. A breach classified as not being a risk to data subjects rights at the time of the incident might become a risk later on.

And lastly, the direct identification of the concerned data subjects may not be possible. For instance, the provided service may not require contact information or the individual is no longer using the service. This could mean there is no contact information or the available information is outdated. Of course, the incident itself can be the reason for missing contact information, e.g. when after a malware attack the relevant data is encrypted [19]. Article 34 GDPR, as a backup, requires information via public communication when no contact information is available. Yet, this requires the same level of effectiveness as a direct communication. It is doubtful that any public annunciation meets this high standard.

4 Hands-On: Assessment of Case Studies

After the input statements, participants were divided into two groups to discuss the two following case studies, identify risks for the rights and freedoms of natural persons and discuss them. These were then summarized by participants of each group and discussed with all participants.

4.1 Case Studies

Case Study 1: Public Hospital

A public hospital wants to combat cases of an acute disease, which can cause severe damage to the nerve system if not diagnosed within 24 h of the first symptoms. As some of these symptoms are similar to the flu, the other characteristic symptoms are often not properly recognized. In order to assist doctors in the diagnosis, the hospital wants to develop an app for its own managed mobile devices used by clinicians, which recognizes these symptoms and alerts doctors to this potential diagnosis. The app uses machine learning technology, which, based on patient data, constantly improves the recognition of relevant symptoms.

In order to train the algorithm, the hospital gives the department that carries out the development of the app full access to all of its 1.5 million patient data records. These data consist of records of former as well as current patients dating back to the 1980s and include the address, date of birth, phone number, occupation as well as the patient's medical history concerning all treatments at the hospital. When the data are collected, the hospital informs patients that their data will be processed in order to facilitate the treatment of their medical conditions at the hospital.

In order to process the patient data, the hospital moves the data, which it encrypts beforehand, in its own cloud environment. However, an unencrypted backup of the data is stored on a server with an open port, which leads to all the patient files that are analysed by the algorithm being available online. This concerns 150,000 of the hospital's patient files.

Task for Case Study 1. 23 h after this incident is detected, the hospital submits a notification to you (Table 1), the competent data protection authority. Determine whether the notification conforms to the requirements of Article 33 GDPR, focusing especially on the assessment of likely consequences of the breach, and covers the entirety of data breaches. Then consider which actions you would take next.

Discussions on Case Study 1. In the discussion of the first case study, participants quickly discovered flaws in the envisaged processing operation: they pointed out that the breach that had been notified to the supervisory authority might, in fact, not be the first data breach that had occurred during the processing operation. The definition of

Table 1. Notification form submitted to data protection authority

1. Controller
Contact Data Data Protection Officer of a Public Hospital in an EU Member State DPO@public-hospital.health
2. Timeline
When did you discover the breach? On Tuesday 21/8/2018 at 10:43
When did the breach occur? On Monday 20/8/2018 at 17:19
This notification is made within 72 hrs of discovery ⊗ Yes O No **If not, why was there no earlier notification?** ---
3. Description of Data Breach
Kind of data breach O Device lost/stolen O Papers lost/stolen/kept in unsafe environment O Unencrypted email sent O Mail was lost/opened accidentally O Hacking/Malware/Phishing ⊗ Accidental disclosure/publication O Wrong recipient(s) O Misuse of access rights O Other, please specify: ---
Please describe the data breach in detail Backup was uploaded to cloud environment; data was encrypted; port on server was

open after maintenance work and data accessible online; encryption can be broken due to security flaw in algorithm

Categories of personal data

⊗ Basic personal identifier, e.g. name, contact details
O Passwords
O Data revealing racial or ethnic origin
O Political opinions
O Religious or philosophical beliefs
O Trade union membership
⊗ Data on sex life or sexual orientation
⊗ Health data
⊗ Genetic or biometric data
O Criminal convictions, offences
O Location data
O Not yet known
O Other, please specify: ---

Number of individuals concerned?

150,000

Number of records affected?

150,000 patient records

Description of likely consequences of the personal data breach:

Loss of confidentiality of patient records, potentially identity theft

The personal data were safeguarded by the following appropriate technical security measures:

Encryption of personal data

4. Measure taken to address the personal data breach

Description of measures taken to address data breach:

Patient records have been removed from cloud environment, port on server has been closed

Description of measures proposed to be taken to address data breach:

Encryption with new algorithm

Description of measures taken to mitigate adverse effects of data breach:

Communication of breach to data subjects

Description of measures proposed to be taken to mitigate adverse effects of data breach:

--

5. Communication to data subjects

The personal data breach has been communicated to data subjects

⊗ Yes

On Tuesday 21/8/2018 at 18:00: all patients have received information about the breach to their contact details via email or mail, depending on available information

O No, as

O Appropriate technical and organisational measures have been taken; please describe: ---

O Follow-up measures ensure that the high risk for the rights and freedoms of data subjects does no longer exist; please describe: ---

O Communication to data subjects would involve disproportionate effort; please describe: ---

personal data breach under Article 4(12) GDPR also encompasses a breach of security leading to the unauthorised access to stored personal data. As the participants noted, the hospital, when collecting the data of patients stated that their data would be processed in order to facilitate their treatment at the hospital. While this purpose was rather generic, it could refer to both a contract on medical treatment according to Article 6(1)(b) in conjunction with Article 9(2)(h) or, for medical emergencies, the protection of vital interests of data subjects according to Article 6(1)(d) in conjunction with Article 9(2)(c) GDPR.

However, under the principle of purpose limitation of Article 5(1)(b) GDPR, the data could not be further processed for purposes that are incompatible with this initial purpose of treating a medical condition. It could be argued that the use of the data to train the algorithm was a processing of personal data for the compatible purpose of medical research under this provision, which would, however, be subject to the test of Article 6(4) GDPR, which requires that the controller takes into account inter alia the link between the purposes, the nature of the data, the possible consequences for data subjects and the existence of appropriate safeguards. As the hospital granted access to the full patient files, the participants found that with regard to the principle of data minimisation, which requires that only data necessary to achieve a specific purpose are processed according to Article 5(1)(c) GDPR, the hospital did not conform to the legal requirements for further processing. Therefore, the sharing of the patient record data with a different department of the hospital, which was not tasked with the treatment of the patients was an unauthorised and unlawful processing of personal data and hence constituted an independent data breach, which was not notified to the supervisory authority, even though there was a risk that patient data would be further disseminated than necessary. While this first breach was limited to an internal department of the hospital, it did concern all of the 1.5 million patient data records and thus occurred on a very large scale. Ultimately, with the subsequent data breach this specific risk to the rights and freedoms of patients even materialized with regard to the files of 150,000 patients.

Concerning the notification of the second data breach, the participants of the workshop noted that the hospital described neither the processing operation nor the breach accurately or in much detail. From the perspective of the supervisory authority, participants found that it would be helpful to receive a Data Protection Impact Assessment concerning the relevant processing operation in order to have a more concrete idea of the systems and data used as well as the controller's initial assessment of the risks that the processing operation entails. While this is not foreseen by Article 33 GDPR, the supervisory authority may request any relevant information from a controller under Article 58(1)(a) GDPR. However, participants pointed out that in a time-sensitive situation where the rights of individuals could be in jeopardy, requiring a formal request only after the submission of the breach notification

Most notably, the controller made the counterfactual statement that the backup stored in a cloud environment could be decrypted due to a security flaw in the algorithm, whereas the backup that was accessible through a port opened for remote maintenance was actually not encrypted at all. This extended to the claim that the personal data affected by the breach were safeguarded by encryption. While this was the case for the live data, the hospital stored its backup without any encryption, which seriously undermines the protection.

Participants further found the risk analysis of the controller to be lacking. By recourse to the framework for risk assessment provided in the first part of the workshop, they could easily identify risks beyond those noted by the controller, which encompassed only the loss of confidentiality of patient records and the potential for identity theft. In this regard, it could be seen that the controller was very much focused on an information security perspective. As the concept of breach notification originated in this field, this is not surprising. However, it is a common pitfall to equate an information security breach with a personal data breach. Instead it must be seen from the perspective of data protection law, which, unlike information security is not concerned with the protection of the controller, but rather states under Article 1 GDPR that it serves to protect the rights and freedoms of natural persons and especially data subjects. Therefore, the risks to the rights and freedoms of individuals, which must be assessed for a personal data breach, differ from those of information security.

Applying these principles to the case at hand, participants pointed out that beside the contact information, the files also included the medical records and thus a patient's medical history, which themselves constitute health data and are this covered by Article 9 GDPR as a special category of personal data. With regard to the patient's medical history, risk sources were not limited to criminal third parties engaging in identity theft. The health data could potentially be of interest to the data subject's employers, pharmaceutical companies, insurance companies as wells as banks or credit scoring agencies. From the identification of risk sources alone, several other risks to the rights of individuals could be deduced, such as an employer terminating the contracts of severely or chronically ill employees, while pharmaceutical companies could be interested in contacting individuals in order to market medicinal products. Furthermore, insurance companies could individualize the cost of insurances, such as life or health insurance and thus increase prices. Similarly, a credit scoring agency or bank could downgrade an individual's score where they are aware of heavy costs incurred by illness or decreased life expectancy.

Participants also criticized that due to the poor risk assessment carried out by the controller, the potential measures taken to address these risks were insufficient. Especially with regard to the communication of the data breach, the workshop participants were sceptical whether it would be possible for the hospital to reach all of the affected individuals, due to the fact that the patient's files dated back to the 1980s and the contact information of former patients may have since changed.

In order to improve the handling of personal data by the hospital, the participants argued that the hospital should process only the necessary information from the patient records and ensure proper pseudonymisation of these data. While pseudonymised data is still personal data in the sense of Article 4(2) GDPR, as an individual can still be identified with reference to the assignment function, pseudonymisation is a technical and organisational measure which helps to reduce the risks to the rights of individuals. Participants found that proper pseudonymisation should ensure that the individuals cannot be identified by the department carrying out the relevant research, i.e. the assignment function should remain in the department which initially collected the data, or, in order to provide an additional layer of security, be stored by a third party, which in turn has no access to the pseudonymised data. This would also serve to reduce the risk of a data breach, as it would be hard to identify individuals, if the data collected were reduced appropriately and perhaps randomized in order to hamper attempts to identify individuals by drawing inferences.

Case Study 2: Online Shop

The online shop "Fancy Foods" offers its European customer a wide variety of delicacies. To get to know its 3.5 million customers better and attract new customers. Fancy Foods' management launched the application my favourite poison for mobile devices where people can share and rate their favourite recipes.

In order to be able to share their own recipes a user first has to answer five questions about personal eating habits and then consent to the Terms and Conditions (including privacy policy) of the app. In the next step the user can swipe to the left to like a picture and to the right to dislike a dish. Afterwards the user gains access to the liked recipes' list of ingredients and may add these to a personal cookbook or delete the recipe. In both cases, the user needs to select from a variety of reasons why the dish was chosen or erased, e.g. categories like allergies, religious diet limitations, love for sugar and sweets or childhood memories, before access to new pictures is granted. To purchase ingredients for a certain recipe online the app user may just enter an email address, a credit card number and shipping address or log into the online account.

In 2017, a leading health insurance company (HIC) started a project together with Fancy Foods to counter the effects of an unhealthy diet. Thus, Fancy Foods created a separate database accessible for HIC containing pseudonymised costumer profiles. Aside from a user-ID (instead of login credentials) the database contains the whole user profile including address, credit card number and the reasons for selection and rejection of all recipes. After the subsequent update (1st August 2018) every registered member of the Fancy Foods online shop automatically receives dishes chosen according to their individual health needs in their personal cookbook.

Task for Case Study 2. Due to a wrong setting in the database the ID is not permanently linked to the rest of a data set. After the next database access (18th August 2018) the mix-up was detected and the insurance company immediately informed "Fancy Foods". Fancy Foods' IT department fixed the problem this morning and re-established the link between ID and data set.

Assume the position of the controller and decide whether the mix up requires communication with the data subjects. Use the attached form for documentation.

Discussions on Case Study 2. The second case study is based on a different breach type. Instead of unwanted disclosure of personal data, here the integrity of the data is compromised. This approach was specifically chosen to raise the participants' awareness for varying incidents. The task for group II combined the practical application of Articles 33 as well as 34 GDPR and the risk assessment introduced at the beginning of the workshop.

The participants had no problem with the identification of the integrity breach as this was already hinted at in the task. Article 34(2) GDPR states which information must be included in the communication of a breach to the data subjects. It refers to Article 33(3)(b) to (d) GDPR. The group used all provided information to describe the breach in as much detail as possible.[2] The counter measure re-establishment of the correct link between user ID and database was recognized as well. The distinction of awareness and occurrence was not debated. Yet, even with precise numbers some discussion was necessary. The affected individuals and records were at first set with 3.5 million, but then corrected to unknown. Here a few group members correctly objected to the figure of 3.5 million, because the database consists of the app-profiles and not the online shop costumers. The case did not state any user or download numbers

[2] The provided notification form differed in two aspects from the first one (see Table 1). Number three only referred to the basic IT-security incidents and did not mention specific examples and number five included further communication channels.

concerning the app. Also the affected categories, in particular the special categories according to Article 9(1) GDPR, like data on health and religion, were correctly identified.

Aside from these special categories, the risk assessment required that aggravating factors given in the case description were discovered first, e.g. missing or insufficient IT-security measures and the unlawful processing. This was partly directed to the basic principles of data processing referred to in Article 5(1) GDPR. Several of these were violated by the processing. The first five questions about data subjects' personal eating habits represent an unlawful processing, because the users consent was given after these data were processed. Data minimisation in the app could be increased: It is questionable why the user always needs to justify the decision to store or delete a recipe. The getting-to-know-you purpose may as well be accomplished with less personal data.

Participants further found a violation of the purpose limitation principle poses the project with the health insurance company: The participants indicated the purpose change and the missing consent for the disclosure of data from the app and the online shop, but did not raise questions concerning the risk of disclosing the complete user profile to the insurance company. Due to the groups' lively exchange of arguments the aspects of the profiling could not be discussed any further. The unlawful processing would have been the second data breach in this case study. The joint controllers neither asked the users for consent nor did "Fancy Food" limit the access to the stored user data and user profiles. The available database consists of almost the same personal data as the user profiles. HIC therefore processes personal data not necessary for the purpose of countering the effects of an unhealthy diet. Contact information, credit card numbers and addresses are not covered by this purpose. However these categories enable the company to link their own costumer database to app users. This could lead to higher insurance fees for those who are branded as unhealthy eaters, because their risk to suffer from diet-related diseases may be considered higher. Policy holders with a similar lifestyle that do not use the app would not need to pay the higher fees.

The risk assessment in this case study was quite challenging for the group. Discrimination based on health or religious data processed was discussed only in the context of the processing within the app. However, further negative consequences related to the processing of the insurance company were not as easily recognized. Furthermore, the database mix up poses not only a risk to data protection but could even be fatal to "Fancy Food" costumers with food allergies: For three weeks registered online shop members automatically received recipes chosen in according to their health requirements. As the selected dishes were based on another person's data and thus may have contained ingredients they could have caused an anaphylactic shock to the recipients who did not check the recipe.

A further risk of identity fraud can be identified with regard to the low level of security measures in the food ordering process. As no authentication procedure is mentioned, an attacker may use any email address, credit card and shipping address to order from the online store.

5 Conclusion

As can be seen from the introduction of the framework for the assessment of the risk to the rights and freedoms of natural persons, the GDPR introduces a new concept, which has been adopted from information security. However, it is important to stress that the concept has been adopted from its former context and has been fully adapted to the requirements of data protection law. This concerns most importantly the object of protection, which has shifted from the organisation using information technology to the protection of individuals subject to the processing of data. Like all data protection law, this concept thus serves to protect these individuals' rights and interests.

From the practical exercise carried out by way of participants using the risk framework of the GDPR to assess data breaches in two case studies several lessons can be learned:

The risk assessment in case of a data breach is crucial. Firstly, the initial risk assessment which has to be carried out to conform to the responsibility of the controller and ensure the security of the processing under Articles 24 and 32 GDPR is important in order to determine measures which prevent a data breach from happening. Secondly, the measures to be taken in cases where a data breach has occurred are dependent on a comprehensive assessment of the risks emanating from the specific data breach in question.

Furthermore, the information contained in a notification to the supervisory authority is very limited and should be supplemented by the initial risk assessment carried out by the controller and include a description of the processing operation or, in cases of a high risk processing operation, the Data Protection Impact Assessment.

The notification process itself depends not only on a correct risk assessment but also the right timing, detection methods and were necessary the communication channel. While Articles 33 and 34 GDPR provide some clues for controllers as to if, when, and how to react in case of a data breach there are many instances in which standard procedures may not be applicable. The case study illustrated how controllers can fail to consider data protection risks arising from inside their own organisation. In these cases, the controller will be the single point of failure.

On the other hand, legislators need to reconsider their focus on the controller in terms of breach notification procedures. An unwilling or unknown controller leads to a dead end in the notification process and leaves data subjects unprotected. Future work should thus develop ways to address this very practical issue for data subjects.

Acknowledgement. This work is partially funded by the German Federal Ministry of Education and Research through the project 'Forum Privacy and Self-determined Life in the Digital World', https://www.forum-privatheit.de/forum-privatheit-de/index.php and the project EIDI (efficient notification after digital identity fraud), https://itsec.cs.uni-bonn.de/eidi/.

References

1. Khaira, R.: Rs 500, 10 minutes, and you have access to billion Aadhaar details. The Tribune, 4 January 2018. http://www.tribuneindia.com/news/nation/rs-500-10-minutes-and-you-have-access-to-billion-aadhaar-details/523361.html
2. Barret, D., Yadron, D., Paletta, D.: U.S. Suspects Hackers in China Breached About 4 Million People's Records, Officials Say. Wall Street Journal, 5 June 2015. https://www.wsj.com/articles/u-s-suspects-hackers-in-china-behind-government-data-breach-sources-say-1433451888
3. Donelly, L.: Security breach fears over 26 million NHS patients. The Telegraph, 17 March 2017. https://www.telegraph.co.uk/news/2017/03/17/security-breach-fears-26-million-nhs-patients/
4. Swedish authority handed over 'keys to the Kingdom' in IT security slip-up. The Local, 17 July 2017. https://www.thelocal.se/20170717/swedish-authority-handed-over-keys-to-the-kingdom-in-it-security-slip-up
5. Goel, V., Perlroth, N.: Yahoo Says 1 Billion User Accounts Were Hacked. New York Times, 14 December 2017. https://www.nytimes.com/2016/12/14/technology/yahoo-hack.html?action=Click&contentCollection=BreakingNews&contentID=64651831&pgtype=Homepage&_r=0
6. Haselton, T.: Credit reporting firm Equifax says data breach could potentially affect 143 million US consumers. CNBC, 7 September 2017. https://www.cnbc.com/2017/09/07/credit-reporting-firm-equifax-says-cybersecurity-incident-could-potentially-affect-143-million-us-consumers.html
7. Cadwalladr, C., Graham-Harrison, E.: Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. The Guardian, 17 March 2018. https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election?CMP=twt_gu
8. Ghorayshi, A., Ray, S.: Grindr Is Letting Other Companies See User HIV Status And Location Data. BuzzFeedNews, 2 April 2018. https://www.buzzfeed.com/azeenghorayshi/grindr-hiv-status-privacy?utm_term=.oj8dJKebLJ#.hwOGAMBZKA
9. DSK, Kurzpapier Nr. 18: Risiko für die Rechte und Freiheiten natürlicher Personen. https://www.datenschutzzentrum.de/artikel/1225-Kurzpapier-Nr.-18-Risiko-fuer-die-Rechte-und-Freiheiten-natuerlicher-Personen.html
10. ECJ, Judgment of 9 November 2010, Volker und Markus Schecke und Eifert, C-92/09 and C-93/09, ECLI:EU:C:2010:662, paras. 60–63
11. Bieker, F.: Die Risikoanalyse nach dem neuen EU-Datenschutzrecht und dem Standard-Datenschutzmodell. Datenschutz und Datensicherheit (DuD) **42**, 27–31 (2018)
12. Bieker, F., Bremert, B., Hansen, M.: Die Risikobeurteilung nach der DSGVO. Datenschutz und Datensicherheit (DuD) **42**, 492–496 (2018)
13. Friedewald, M. u.a.: White Paper Datenschutz-Folgenabschätzung, 3rd edn (2017). https://www.forum-privatheit.de/forum-privatheit-de/texte/veroeffentlichungen-des-forums/themenpapiere-white-paper/Forum-Privatheit-WP-DSFA-3-Auflage-2017-11-29.pdf
14. Article 29 Working Party, Guidelines on Personal data breach notification under Regulation 2016/679 of 3 October 2017, WP250rev.01. http://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612052
15. The Standard Data Protection Model (SDM), V.1.0 EN1 (2017). https://www.datenschutz-mv.de/static/DS/Dateien/Datenschutzmodell/SDM-Methodology_V1_EN1.pdf
16. Henseler-Unger, I., Hillebrand, A.: Aktuelle Lage der IT-Sicherheit in KMU, Datenschutz und Datensicherheit (DuD) (2018)

17. Malderle, T., Wübbeling, M., Knauer, S., Sykosch, A., Meier, M.: Gathering and analysing identity leaks for a proactive warning of affected users. In: Proceedings of the ACM International Conference on Computing Frontiers (CF 2016). ACM, New York (2018). https://itsec.cs.uni-bonn.de/eidi/files/malderle-cf18.pdf
18. EIDI. https://itsec.cs.uni-bonn.de/eidi/
19. Blinder, A., Perlroth, N.: A Cyberattack Hobbles Atlanta, and Security Experts Shudder. New York Times, 27 March 2018. https://www.nytimes.com/2018/03/27/us/cyberattack-atlanta-ransomware.html

Design and Security Assessment of Usable Multi-factor Authentication and Single Sign-On Solutions for Mobile Applications
A Workshop Experience Report

Roberto Carbone, Silvio Ranise, and Giada Sciarretta(✉)

Security and Trust, FBK, Trento, Italy
{carbone,ranise,giada.sciarretta}@fbk.eu

Abstract. In this interactive workshop we focused on multi-factor authentication and Single Sign-On solutions for mobile native applications. The main objective was to create awareness of the current limitations of these solutions in the mobile context. Thus, after an introduction part, the participants were invited to discuss usability and security issues of different mobile authentication scenarios. After this interactive part, we concluded the workshop presenting our on-going work on this topic by briefly describing our methodology for the design and security assessment of multi-factor authentication and Single Sign-On solutions for mobile native applications; and presenting a plugin that helps developers make their mobile native application secure.

1 Introduction

This paper is a report of the workshop "Secure and Usable Mobile Identity Management Solutions: a Methodology for their Design and Assessment" presented at the *13th International IFIP Summer School 2018 on Privacy and Identity Management - Fairness, accountability and transparency in the age of big data* held in Vienna, Austria.

Context. We focused on the design and security assessment of solutions for mobile native applications (hereafter native apps) with two features: Single Sign-On (SSO) and Multi-Factor Authentication (MFA); these two features are extremely important indeed: SSO allows users to access multiple apps through a single authentication act performed with an identity provider (IdP), for example Google or Facebook; while a MFA is a procedure that enhances the security of an authentication process by using two or more authentication factors (e.g., a password combined with the use of a fingerprint). A good design choice is to combine these features to have a good balance between usability and security.

While there exist many secure MFA and SSO solutions for web apps, their adaptation in the mobile context is still an open challenge. The majority of mobile MFA and SSO solutions currently used are based on proprietary protocols

E. Kosta et al. (Eds.): Privacy and Identity 2018, IFIP AICT 547, pp. 51–66, 2019.
https://doi.org/10.1007/978-3-030-16744-8_4

and their security analysis lacks standardization in the structure, definitions of notions and entities, and specific considerations to identify the attack surface that turns out to be quite different from well understood web scenarios. This makes a comparison among the different solutions—in order to choose the proper solution for a specific scenario—very complex or, in the worst case, misleading. Due to this lack of specifications and security guidelines, designing a mobile MFA and SSO solution from scratch is not a simple task; and as its security depends on several trust and communication assumptions, in most cases, could result in a solution with hidden vulnerabilities. In addition, it is necessary to take into account the legal aspects of the country where the MFA and SSO solution will be deployed. However, when innovative solutions are analyzed it is not an easy task to understand which legal obligations follow.

The main goal of the workshop was to make participants aware of the current security and usability issues of MFA and SSO solution in the mobile context. Participants were first introduced to the context and then, through the use of exercises we openly discussed several illustrative scenarios. Finally, we described our methodology for the design and security assessment of mobile MFA and SSO solutions. The design space is characterized by the identification of: *(i)* national (e.g., Sistema Pubblico di Identitá Digitale - SPID for Italy [3]) and European (e.g., electronic IDentification Authentication and Signature - eIDAS [15]) laws, regulations and guideline principles that are particularly relevant to digital identity; *(ii)* a list of security and usability requirements that are related to authentication solutions; *(iii)* a set of implementation mechanisms that are relevant to authentication and authorization on mobile devices and provide an easy way to satisfy the requirements in *(ii)*. To validate our approach, we applied it to a number of real-world scenarios that represent different functional and usability requirements. In this workshop, we applied our methodology to a real use-case scenario (called *TreC*) that supports the usage of mobile health apps. TreC (acronym for "Cartella Clinica del Cittadino", in English Citizen's Clinical Record) is an ecosystem of services that supports doctors and patients in the health-care management, by enabling all citizens living in the Italian Trentino Region to access, manage and share their own health and well-being information through a secure access (currently used by more than 80.000 patients).

Workshop Objectives. The main objectives of the proposed workshop were the following:

- to enable the audience to acquire the basic notions and the state of the art of MFA and SSO solutions for native apps;
- to create awareness of usability and security problems together with legal provisions related to authentication in mobile computing;
- to provide an overview of the techniques commonly used to analyze the security of an authentication solution;
- to perform an experimental evaluation of security and usability of MFA solutions for native apps.

Expected Contributions from the Audience Members. To raise the participants' awareness of the current possible limitations on usability and security of mobile MFA and SSO solutions, several questions and exercises were discussed together. The information extracted from the discussion has been useful in two ways. On the one hand, we were able to validate our hypothesis on usable solutions and to understand which security level is perceived. On the other hand, we were able to evaluate our methodology asking feedback to possible user.

Intended Audience, Including Possible Assumed Background of Attendees. The workshop was oriented to academic researchers, (PhD) students, security experts from industries that work on or want to approach the field of identity management. The attendees did not require a specific background on authentication to follow the main part of our workshop, as our step-by-step teaching approach enabled them to grasp the information presented even if some of the concepts were new or not consolidated. A dozen technical (IT security) and legal researchers took part to our workshop. The slide of the workshop are available at https://st.fbk.eu/workshop-ifipsc-18.

Paper Structure. In Sect. 2, we describe the content of the workshop on MFA and SSO solutions for native apps. Section 3 details the workshop structure and the assigned exercises. In Sect. 4, we present the exercises and discuss the outcomes. Finally, in Sect. 5 we discuss some lessons learned and describe our on-going work on this topic.

2 Content of the Workshop

We make a large use of our digital identities in our everyday life, from accessing social apps to security critical apps like e-health or e-banking apps. Underlying these transactions there is the exchange of personal and sensitive data, which could be exploited by a malicious intruder to impersonate or even blackmail a user. For this reason, many Identity Management (IdM) solutions have been designed to protect user data. In general, IdM refers to different aspects of the digital identity life-cycle (e.g., the creation and the provision of identities, password management and so on). In this workshop, we focused on the aspects related to authentication.

Password-Based Authentication. A common authentication mechanism is the password-based authentication, however its use is resulting in many attacks (e.g., identity theft). There are two main reasons. First, users are very bad in inventing and remembering passwords. [11] shows the list of the top 100 worst passwords of 2017, where at the first place there is "123456". This password is very easy to remember but it is also easy to crack; attackers have rainbow tables and dictionaries that contain this kind of credentials. Second, users re-use their passwords on several services: as reported in [9], more than the 54% of people use only 5 or fewer different passwords across their entire online life. This means that, if their credentials are compromised, for example after a guessing attack (easily performed by an hacker if the password selected is one of the 100 worst passwords [11]) then the attacker can access all the user data in different online services.

There exist "complexity tips" which allow users to choose proper passwords. For example, though being nine digits long, a password such as '123456789' could be instantly and easily cracked. According to former NIST recommendations, properly complex passwords should contain lower- and upper-case letters, as well as special symbols and numbers, for them to be secure enough that it could take years to crack them (via brute-force or dictionary attacks). However, the current NIST guidelines [6] recommend to follow an entirely different scheme, i.e. a good password should be made of a random, long phrase.

Password-Based Authentication and SSO. To permit the user to choose a complex password and access different services, an advisable design choice is to combine the password-based authentication with a Single Sign-On (SSO) solution. SSO allows users to access multiple apps through a single authentication act performed with an IdP, for example Google or Facebook. A common practice is to adopt the state-of-the-art standards, like SAML 2.0 [8] and OpenID connect [20] (OIDC). SAML 2.0 is pervasively used in the corporate environment, while OIDC is mainly used in social apps. There are two main advantages of using SSO. First, users do not need to register with an app to access it. Thus the user can choose a single complex password (providing usability and security). Second, if a user has already an active login session with an IdP, then she can access new apps without entering her IdP credentials anymore (providing usability). A SSO security drawback is that users are using only one password to access many services. Again, the user could select a more complex password, but still there is a single point to failure.

Multi-factor Authentication and SSO. A better design choice is to combine SSO with MFA solutions. Where a MFA is a procedure that enhances the security of an authentication process by using two or more authentication factors, such as combining a password that is something you know, with a hardware token that is something you have or the use of a fingerprint that is something you are.

Usually a MFA procedure requires the generation of a OTP (One Time Password). That is an una-tantum code that proves the possession of the OTP generator and optionally, if protected by a PIN, proves the knowledge of the PIN as well. There are different OTP generation approaches, in this workshop we focus on the Time-based OTP approach, where the OTP is generated starting from the current time of the operation and a secret key shared between the OTP generator app and the IdP. IdP must validate this value: only OTPs that fall into a short temporal range are accepted.

There are many MFA solutions on the market, and some of them are based on FIDO [5]. FIDO is a standard for password-less and MFA authentication that allows online services to augment the security of their existing password-based solution by adding a MFA procedure. To provide a FIDO two-factor solution, the organization must provide to its users a physical FIDO U2F device, like a USB key.

MFA and SSO Solutions for Native Apps. In this workshop, we focused on the design and security assessment of solutions for native apps with two features: SSO (for usability) and MFA (for increasing security). We focused on native apps—which differ from browser-based mobile apps as they are not accessed through a browser but they need to be downloaded from a marketplace—as the market is pointing on their use. Think about the many times you are suggested to download the app while you are navigating it in the browser; or the limitations that you can have from a browser-based version of an app. For example, this is the case of the browser-based version of TripAdvisor that provides less functionalities compared to the native app, such as the management of the reviews: users can read reviews on the browser version, but if they want to contribute on one review they have to download the native app.

As we will detail in Sect. 4, the known standards and solutions currently available for browser-based authentication (e.g., SAML or OIDC) cannot be easily reused in the mobile context, as browser-based and native apps are based on different security assumptions. Even if these are very good solutions there are still some limitations. Being proprietary protocols they cannot be customized and they do not necessarily satisfy the security requirements of a company. For example, an identity used in the social network solution is self-declared by the user. And so, it cannot be used by a company which needs to be sure of the real identity of the user. A first attempt of designing a solution for mobile authentication was carried on by big companies (e.g., Google and Facebook) that have designed their own solutions based on their security assessment. At the same time, the OAuth working group has released some guidelines. The current best practice was released in 2017. The solution proposed is called "OAuth 2.0 for Native Apps" [18]. Even if it is a good starting point, it does not cover some aspects. For example, it does not mention how to extend the protocol to support MFA or more complex environments, where for example different standards are used. Thus, in some specific cases and based on the requirements, a company is required to design a new ad-hoc solution.

The Importance of a Careful Design Phase. Designing a security protocol from scratch is not a simple task, as many aspects must be taken into account, such as how to establish trust, how to choose the right communication channel or how to evaluate the compliance of the designed solution with the current legal obligations, thus it is not recommended. Moreover, after the design, it is not simple to choose the right method to evaluate the corresponding security. Given all these aspects, it is clear that the design phase is not trivial and wrong design choices could lead to serious security and usability problems.

An example of a wrong design choice is the use of SMS as a second-factor authentication. This authentication method consists of the following steps: first the user has to enter her credentials into the app, then she will receive an SMS containing a OTP, and finally this OTP is entered by the user in the app. NIST [7] points out that this solution could be vulnerable to two kinds of attacks:

Social Engineering. "An out of band secret sent via SMS is received by an attacker who has convinced the mobile operator to redirect the victim's mobile phone to the attacker" (e.g., using SIM swap [14]).

Endpoint Compromise. "A malicious app on the endpoint reads an out-of-band secret sent via SMS and the attacker uses the secret to authenticate".

Even if these attacks are well known in the security community, there are still many companies that are using this authentication method. This is causing the spread of many security breaches. For example, this is the case of the social network platform Reddit, attacked in August, 2018. In [10], the Reddit security experts specified that the attack was related to the use of SMS and that they are now moving to a token-based second factor authentication method. Another example is described in [2], where a user was victim of two SIM hijacking attacks and now he is suiting the telecommunication company for a total of 224 million dollars. This example clearly demonstrates how a wrong design choice could damage not only the end-user but also the company.

Our Methodology. Given that designing a new security protocol from scratch is not an easy task and could result in a solution with hidden vulnerabilities, we have contributed with the definition of: a *reference model* for MFA and SSO for native apps, and a *methodology* to assist a designer in the customization of our model and in the analysis of its security and usability.

Our reference model is inspired by the Facebook solution [4] and OAuth 2.0 for native app [18]. We have extended these solutions in a way that they can be used by any IdP willing to provide its own SSO solution, meaning that the resulting SSO does not necessary leverage on identity provided by social IdP, and (optionally) a MFA. Currently, our models support two different OTP generation approaches: TOTP and Challenge-Response. Full details about the reference model based on TOTP can be found in [21].

Together with the reference model we have defined a methodology to assist a designer in the customization of our reference model and in the analysis of the resulting security and usability. In the first phase, we ask the designer to clarify the application scenario by filling a table that we provide (specifying the entities involved, the type of data that will be processed and which are the authentication requirements). Given this table, in the customization phase we are able to instantiate our model for this specific scenario and we provide as output a message sequence chart of the flow and a set of assumptions and security goals. These values are then given as input to the security analysis phase. We provide a semi-formal and a formal analysis. The output of this phase is a security analysis report. If some serious attacks are found then the designer has to go back in the customization phase and change the design otherwise the designer can proceed with the last phase. In the usability analysis phase we are asking to validate the usability satisfaction. If no problems are reported then the final solution is generated; otherwise the designer has to go back to the definition of the requirements and refine the design accordingly.

To validate our methodology, we have applied it to different real-world scenarios which consider different authentication and usability aspects. During this workshop we had detailed the e-health scenario TreC.

3 Structure of the Workshop

To raise the participants' awareness of the current limitations on usability and security of mobile MFA and SSO solutions, together with the background context described in Sect. 2, several questions (labeled with 🗪) and exercises (labeled with 📝) were discussed together. Figure 1(a) shows the background and current position of the participants (divided in two groups) and the outline of the exercises. The workshop followed the following structure:

Introduction and Problem Statement. In this introductory part we provided participants with the background described in Sect. 2 and we pointed out which are the current limitations related to the development of usable authentication solutions that are also secure in the mobile context.

🗪 *Browser-based vs Mobile Authentication.* After the description of the browser-based OIDC standard, we asked the participants if—in their opinion—this solution (and more in general, any browser-based solutions) can be reused also in the mobile context. We decided to ask this question to evaluate the participant's awareness and understanding of the differences between a mobile and a browser-based solution and the fact that we cannot easily reuse solutions that are developed in one context in another.

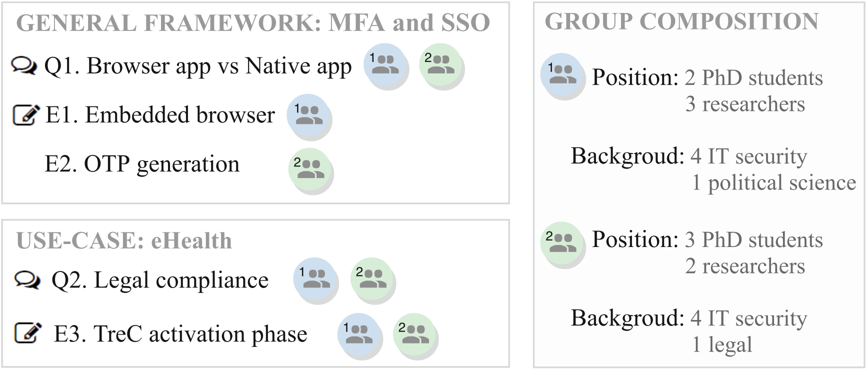
Design Choices: Security and Usability Problems. Designing a security protocol from scratch is not easy and the design phase is very important in terms of striking the right balance between security and usability. To raise this warning, during the workshop we asked participants to identify which are the security and usability problems related to two wrong design choices:

📝 *User Agent Choice: embedded browser.* This exercise is related to the choice of using an embedded browser as user agent. So we asked the participants to evaluate the usability and security of a solution where users enter their credentials in an embedded browser, namely a browser that is managed by an app.

📝 *OTP Choice: app that shows the OTP value to the user.* During this exercise, we asked the participants to evaluate the usability and security of an OTP generator app that shows the OTP value on the smartphone screen.

Methodology Overview: TreC Scenario. In this part, we described our methodology for the design and security assessment of mobile authentication solutions, applied directly to a real-world use case scenario, called TreC.

🗪 *e-Health Legal Compliance.* Being TreC a personal health record platform, we briefly mentioned the Italian legal aspects concerning health data, and more in general to sensitive data. Then, with the aim of having a broader



(a) Question and exercise structure. (b) Group Composition.

Fig. 1. Interactive workshop structure.

view on the legal aspects, we asked the participants to present the legal obligations of their country.

Usability Discussion on TreC. In relation to the TreC solution, we discussed with the participants some usability problems that resulted from proposing our solution to patients.

✎ *TreC activation phase.* We asked the participants to suggest some changes in order to simplify the TreC activation phase taking into account security.

Conclusions and On-Going/Future Work. Finally we summed up the main points of the workshop and presented our on-going and future work.

The answers and the discussions are reported in the following section.

4 Outcomes of the Exercises

In this section, we report the solutions to the exercises and discussions introduced in Sect. 3. To promote an interdisciplinary approach, we divided the participants in two groups based on their backgrounds and current position (see Fig. 1(b)). Each group included four men and a woman.

4.1 🗨 Browser-Based vs Mobile SSO Solutions

Q1. Can we use browser-based authentication and SSO solutions for native apps? We gave 5 min to discuss the problem and then we asked for the individual answers.

Participant's Answers: 5 "yes" and 5 "no". With 2 participants that voted "yes" saying that actually they would prefer to vote for "it depends", clarifying that it depends on the scenario and the security level required.

Our Answers: The use of a browser-based authentication protocol in the mobile context requires a detailed understanding about the differences between the two scenarios [13,22], and at the end it is pretty clear that we need to design a new flow ad-hoc for the native apps. Let us have a look at the differences.

The first difference is the Service Provider (SP) type: SP is not an app running in the browser but it is a native app. So we have to consider all the vulnerabilities related to a mobile platform.

Secondly, the User Agent (UA) that is used by the user to interact with the SP could be of different types: it could be a browser embedded inside the SP app, or an external browser that is installed in the user smartphone, or even an app released by the OIDC provider. So, we must take into consideration that now the SP app and the UA could not be played by the same entity as was in the browser case.

In addition, the redirection mechanisms between the OIDC provider and the app are different. Indeed, in a browser redirection you can uniquely identify the SP using its hostname. This ability is not always available in the mobile case as you are required to redirect to a specific app in the user's smartphone.

Finally, in the mobile case, the use of a SP backend is optional (you can have a native app that does not require a backend), so we have to adapt the flow by directly managing the authentication from the mobile. So in this case, we cannot use the client secret for authenticating the SP since in a mobile device we cannot store a secret as all the stored values are readable, at least by the owner of the smartphone.

For all these reasons, the reuse of available browser-based solutions in a mobile context is not obvious and it is necessary to redesign them taking into account the differences highlighted above.

Evaluation: The question was not fully clear and some participants were not able to answer.

4.2 Wrong Design Choices

In some specific cases and based on the requirements, a company or an organization could be forced to design a new authentication solution to provide secure mobile authentication solutions to their employees. However, designing a security protocol from scratch is not a simple task, as many aspects must be taken into account. In this section, we report the two scenarios that have been proposed to the workshop participants to highlight examples of wrong design choices. We proposed two exercises and asked Group 1 to tackle Exercise 1 (*E1*) and Group 2 to solve Exercise 2 (*E2*) in parallel. We allowed 10 min to elaborate a solution to each group.

📝 User Agent (UA) Choice: Embedded Browser

E1. We asked the participants to evaluate the usability and security of a solution where users enter their credentials in a browser that is managed by a native app (so called embedded browser). In detail, we asked the members of Group 1 to focus on the following questions:

1. How does an embedded browser work?
2. Are there any security issues?
3. How would you rate the user experience when accessing multiple apps?

Participant's Answers:

1. An embedded browser is a component inside your app that opens a website URL.
2. A native app has full control of the embedded browser, so if you are typing a password, the attacker can read it.
3. Not so good. Since an embedded browser is a separate browser-instance for all the native apps, users have to re-enter their password for all the native apps.

Our Answers:

1. An embedded browser is defined in [18] as "a user-agent hosted inside the native app itself (such as via a web-view), with which the app has control over to the extent it is capable of accessing the cookie storage and/or modifying the page content". The relevant bit from the point of view of security is that a native app is in control of the embedded browser.
2. The use of this type of browser is widely discouraged, as there is a loss of isolation between the app and the browser [19]. If the app is malicious, then it can steal the user credentials or change the authorization permissions. This is an example of a JavaScript added by a malicious app to steal user credentials:

```
webView.evaluateJavascript(
‘‘(function() {return document.getElementById(‘pwd’).value;})();’’,
new ValueCallBack<String>() {
@Override public void onReceiveValue(String s){
Log.d(‘‘WebViewField’’,s);
}
});
```

3. An additional limitation when choosing an embedded browser is that it does not provide a SSO experience. Indeed, if the browser is integrated within the app, then the login session information is stored in (and only accessible to) the app and it is therefore not available to other apps. This forces the user to re-enter credentials even if she has an active login session with an IdP. This is a frustrating experience, especially due to the small-virtual keyboard of a smartphone.

Evaluation: The participants were able to answer in the correct way.

OTP Choice: App that Shows the OTP Value to the User

E2. This exercise is related to the choice of using a native app for showing an OTP value to the user.

1. How would you rate the user experience when accessing a native app?
2. Is there any security issue?
3. List one or more OTP choice alternatives.

Participant's Answers:

1. We thought that usability really takes a hit if you have to switch between your native app and the OTP app all the time. This is the main thing. The usability can also be affected if you have multiple devices lost or broken.
2. Malicious apps trying to exfiltrate the OTP code or shoulder-surfing attacks (you are just behind someone and take a look at the OTP code that can be memorized or can be picked using a mobile camera).
3. Hardware tokens.

Our Answers:

1. Moving from an app to another is burdensome for the user in terms of time and difficulty.
2. To avoid the burden of remembering the OTP value, some apps (e.g., MySielteID[1] app of the Sielte SPID IdP) have a button for coping the OTP value. This is a serious security issue as the clipboard can be accessed by any app installed in the smartphone so that malicious apps can easily steal the OTP that has been copied.
3. An alternative choice is the use of a solution that does not ask the user to enter the OTP value, but after the PIN input, the OTP value is sent to the IdP server in a transparent way (namely, without any involvement of the user). The other alternative is the use of external OTP generators, such as FIDO keys or eID smartcard with capability of implementing a challenge-response OTP approach (using NFC and a smartphone as a card reader). The advantage of using an eID card over other external hardware tokens is that usually users bring it with them to permit an in-person identification; thus we can assume that an eID card is always available to the user also for online access.

Evaluation: The participants were able to answer in the correct way.

4.3 e-Health Legal Compliance

Being TreC a personal health record platform, we briefly mentioned the Italian legal aspects concerning health data. Then, with the aim of having a broader view on the legal aspects, we asked the participants to present the legal obligations of their country.

[1] https://play.google.com/store/apps/details?id=it.company.sielte.

At the time of running the workshop, when dealing with sensitive health data—that are a particular type of personal data—in Italy, we had[2] to follow the Data Protection Code [16]. [16] says that the data controller shall adopt the minimum security measures in order to protect these data. More technological details can be found in CAD [12] that is the Italian code for the public administration. In particular, [12] specifies which digital identities can be used in this context: CNS (a smartcard used to access online services of the public administration), CIE 3.0 (the Italian electronic identity card) or the Italian national ID scheme called SPID (from the second level up).

Q2. Which legal obligations do you have to follow when dealing with e-health data in your country?

Participants' Answer (from Germany): We apply the GDPR, especially for sensitive health data. The basic approach is: you are not allowed to process any data if you cannot provide the level of security that is necessary to protect the data; it is not a minimum standard but it must be the state-of-the-art. Additionally to the GDPR there are also rules/guidelines specific for health data.

Extra Discussion: we report here different interesting discussions that came up in relation to this question.

The first is a discussion on the minimum security measures required in [16]. We explained that the Annex B of [16] specifies a set of specific security measures, such as: sensitive data must be protected with an authentication method with passwords of at least 8 characters. [16] was in force pre-GDPR and we agree with the participants on the fact that the GDPR approach is the opposite: they do not suggest any kind of measures; they should be state-of-the-art, and it should be proved that companies have done their best to comply and provide enough security.

Then, we have discussed who is in charge of managing health data. In Italy, each region is responsible for applying and developing solutions for healthcare. In Germany, it depends of what you do. For example, hospitals have to follow the federal state law, while the health insurance is the same for all the country.

Finally, we briefly discussed about the privacy concerns arising by the adoption of a SSO solution. A participant draw our attention to the fact that the usage of a SSO protocol every time a user wants to log-in creates a single point of knowledge that can become a major threat to privacy. This problem is particularly acute when considering the so-called "consumer SSO": what happens to consumers data when they enable access to other applications or accounts from Facebook, Gmail, LinkedIn, Twitter, or a host of other providers? Indeed,

[2] During the preparation of this workshop, the Italian government was in the process of adopting the EU General Data Protection law (GDPR [17]), with some delay compared to other member states due to the national elections. Now, [16] was amended by the decree adapting the national legal system to the GDPR 2016/679 (Legislative Decree No. 101 of 10 August 2018).

this enables service providers with the capability of monitoring and collecting information on consumers habits and preferences as they browse the Web. While acknowledging the importance of these and related privacy concerns, we consider them outside the scope of the workshop which focuses on the trade-off between security and usability of SSO solutions. We just observe how security and privacy may be contentious even in SSO solutions: the capability of tracking and profiling users can be used by an identity provider to spot when an attacker is trying to impersonate a legitimate user. This is known as behavioral authentication and is provided by, e.g., Google which alerts if a user account is accessed from a location which is not among the usual ones.

Evaluation: Different interesting discussions arose from this question.

4.4 ✍ TreC Activation Phase

TreC activation phase is performed by the patient, only once and after she download the OTP-PAT app. It is performed partially on her laptop and partially on her smartphone:

- On her laptop, patient logs in using her CNS (the card is read by desktop smartcard reader), and generates a temporary code (it lasts 5 min).
- On her smartphone, patient downloads the OTP-PAT app using an official marketplace. Then, she enters the temporary code together with her credentials into OTP-PAT. If the login is successful, the activation phase is completed by patient with the creation of a PIN code.

As a consequence of this phase, OTP-PAT obtains two values: a token that is used as a session token in place of the user credentials to provide a SSO experience; and a seed value—stored encrypted with the PIN code selected by patient during this phase—that is used to generate OTPs.

For the TreC scenario, we have performed a pilot involving a controlled set of patients. Regarding usability, we found out that the activation phase is considered too complex. The main reason was the use of a smartcard reader that for being used, needs the installation of a specific software, which sometimes does not work properly. In addition, users were annoyed by the requirement of choosing complex passwords inside the mobile and they tend to easily forget them.

E3. What would you suggest to change in order to simplify the activation phase taking into account security? Note that the activation phase must provide a good level of assurance on the real identity of the user. In the previous solution this was implied by the use of a smartcard and the generation of an activation code, specific for the particular installation of the app. We gave 10 min to debate this issue.

Participant's Answers: they came up with two ideas:

1. Use one of the existing identity infrastructures. For example, in Austria they can use an identification method provided from the post system or in Sweden, if you have a bank account, then you can request a BankID that you can use for accessing online services;
2. Perform a face-to-face authentication and then send an activation code via email.

In addition, they observed that the objection on the password complexity is just an educational thing: people do not understand the security behind this design choice.

Our Answer: We change the activation phase as follows:

- On her laptop, patient logs in with one of the authentication solutions that are available (for example SPID) by using a high level of assurance (e.g., requiring a second factor authentication) and obtains a QR code.
- On her smartphone, patient downloads the OTP-PAT app using an official marketplace. Then, she scans the QR code using OTP-PAT app and enters a temporary code obtained on her email. If the login is successful, then the activation phase is completed by patient with the creation of a PIN code.

As an alternative, if some users do not want to activate the app online, they can go to one of the office of the healthcare organization, prove their identity by using an identity card (in-person identification) and finally they receive a printed version of the QR code. Note that this solution avoids the need to manually enter username and password since the identity information of the patient is inside the QR code obtained after a strong authentication (in-person or online).

Evaluation: The participants were able to propose valid alternatives.

5 Lesson Learned and On-Going Work

During this workshop, we had the opportunity to discuss our work with researches from technical and legal backgrounds. On the one hand, our main goal was to create awareness of usability and security issues related to authentication in the mobile context. On the other hand, the information extracted by their answers and discussions gave us the opportunity to validate our hypothesis on the usability and security of the solution that we have proposed for the TreC scenario.

Regarding the exercise session, we observed a high interest and participation. The participants were able to answer in the correct way to almost all questions. Only the first question related to the possibility to re-use a browser-based solution in the mobile context was not fully understood. The problem was that—being us security experts—we intended that question from a security perspective, while in a broader context this was unclear.

In the workshop, we made an effort to clearly define the security and usability problems in each one of the proposed exercises and questions. Indeed, this is rarely the case in a real-world scenario whereby striking the best possible balance between security and usability turns out to be a daunting task. This is so because of tight development schedules, focus on functionalities rather than security-by-design, and the unawareness of developers about the security implications of certain implementation decisions. The combined effect of these factors results in the presence of severe (and exploitable) vulnerabilities in a large amount of applications. To alleviate this state of affair, we developed a plug-in for the automated synthesis of secure authentication solutions in mobile applications in the context of the EIT Digital activity "Security Tools for App Development" [1] (STAnD). The basic idea underlying STAnD is to provide native app developers with tools that help them to take into consideration more and more security aspects. STAnD will be composed by a plugin for code hardening and a wizard that allows developers to configure and customize their solutions: developers are presented with a series of choices and then the code is automatically produced according to their answers. A second feature of STAnD will be the possibility to validate the app code by submitting the APK to a managed service that will check if there are security problems (e.g., the need for obfuscating the code related to the handling of a key).

Acknowledgments. This work has partially been supported by the Activity no. 18163, "API Assistant - Automated security assessment of 3rd party apps for the API economy", funded by the EIT Digital.

References

1. API Assistant: automated security assessment of 3rd party apps for the API economy. https://st.fbk.eu/projects/api-assistant/
2. AT&T Sued Over $24 Million Cryptocurrency SIM Hijack Attacks. https://www.databreachtoday.com/att-sued-over-24-million-cryptocurrency-sim-hijack-attacks-a-11365
3. DPCM of 24 October 2014, (SPID). http://www.agid.gov.it/agenda-digitale/infrastrutture-architetture/spid
4. Facebook: Getting started with the Facebook SDK for Android, May 2017. https://developers.facebook.com/docs/android/getting-started/facebook-sdk-for-android/
5. FIDO. https://fidoalliance.org/about/what-is-fido/
6. NIST Special Publication 800–63B: Appendix A - Strength of Memorized Secrets. https://pages.nist.gov/800-63-3/sp800-63b.html#appendix-astrength-of-memorized-secrets
7. NIST Special Publication 800–63B: Section 8.1: Authenticator Threats. https://pages.nist.gov/800-63-3/sp800-63b.html#81-authenticator-threats
8. Profiles for the OASIS: Security Assertion Markup language (SAML) V2.0. http://docs.oasis-open.org/security/saml/v2.0/saml-profiles-2.0-os.pdf
9. Telesign Consumer Account Security Report. https://www.telesign.com/resources/research-and-reports/telesign-consumer-account-security-report

10. We had a security incident. Here's what you need to know. https://www.reddit.com/r/announcements/comments/93qnm5/we_had_a_security_incident_heres_what_you_need_to/
11. Worst passwords of 2017 - Top 100. https://s13639.pcdn.co/wp-content/uploads/2017/12/Top-100-Worst-Passwords-of-2017a.pdf
12. CAD: Codice dell'Amministrazione Digitale - D.Lgs.n. 82/2005 (2014). http://www.altalex.com/documents/codici-altalex/2014/06/20/codice-dell-amministrazione-digitale
13. Chen, E., Pei, Y., Chen, S., Tian, Y., Kotcher, R., Tague, P.: OAuth demystified for mobile application developers. In: Proceedings of the ACM Conference on Computer and Communications Security (CCS) (2014). https://doi.org/10.1145/2660267.2660323
14. Cranor, L.: Your mobile phone account could be hijacked by an identity thief. https://www.ftc.gov/news-events/blogs/techftc/2016/06/your-mobile-phone-account-could-be-hijacked-identity-thief
15. European Parliament: eIDAS. http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32014R0910&from=EN
16. Garante Privacy: Personal Data Protection Code. Legislative Decree no. 196 of 30 June 2003 (2003). http://www.privacy.it/archivio/privacycode-en.html
17. General Data Protection Regulation: Regulation EU 2016/679. http://www.eugdpr.org
18. OAuth Working Group: OAuth 2.0 for Native Apps (2018). https://tools.ietf.org/html/rfc8252
19. Luo, T., Hao, H., Du, W., Wang, Y., Yin, H.: Attacks on WebView in the android system. In: Twenty-Seventh Annual Computer Security Applications Conference, ACSAC 2011, Orlando, FL, USA, 5–9 December 2011, pp. 343–352 (2011). https://doi.org/10.1145/2076732.2076781
20. OpenID Foundation: OpenID Connect Core 1.0. (2014). http://openid.net/specs/openid-connect-core-1_0.html
21. Sciarretta, G., Carbone, R., Ranise, S., Viganò, L.: Design, formal specification and analysis of multi-factor authentication solutions with a single sign-on experience. In: Proceedings of the 7th International Conference on Principles of Security and Trust (POST), pp. 188–213 (2018). https://doi.org/10.1007/978-3-319-89722-6_8
22. Shehab, M., Mohsen, F.: Towards enhancing the security of OAuth implementations in smart phones. In: IEEE International Conference on Mobile Services (MS), pp. 39–46 (2014). https://doi.org/10.1109/MobServ.2014.15

Towards Empowering the Human for Privacy Online

Kovila P. L. Coopamootoo(✉)

Newcastle University, Newcastle upon Tyne, UK
kovila.coopamootoo@newcastle.ac.uk

Abstract. While it is often claimed that users are more and more empowered via online technologies [4,16,17,31], the counterpart of privacy *dis*-empowerment is more than a suspicion [27]. From a human-computer interaction perspective, the following have previously been observed (1) users still fail to use privacy technologies on a large scale; (2) a number of human-computer interaction mismatches exist that impact the use of privacy technologies [1,5,6,15,22,32]; and (3) the user affect dimension of privacy is fear focused [14].

This paper reports on a experts' perspectives on empowering users towards privacy. We facilitated a workshop with $N = 12$ inter-disciplinary privacy experts to gather opinions and discuss empowering case-studies We reviewed literature focusing on the empowering versus dis-empowering impact of online technologies, and looked into psychological empowerment and usable privacy research.

The workshop participants pointed to a state of privacy *dis*-empowerment online, with human-computer interaction and business models as major themes. While it was clear that there is no clear-cut solution, supporting clearer communication channels was key to experts' mental models of empowered privacy online. They recommended enabling user understanding, not only for privacy threats but also in using privacy technologies and building user skills. To facilitate user interaction with complex secure communication tools, they suggested a transparency enhancing tool as a bridge between the user and encryption technology.

The outcome of the workshop and our review support the need for an approach that enables the human user, as well as their interactions and their active participation. For that we postulate the application of psychological empowerment [33]. To our knowledge, this paper provides the first known discussion among inter-disciplinary privacy experts on the topic of privacy *dis*-empowerment online, as well as the first categorisation of HCI mismatches impacting the use of privacy technologies.

1 Introduction

The internet is often seen as an empowering environment for consumers impacting personal, interpersonal, group and citizen-wide dynamics [4,16], and enabling consumer influence on product design, choice and decisions [17] and

Published by Springer Nature Switzerland AG 2019
E. Kosta et al. (Eds.): Privacy and Identity 2018, IFIP AICT 547, pp. 67–80, 2019.
https://doi.org/10.1007/978-3-030-16744-8_5

co-creation [31]. However, the indiscriminate amount of information collected for this purpose is also seen to come with privacy-, identity- and empowerment-related issues [25]. In literature, the potential for privacy empowerment has been linked with awareness of threats [26], confidence in behaviour [7], and perception of control on distribution and use of personal information [23].

While Human-Computer Interaction (HCI) research in the area of Privacy has seen great progress under the flagship of Usable Privacy, enabling large-scale use of privacy enhancing technologies (PETs) still remains a challenge. Given the need for an approach that caters for the individual, their HCI interaction and their engagement online, we postulate the need for discussions on the concept and practice of 'empowering users online' from an HCI perspective. So far empowerment literature has focused on the workplace and has mainly come from organisational research and with a psychology angle [29,33].

Workshop. We facilitated a workshop entitled "Empowering the Human for Privacy Online" at IFIP Identity & Privacy Management Summerschool 2018, in Vienna. The workshop was organised with a presentation and an interactive component.

Contributions. To our knowledge, we provide the first known inter-disciplinary discussion on the topic of privacy empowerment. Our workshop points towards themes of HCI and skills, business models and trust, and choice and legal contexts. We also categorise and summarise HCI barriers to adoption of PETs under mismatches that need research attention.

2 Unpacking *Dis*-Empowerment Online

Definition. The Oxford dictionary defines empowerment as the

> *"authority or power given to someone to do something"* which includes
>
> *"The process of becoming stronger and more confident, especially in controlling one's life and claiming one's rights".*

2.1 Empowerment via Online Technology

Empowerment via the internet has been referred to as *e-empowerment* or *consumer empowerment*. Amichai et al. [4] proposed a conceptualization of ways the internet is used as an empowering tool, coining *e-empowerment*, and referring to four levels, namely personal, interpersonal, group and citizenship, while Fuglsang [16] showed how IT and the internet can be used in social experiments to enable active citizenship for seniors.

For their part, Füller et al. [17] proposed the concept of consumer empowerment to describe consumers' perceived influence on product design and decision-making. They investigated perceived empowerment through internet-based co-creation activities. They observed that consumers engaging in co-creation felt more or less empowered, depending on the design of the virtual interaction tool, the related enjoyment, participants' task involvement as well as their creativity.

This is in line with Wathieu et al.'s suggestion that consumer empowerment is facilitated by consumers' ability to shape the composition of their choice set, where progress cues and information about other consumers are likely to enhance the experience [31]. Finally, Pires et al. argue that ICT is shifting market power from suppliers to consumers, with the ensuing consumer empowerment as unintended consequence of marketing [28].

2.2 Privacy Trade-Off

However, scholars have observed that mediated connections are more and more part of the infrastructure of people's lives in the internet age, where Pierson [27] argues that individuals' vulnerability is changing in relation to online consumer privacy when engaging with new network technologies, in particular those of mass self-communication. He posits greater external vulnerability for individuals induced by scalable systems, data replicability, persistence and searchability, and difficulties coping with internal vulnerability due to the increased complexity of the online environment.

In terms of trading privacy, O'Hara et al. discussed that while lifelogging, the indiscriminate collection of information concerning one's life and behaviour could be valuable in empowering the individual by providing a new locus for the construction of an online identity, it can also present some privacy, identity and empowerment-related issues [25].

In addition, literature also suggests that the potential for privacy empowerment includes awareness of threats [26], confidence in behaviour [7], and perception of control on distribution and use of personal information [23]. We find that perception of threats and risks [22] are often not accurate, confidence in behaviour is impacted by the fear dimension of privacy [8] and perception of control and use of personal information is often missing.

2.3 Why Privacy Empowerment?

Although privacy is implicit within human behaviour offline, in the online environment, it is mediated by technology and its human-computer interaction and thereby introduces a number of behavioural challenges not necessarily obvious and seamless to the human and to designers [9,12].

Bellotti & Sellen point to the problems of disembodiment (the actors are invisible in actions) and dissociation (actions are invisible to actors), both leading to visibility issues in privacy and security [5]. dePaula et al. examined the interaction problem of facilitating the understanding and effective use of PETs, by turning away from expression and enforcement and towards explication and engagement. These human-centred strategies towards enabling effective use of PETs, dubbed "user empowerment", have previously been raised by Wang and Kobsa [30], in particular with regards to empowering users in their privacy decisions.

3 Psychological Empowerment

We identify and review two main approaches of theorising about psychological empowerment in literature: (1) a cognitive model based on task assessments that impact intrinsic task motivation [29]; and (2) a nomological network including intrapersonal, interactional and behavioural components that also distinguish between empowerment processes and outcomes [33].

3.1 Cognitive Model

Thomas and Velthouse [29] defined *Psychological Empowerment* as increased intrinsic task motivation and proposed a theoretical model with four or cognitions or task assessments, namely *sense of impact*, *competence*, *meaningfulness*, and *choice* [29] that produce the motivation. The model captures individuals' interpretive processes via which they arrive at the task assessments. Psychological empowerment here focuses on intrinsic motivation and not on the managerial practices used to increase individuals' level of power.

Intrinsic Task Motivation involves positively valued experiences that individuals derive directly from a task. The core of the model therefore involves identifying these cognitions called task assessments (sense of impact, competence, meaningfulness, and choice [29]). These occur within the person and refer to the task itself, rather than the context of the task or rewards/punishments mediated by others. A task includes both **activities** and a **purpose**. The intrinsic value of goal or purpose accomplishment is produced by the articulation of a meaningful vision or mission.

3.2 Nomological Network

Zimmerman's nomological network extends the focus from intrapersonal aspects of the cognitive model to also include interactional and behavioural components [33].

In particular, the three components merge to form a picture of a person who believes that he or she has the capability to influence a context, understands how the system works in that context and engages in behaviours to exert control in that context. The intrapersonal aspect refers to how people think about themselves and includes domain-specific perceived control and self-efficacy, motivation to control, perceived competence, and mastery. The interactional aspect suggests that people are aware of their behavioural options, and includes critical awareness, understanding of causal agents, skill development, skill transfer and resource mobilisation. The behavioural aspect refers to actions taken to directly influence outcomes, including community involvement and participation and coping behaviours.

4 Usable Privacy

There has been roughly 19 years of research into approaches for aligning research in Human Computer Interaction with Computer Security, colloquially under *usable privacy and security.* This body of research was established to investigate the usability issues that explain why established security and privacy mechanisms are barely used in practice.

State-of-the-art research in usable privacy has mainly spread across (1) usability of website privacy policy, platform for privacy preferences, and behavioural advertising, (2) policy specification and interaction, (3) mobile privacy and security, and (4) social-media privacy [18].

4.1 Human-Computer Mismatches

A key goal of Usable Privacy research has been to bridge the gap between human users and technology [18], where a number of HCI challenges have been identified that act as obstacles to adoption of privacy technologies. We refer to them as mismatches between the Human and the Computer system. These include:

- user needs vs structure of tools [32];
- user mental models vs conceptual models of tools [1];
- user perceptions of system risks vs actual risks [22];
- visible disclosure vs invisible threats [5];
- visible security and privacy action vs invisible impact [15];
- skills needed vs actual user skills [6].

In addition, recent work found that perceived anonymity and trust were strong determinants of behavioural intentions and actual use behaviour [19]. Yet another challenge to the adoption of privacy technologies, that relates to usable privacy, is the distinction in the cognitive and affective components of privacy and sharing attitudes [14], such that human-computer designs that induces cognitive aspects of close connections and joy affect may be incongruent with privacy appraisal online and subsequent protective behaviour.

4.2 Assisting the User

There have been endeavours to address certain mismatches in specific context. For example, for invisible privacy threats, Ackerman and Cranor [2] proposed privacy critics, that are semi-autonomous agents that can monitor users' actions, warn them about privacy threats and suggest suitable countermeasures.

To address skills development, Brodies et al. proposed allowing users to create privacy policies suitable to their skills and background and to visualise the policies they have created [6].

To aid user comprehension and assessment of actions, dePaula et al. deliberately proposed dynamic real time visualisation of system state [15] for integration of configuration and action (aid flexible and effective control), for peer to peer file sharing application.

In addition, various strategies have been proposed, in particular those designed to counter decision-making hurdles [3].

4.3 Towards Empowering the User

The first step to successfully protecting one's privacy online is effectively using privacy technologies. We however note that privacy technologies have not yet reached large scale nor mainstream use.

We perceive that human factors of privacy research has progressed within distinct components of the nomological network of psychological empowerment, and identified or addressed the intrapersonal, interactional or behavioural components separately. In addition, while the mismatches point to the interactional aspects of the network only, we position that to enable and sustain effective use of PETs on a large scale, addressing one mismatch at a time does not ensure use of PETs. For example supporting user understanding of privacy threats only and not building skills in how to use the PET is futile, as threat appraisal may accentuate fear and helplessness with regards to privacy and impact protection motivation [8, 14].

We also postulate supporting users throughout their lifetime, that is, to not only investigate intrapersonal aspects of attitudes [14], concerns [21], affect [24], cognition [10–13] and confidence in competency [8], but also (1) to investigate how the HCI enables the evaluation of risks in interaction, the gathering of skills and resources, and the awareness of the impact other agents online, as well as (2) to promote users' active contributions privacy protection online.

5 Workshop

We conduct a workshop at the *IFIP Summerschool in Privacy and Identity Management 2018*, in Vienna. The workshop was facilitated with both a presentation component and an interactive component.

5.1 Aim

To enable a discussion on privacy *(dis)*-empowerment online and how psychological empowerment may enable the use of privacy technologies.

5.2 Workshop Method

Procedure We first elicited participants' awareness of usable privacy, of supporting and enabling the user for privacy online, as well as their opinions on the state of *dis*-empowerment online. In particular we asked:

- What are your privacy research area/interests?
- What does Usable Privacy mean to you?
- What do you already know of ways to support and enable the user for privacy online?
- Do you think online users are currently empowered or dis-empowered wrt privacy? Why is this the case?
- What would empowered privacy online look like?

Second, we gave a presentation covering the history of usable privacy research and a review of state-of-the-art research. We introduced the concept of empowerment and made a case for privacy dis-empowerment online via literature and previous studies. We summarised the human-computer mismatches identified within Usable Privacy research.

Third we facilitated a discussion on ways to empower users online.

Fourth, we facilitated a longer discussion, with participants divided into 2 groups of 6 participants. The topics of the group exercise and discussion included to

- select a context between either the social web with transparency enhancing technologies (TETs) or anonymous communications which can use more traditional privacy enhancing technologies (PETs) or a combination of PETs and TETs;
- select PETs and TETs applicable to chosen context based on a list provided and add others if needed;
- discuss how an empowered privacy preserving and privacy enabling human-computer interaction may look like, and what are the environmental requirements;
- use the psychological empowerment models presented to create requirements, and identify empowered processes and outcomes.

Participants. $N = 12$ participants joined the workshop, with a mix of male and female PhD researchers and academics. We did not elicit demographic data, as it is not relevant to the workshop aim. Instead we asked participants about their privacy research interests and their opinions on usable privacy and privacy *dis*-empowerment online (as described above).

Participants' research area and/or interests were spanned follows: we had 2 participants with connections to fairness (fairness in general and in relation to machine learning); 2 participants connected to privacy decision-making and risks; 3 participants either connected to the legal aspects of privacy or mapping legal requirements to human-computer requirements or into mechanisms implementation; 2 participants related to IoT and smarthome privacy; and the 3 others involved in areas of anonymity from a systems angle, trust building artificial intelligence and information security. These were gathered at the start of the workshop.

5.3 Opinions on Usable Privacy and Dis-/Empowerment Online

We provided a questionnaire at the beginning of the workshop with the questions as provided in Sect. 5.2 above. We report on the responses elicited which constitute participants' individual opinions. We refer to participant 1 as P1, participant 2 as P2 and so on till P12.

Usable Privacy Participants were queried on their perception of usable privacy.

According to 5 participants, usable privacy refers to *understandable methods of interacting with users.* They referred to communicating with users via ways that are not cognitively demanding (P2), enabling stakeholders to make decisions with full understanding of tradeoffs (P6), providing privacy technologies that are easy to understand and use (P7), as well as understandable data treatment and legal contexts (P11).

3 participants connected usable privacy with *ease of use.* These included P1 "interactive and easy to use", P7 "Privacy enhancing technologies (PETs) that are easy to use ... by users" and P8 "ease of navigation of settings and controlling these". In addition P10 referred to striking the best possible tradeoff between user experience and level of security.

2 participants referred to *efficiency* and/or *effectiveness* to explain usable privacy, with P1 "efficient way to exercise rights" and P9 "privacy that is effective and efficient for the user".

2 participants made a link between *the user and legal aspects*, with P5 "combine legal compliance with usability" and P11 "...explain legal contexts in a simpler way".

1 participant perceived usable privacy as related to *trust*, with P5 "making PETs trustworthy and deployable".

Ways to Support and Enable the User for Privacy online. We queried participants' awareness of ways to support and enable the user online.

3 participants responded with *human-computer interaction methods*, with P2 referring to "online nudges" and "feedback mechanisms", P7 referring to "transparency enhancing technologies can help users to understand the impact of using or not using PETs" and P5 "HCI challenges and requirements".

3 participants made a connection with *legal requirements of privacy*, with P2 "...providing opt-in choices (not opt-out)", P8 "currently most privacy is 'hidden' or non-existing although that is starting to change with GDPR" and P11 "GDPR and ... privacy protection code".

3 participants referred to *specific technological solutions*, such as P3 "VPN, proxy, TOR", P5 "PETs & TETs" and P10 "Auth protocols (e.g. SSO, OAuth, OpenID Connect) and their deployment in different technological scenarios (mobile, desktop, cloud), access control policies (specification and enforcement)".

In addition, 2 participants expressed a *lack of accessible methods*, with P9 "the existing ways are usually supporting and enabling users that are specially interested in their privacy and will put some considerate effort into maintaining it", or P8 "currently most privacy is 'hidden' or non-existing ... ".

Dis/Empowerment with Respect to Privacy Online. We asked participants if they felt users are currently empowered or dis-empowered with respect to their privacy online and to explain why they thought so. 11 participants responded while 1 did not provide an answer. The 11 pointed to users currently being dis-empowered with respect to their privacy online.

8 of the 11 participants pointed clearly towards dis-empowerment. 2 participants explained their verdict with a *lack of choice*, for example P11 said "...many times the user does not have a choice with respect to her privacy if she wants to use the online service". 2 participants pointed to the *complexity*, with P1 "...data processes are too complex" and P7 "they do/can not understand what people can do with lots of data". 3 participants supported their answer by referring to *business models*, with P2 "power dynamic favors business/state [who] favor profits", P3 "due to business benefits specially for large companies" and P10 "...due to business models ... and more in general companies [are] in the data monetization business". In addition, P6 explained linked dis-empowered online users with "mistrust and all its consequences".

3 of the 11 participants hinted towards *empowering possibilities*, with P5 explaining that "in theory there are many tools, but usability is indeed a problem", P9 making a distinction by the type of users, with "the lay users are currently feeling they have no power over their privacy anymore", and P12 hinting to a potential change in the future, with "generally dis-empowered maybe better with GDPR?".

Mental Models of Empowered Privacy Online. We elicited participants' mental model of empowered privacy online by asking them to describe how empowered privacy may look like.

6 participants' mental models included aspects of human-computer interaction, with 4 pointing towards clarity of communication and understandability, 2 pointing to intuitiveness, 1 pointing to having a choice to react and 2 towards control.

For *clarity of communication and understandability*, P1 depicted empowered privacy online with "provide adequate ... in clear and understandable manner", P2 referred to "clear un-conditional controls, clear online communication", P7 offered "users understand problems with respect to online privacy and have a choice, can re-act" and P3 "I can decide what data of mine to be shared in a more intuitive and straight-forward way".

For *intuitiveness*, P5 listed "intuitive, privacy by default, adapting to user needs" and P3 responded as above.

Choice was mentioned by P7 as above, while *control* was elicited from P2's response as above and P8 "ease of navigation of settings and controlling these".

In addition, 5 could not provide a response with P4, P9 and P11 not responding and P10 expressing "hard question, difficult to answer because of many contradicting interests from stakeholders" and P12 "not sure".

5.4 Empowered Privacy Human-Computer Interaction

We facilitated a 20 min group discussion, where participants gathered in two groups of 6 each. They were guided by the questions in the fifth part of the procedure described in Sect. 5.2. We report on the two empowering design solutions created by the two groups. Group 1 selected an anonymous communications case-study whereas Group 2 selected a social web case-study.

Anonymous Communication. Group 1 participants chose a mix of privacy and transparency enhancing technologies to facilitate anonymous communications, in particular AN.ON and Privacy Score. AN.ON[1] is an anonymity service that ensures anonymity/un-trace-ability of the users machine via a series of intermediaries on a network of mixes. "Instead of connecting directly to a webserver, users take a detour, connecting with encryption through several intermediaries, so-called Mixes. The sequence of linked mixes is called a Mix Cascade. Users can choose between different mix cascades. Since many users use these intermediaries at the same time, the internet connection of any one single user is hidden among the connections of all the other users. No one, not anyone from outside, not any of the other users, not even the provider of the intermediary service can determine which connection belongs to which user".

PrivacyScore[2] is a browser add-on that acts as an automated website scanner to investigate websites for security and privacy issues [20]. The beta version reports on whether the website is tracking the user or allowing others to do so, whether the webserver offers HTTPS connections and the security of their configurations, whether the website has obvious security flaws, and whether the mail servers of the website support state-of-the-art encryption.

Participants discussed their choice for anonymous communication to target users who "care about their privacy", and for users who "trust the tools [sic] and want to pay for them [sic]". They observed that with AN.ON, human-computer interaction is a concern, in particular for users on the network and that currently the service is for "users who [sic] understand privacy". They therefore proposed PrivacyScore as TET to enhance users' trust in efficacy of AN.ON and facilitate adoption. They proposed "Privacy by Default" because they assessed that AN.ON was "are quite technical". Hence having PrivacyScore as TET would "show that a privacy-enhancing technology is working", and support average skills users. They also added that the anonymous communication setup of AN.ON with PrivacyScore may not work (another word) in all environments such for an employee in a work context, therefore highlighting the question of how individuals' right to privacy apply in a work context.

Social Web. Group 2 participants chose Google Dashboard[3] as transparency enhancing technology (TET) for the social web. The TET supports awareness of data collection by allowing users to see some of the personal data that Google stores about them, linking to settings where users can influence the storage and visibility of data. However, note that with Google logging all user activities, user behaviour prediction and manipulation is a privacy issue while hacking is a cyber security risk with potentially huge impact.

As empowered privacy HCI, participants explained the need for users to be informed. They mentioned (a) awareness of policies, (b) sign up for data uses

[1] https://anon.inf.tu-dresden.de/index_en.html.
[2] https://privacyscore.org/.
[3] https://myaccount.google.com/dashboard.

information, (c) data portability: move from provider to provider, (d) what's going on: who's looking at their data, (e) wider implications of data storage.

They also spoke about choice on data use by others.

6 Discussion

The outcomes of the workshop depict the complexity of the problem of enabling effective use of PETs on a large scale.

6.1 HCI and Skills

Workshop participants expressed the importance of the human-computer interaction in meeting the users, such as through clarity of communication and understandability to aid decisions (P1, P2, P3), feedback mechanisms (P2) for visible impact and ease of use of PETs (P1, P7, P8). In addition, complexity of interaction with PETs were also explicitly pointed out (P2, P7) as well as the current need of specialist skills and effort if one were to use PETs (P9). As example, the in the case-study discussion, Group 1 participants offered facilitating interaction between users and anonymous communication technology via TETs.

Workshop participants stressed on the need for more user understanding throughout the workshop, well beyond just understanding the threats of data usage (P1, P7) but also understanding how to use PETs and ease of use (P7), as well as understanding of legal contexts (P11).

These link directly with the incongruency observed via the HCI mismatches of Sect. 4.1. While the set of HCI mismatches point to the hard problem of enabling privacy online and sustaining use of privacy technologies, we postulate that the HCI complexity and enabling the individual user can be met by the three components of the nomological network of psychological empowerment [33].

6.2 Choice and Legal Requirements

While participants pointed to a (perceived) lack of choice with regards to privacy if users wanted to use online services (P11), the skills challenge, and poor awareness of resources can also be thought to contribute to a lack of choice. In addition, the intrapersonal aspect of psychological empowerment of confidence in competency or skills also constitute an impact on privacy motivation [8] and therefore a lack of choice.

Participants also postulated that enabling and supporting user privacy online requires provision of legal requirements, where there was hope that there would be more visible privacy solutions with the GDPR (P8, P9).

6.3 Business Models and Mistrust

Similar to previous research pointing to a general culture of fear with regards to privacy online, mainly associated with mistrusting the actions of businesses and governments [14], workshop participants pointed to power dynamics favoring state and business profits (P2), benefits for large companies (P3) and business models based on data monetisation (P10) and mistrust (P6).

The power balance could be shifted if PETs design addressed the HCI mismatches for example, via better congruency with user needs, skills development, better perception of risks. On their part, participants postulated the need to make PETs more trustworthy and deployable (P5), as well as giving users a share of the profits (P6).

7 Conclusion

While it was clear from workshop participants that there currently exists a problem of privacy *dis*-empowerment online, there were no obvious solution on how to tackle the problem. There was a clear reliance on bridging the gap between users and complex privacy technologies via enhanced human-computer interaction. While power dynamics emerging from business models of data monetisation was found to be a major hassle, there was a sense of hope with the advent of the GDPR.

By enabling a discussion among inter-disciplinary privacy experts on the topic of privacy *dis*-empowerment online, as well as providing the first categorisation summary of HCI mismatches impacting the use of privacy technologies, this paper highlights avenues for future investigations in the area of human factors of privacy. For instance, investigating aspects of HCI that promote privacy empowerment as detailed in this paper and finding ways to promote large scale adoption of privacy technologies.

References

1. Abu-Salma, R., Sasse, M.A., Bonneau, J., Danilova, A., Naiakshina, A., Smith, M.: Obstacles to the adoption of secure communication tools. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 137–153. IEEE (2017)
2. Ackerman, M.S., Cranor, L.F., Reagle, J.: Privacy in e-commerce: examining user scenarios and privacy preferences. In: Proceedings of the 1st ACM Conference on Electronic Commerce, pp. 1–8. ACM (1999)
3. Acquisti, A., et al.: Nudges for privacy and security: understanding and assisting users' choices online (2017)
4. Amichai-Hamburger, Y., McKenna, K.Y., Tal, S.-A.: E-empowerment: empowerment by the internet. Comput. Hum. Behav. **24**(5), 1776–1789 (2008)
5. Bellotti, V., Sellen, A.: Design for privacy in ubiquitous computing environments. In: de Michelis, G., Simone, C., Schmidt, K. (eds.) ECSCW 1993, pp. 77–92. Springer, Dordrecht (1993). https://doi.org/10.1007/978-94-011-2094-4_6
6. Brodie, C., Karat, C.-M., Karat, J., Feng, J.: Usable security and privacy: a case study of developing privacy management tools. In: Proceedings of the 2005 Symposium on Usable Privacy and Security, pp. 35–43. ACM (2005)
7. Church, L., Anderson, J., Bonneau, J., Stajano, F.: Privacy stories: confidence in privacy behaviors through end user programming. In: SOUPS (2009)

8. Coopamootoo, K.P.: Work in progress: fearful users' privacy intentions - an empirical investigation. In: 7th International Workshop on Socio-Technical Aspects in Security and Trust. ACM, New York (2017)
9. Coopamootoo, P.L., Ashenden, D.: Designing usable online privacy mechanisms: what can we learn from real world behaviour? In: Fischer-Hübner, S., Duquenoy, P., Hansen, M., Leenes, R., Zhang, G. (eds.) Privacy and Identity 2010. IAICT, vol. 352, pp. 311–324. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-20769-3_25
10. Coopamootoo, K.P., Groß, T.: Mental models: an approach to identify privacy concern and behavior. In: SOUPS 2014 Workshop on Privacy Personas and Segmentation (2014)
11. Coopamootoo, K.P., Groß, T.: Cognitive effort in privacy decision-making vs. 3 × 4: evaluation of a pilot experiment design. In: LASER 2014 Workshop (2014)
12. Coopamootoo, K.P.L., Groß, T.: Mental models for usable privacy: a position paper. In: Tryfonas, T., Askoxylakis, I. (eds.) HAS 2014. LNCS, vol. 8533, pp. 410–421. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07620-1_36
13. Coopamootoo, K.P., Groß, T.: Mental models of online privacy: structural properties and cognitive maps. In: British HCI 2014 (2014)
14. Coopamootoo, K.P., Groß, T.: Why privacy is all but forgotten - an empirical study of privacy and sharing attitude. Proc. Priv. Enhanc. Technol. **4**, 39–60 (2017)
15. De Paula, R., et al.: In the eye of the beholder: a visualization-based approach to information system security. Int. J. Hum.-Comput. Stud. **63**(1–2), 5–24 (2005)
16. Fuglsang, L.: It and senior citizens: using the internet for empowering active citizenship. Sci. Technol. Hum. Values **30**(4), 468–495 (2005)
17. Füller, J., Mühlbacher, H., Matzler, K., Jawecki, G.: Consumer empowerment through internet-based co-creation. J. Manag. Inf. Syst. **26**(3), 71–102 (2009)
18. Garfinkel, S., Lipford, H.R.: Usable security: history, themes, and challenges. Synth. Lect. Inf. Secur. Priv. Trust **5**(2), 1–124 (2014)
19. Harborth, D., Pape, S.: Examining technology use factors of privacy-enhancing technologies: the role of perceived anonymity and trust (2018)
20. Maass, M., Wichmann, P., Pridöhl, H., Herrmann, D.: PrivacyScore: improving privacy and security via crowd-sourced benchmarks of websites. In: Schweighofer, E., Leitold, H., Mitrakas, A., Rannenberg, K. (eds.) APF 2017. LNCS, vol. 10518, pp. 178–191. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67280-9_10
21. Malhotra, N.K., Kim, S.S., Agarwal, J.: Internet users' information privacy concerns (IUIPC): the construct, the scale, and a causal model. Inf. Syst. Res. **15**(4), 336–355 (2004)
22. Mehrnezhad, M., Toreini, E., Shahandashti, S.F., Hao, F.: Stealing PINs via mobile sensors: actual risk versus user perception. Int. J. Inf. Secur. **17**(3), 291–313 (2018)
23. Midha, V.: Impact of consumer empowerment on online trust: an examination across genders. Decis. Support Syst. **54**(1), 198–205 (2012)
24. Nwadike, U., Groß, T., Coopamootoo, K.P.L.: Evaluating users' affect states: towards a study on privacy concerns. In: Lehmann, A., Whitehouse, D., Fischer-Hübner, S., Fritsch, L., Raab, C. (eds.) Privacy and Identity 2016. IAICT, vol. 498, pp. 248–262. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-55783-0_17
25. O'Hara, K., Tuffield, M.M., Shadbolt, N.: Lifelogging: privacy and empowerment with memories for life. Identity Inf. Soc. **1**(1), 155–172 (2008)
26. Olivero, N., Lunt, P.: Privacy versus willingness to disclose in e-commerce exchanges: the effect of risk awareness on the relative role of trust and control. J. Econ. Psychol. **25**(2), 243–262 (2004)

27. Pierson, J.: Online privacy in social media: a conceptual exploration of empowerment and vulnerability (2012)
28. Pires, G.D., Stanton, J., Rita, P.: The internet, consumer empowerment and marketing strategies. Eur. J. Mark. **40**(9/10), 936–949 (2006)
29. Thomas, K.W., Velthouse, B.A.: Cognitive elements of empowerment: an interpretive model of intrinsic task motivation. Acad. Manag. Rev. **15**(4), 666–681 (1990)
30. Wang, Y., Kobsa, A.: Privacy-enhancing technologies. In: Handbook of Research on Social and Organizational Liabilities in Information Security, pp. 203–227. IGI Global (2009)
31. Wathieu, L., et al.: Consumer control and empowerment: a primer. Mark. Lett. **13**(3), 297–305 (2002)
32. Whitten, A., Tygar, J.D.: Why Johnny can't encrypt: a usability evaluation of PGP 5.0. In: USENIX Security Symposium, vol. 348 (1999)
33. Zimmerman, M.A.: Psychological empowerment: issues and illustrations. Am. J. Community Psychol. **23**(5), 581–599 (1995)

Trust and Distrust: On Sense and Nonsense in Big Data

Stefan Rass[1(✉)], Andreas Schorn[1], and Florian Skopik[2]

[1] System Security Group, Institute of Applied Informatics,
Universitaet Klagenfurt, Universitaetsstrasse 65-67, 9020 Klagenfurt, Austria
{stefan.rass,andreas.schorn}@aau.at
[2] Center for Digital Safety and Security, Austrian Institute of Technology,
Giefinggasse 4, 1210 Vienna, Austria
florian.skopik@ait.ac.at

Abstract. Big data is an appealing source and often perceived to bear all sorts of hidden information. Filtering out the gemstones of information besides the rubbish that is equally easy to "deduce" is, however, a nontrivial issue. This position paper will open with the motivating problem of risk estimation for an enterprise, using big data. Our illustrative context here is the synERGY project ("security for cyber-physical value networks Exploiting smaRt Grid sYstems"), which serves as a case study to show the (unexplored) potential, application and difficulties of using big data in practice. The paper first goes into a list of a few general do's and don'ts about data analytics, and then digs deeper into (semi-) automated risk evaluation via a statistical trust model. Ideally, the trust and hence risk assessment should be interpretable, justified, up-to-date and comprehensible in order to provide a maximum level of information with minimal additional manual effort. The ultimate goal of projects like synERGY is to establish trust in a system, based on observed behavior and its resilience to anomalies. This calls for a distinction of "normal" (in the sense of behavior under expected working conditions) from "abnormal" behavior, and trust *can* intuitively be understood as the (statistical) expectation of "normal" behavior.

Keywords: Big data · Trust · Statistics · Anomaly detection · Security · Reasoning

1 Introduction

Trust is a generally familiar, but not a clearly defined term in many contexts. In a simple yet intuitive understanding, trust is the expectation of "correct" behavior (of a system, a person, …). As such, it has some relation to security, since the latter is, in a way, also the assurance that certain requirements are met. Our concern in the following will be security systems. Like in social life, security systems gain trust through their reliable behavior, and lose it in the light of threats or incidents related to the system. To "measure" trust, it is thus necessary to recognize relevant incidents and threats and to find a way of evaluating the impact on the trust in the system. A decent trust model should use the information in a transparent form, so as to support accountability (i.e.,

Published by Springer Nature Switzerland AG 2019
E. Kosta et al. (Eds.): Privacy and Identity 2018, IFIP AICT 547, pp. 81–94, 2019.
https://doi.org/10.1007/978-3-030-16744-8_6

the clear identification of reasons for anomalies) and fairness (i.e., trust should not overproportionally depend on single types or sources of information). *Transparency* is thus hereafter understood as the trust model's artefacts to be explainable, justifiable and interpretability beyond being only the result of complex computations. Methods lacking this kind of transparency are hereafter called "black-box". The two kinds may not differ in their power, but only in the degree to which the results can be explained. For security, it may be enough that the system works as expected; however, when it comes to the aftermath of an incident, it may additionally become necessary to understand the reason why the security did not work as expected (which calls for explainability).

Our position paper opens with an example of a security system related to anomaly detection in energy grids. This then shall provide the context for the further discussion of a simple statistical model to quantify trust and to incorporate continuously incoming information about a system into the system's trust indicator. The aim is to calculate a (always current) confidence index from the history of observed system behavior. The evolution of this trust variable over time is then useful to warn about future risk situations arising from possible series of events that would destroy trust. Worst-case risk, equivalently trust, scenarios then correspond to the shortest sequence of events that makes the trust index drop below a certain threshold (of acceptable risk). One lesson taught by the model is that "fairness" in the sense of how information affects the trust is not necessarily naturally consistent with the human understanding of trust. The statistical trust model is indifferent between positive or negative experience; the trust would change by relatively equal magnitudes into either direction. Between humans, however, trust can be much harder to gain than to lose.

The second part of the paper focuses on the detection of incidents within the history. For this purpose, statistical approaches exist which can uncover an artificial manipulation of data (under suitable conditions). The consideration here lies on the possibility of an automated recognition of manipulations purely on the basis of numerical data and in particular without recourse to (human) domain expertise.

More accurate models of trust as a measure of resilience against or likelihood of abnormal behavior can be established if domain knowledge is available. The statistical toolbox therefor covers a wide spectrum of methods, a categorization of which we will look into in the third part of this tutorial devoted to our case study. Our focus will therein be on reporting practical issues and challenges to overcome when striving for statistical anomaly detection up to predictive analytics.

2 Practical Anomaly Detection by Example – the synERGY Project

The synERGY project [1] aims at constructing an anomaly detection system for energy distribution systems operators (DSOs), which usually maintain highly distributed systems unifying many heterogeneous information technology (IT) and operational technology (OT) components and serving a vast lot of customers. As such, many aspects of the setting are not only very similar to that of general clouds, but also necessarily target of attacks and subject of trust and reputational management.

synERGY is a system to maintain security and hence trust in an energy grid, and to this end, integrates three major components, which are:

1. A security incident and event management (SIEM) system (SecurityAdvisor [2])
2. Two anomaly detection modules, each of which is based on different techniques:
    a. ÆCID [3, 4], which is based on log data parsing and rule-based anomaly detection
    b. The incident detection system "Tuna" (developed by the University of Vienna), which is based on statistical analysis of network packet information.

The system architecture (see Fig. 1) centers on a broker component that collects information from all sources (sensors), and feeds this into the anomaly detection engines, whose results are then – over the same broker – delivered to the SIEM system, where the human operator is informed and supported with rich data in his decision making.

A particular feature of synERGY is the explicit account for cost-benefit tradeoffs in the collection of "big" data (the cost being the computational and human efforts to collect and process information, vs. the benefits of damage prevention by this). The placement of sensors to harvest the data will affect the overall system performance (cost) and must be made w.r.t. aspects of errors in statistical tests and the required amount of data for the analysis (benefits). Regarding the error types in statistical tests, neither false-positives nor false-negatives should occur too often, since they either lead to alert fatigue (hence missing out the alarm when things get really dangerous) or may require unrealistically large amounts of data to be collected, which may be technically or economically infeasible. This is yet another benefit of compound systems such as synERGY, where rule-based detection (that can work with small data) are combined with data driven models that require larger amounts of data (whenever they are available).

While anomaly detection in synERGY, as well as generally most intrusion detection systems, strongly rest on standard statistical tests, the combination of different anomaly detection systems such as in synERGY enables further tests such as the Newcomb-Benford (NB) law for testing data manipulations. Such tests may, however, not necessarily needed to improve anomaly detection itself, but can signalize manipulations of the detection system (bypassing all other standard technical precautions such as encryption, signature techniques, access control, etc.). The NB law most likely kicks in for data being compiled from a complex interplay of at least two sources. This is exactly what an intrusion detection system, and synERGY is one example, may do. Thus, stealthy attacks on the detection system itself can be tested for with the techniques given above. This potential appears yet unexplored and this work may stipulate studies in this direction. Section 3.2 will explain how data manipulation can be tested for. It follows a general discussion on trust quantification in Sect. 3, and a preparatory discussion on data preparation and statistical testing in Sect. 3.1, all of which would integrate in a system like synERGY.

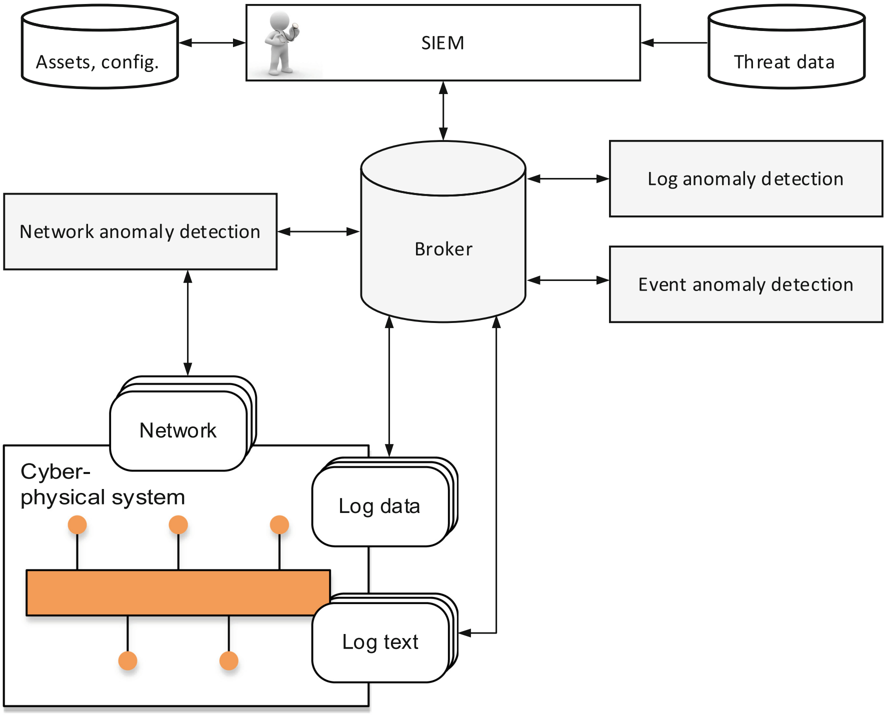


Fig. 1. synERGY architecture (simplified)

3 Quantifying Trust – the Beta Reputation Model

In the simplest setting, we may think of trust as an expectation of correct behavior based on a history of experience. If we note one for a good and zero for a bad experience in the past, the expectation of the so-constructed indicator variable will (in the limit) converge to the probability of a positive experience. The exact likelihood can then be taken as the trust in the event quantified by the indicator variable. Various services on the internet successfully use this scheme, as, e.g., Amazon's ratings on goods, eBay's ratings on sellers, and many other services measure the quality in such terms. A typical representation is on a scale from 1 to 5 "stars", mapping the unit interval [0, 1] linearly to the discrete set $\{1, \ldots, 5\}$, occasionally including the half integers therein (extending the scale to $\{1, 1.5, 2, 2.5, \ldots, 5\}$ with proper rounding to the nearest representative).

The interesting insight about this model is its statistical background, which is surprisingly rich and well-founded under mild hypotheses [5]. The first of these is stochastic independence of events. Let I be the indicator variable of the event in question, say, the adherence to a service level agreement (SLA) in a cloud, or other service. Furthermore, let the SLA be such that the customer can easily and reliably check whether the service provider (SP) has fulfilled its obligation according to the contract (for example, the files in the cloud are still consistently stored, the bandwidth for accessing the cloud is actually provided, or the billing is accurate and neither misses nor exceeds the actual consumption). In the general setting, let a user note $I = 1$ if the service performed satisfyingly, and note $I = 0$ otherwise. Many providers ask their customers for feedback, so as to provide a certificate of customer satisfaction to their prospect customers, so let us assume that every user u reports its individual indicator[1] I_u. Although a user u may indeed inform another user v about her/his personal experience with the SP, the users will generally act independently, and the only choice made upon other user's indicators is whether or not the service is used, but not the assessed quality of experience. This subtle difference is important, as it translates into stochastic independence of indicators I_u and I_v for any two (distinct) users u, v. Extending this view to a large collection of, say N, customers, the feedback to the SP is a set of i.i.d. Bernoulli random variables (r.v.) $\{I_1, \ldots, I_N\}$. The total number of happy customers is then a Poisson random variable with a rate parameter λ being the number of 1-values within the set of N restricted to the most recent unit of time (say, over the last month, 12 months, or similar). The trust value reported to a prospect customer is then the fraction $\frac{1}{N}\sum_{j=1}^{N} I_j$, i.e., the average over all ratings.

The most natural way of updating this trust value is conditioning on new incoming ratings, i.e., a *Bayesian* update. A convenient setup for this uses a Beta-distribution as prior, which is known to be conjugate to a Poissonian likelihood function [6], meaning that the posterior distribution will again be a Beta-distribution. The overall scheme is thus called *Beta-reputation* [5, 7], and roughly works as follows:

1. Initialize the system with a Beta prior distribution with density $f_\beta(x|a, b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$ for $x \in (0, 1)$ and zero otherwise, where B is Euler's Beta-function. The parameters $a, b > 0$ have a natural interpretation exposed by looking at the expectation of $X \sim \beta(a, b)$ r.v., which is $E(X) = \frac{b}{a+b}$. So, under a frequentistic view, b may count the number of positive experience, relative to the total number $a + b$ of events. Thus, if we let our trust variable be Beta-distributed, its first moment can be interpreted as a probability, exactly following the intuition that we developed above.
2. Upon a set $I_1, \ldots, I_k$ of incoming feedbacks, we can set up a likelihood function being a Poisson distribution. By conjugacy, the Bayes-update to the Beta-distribution $\beta(a, b)$ with a number n of negative feedbacks and m positive reports (i.e., $n = \left|\{j : I_j = 0\}\right|$ and $m = \left|\{j : I_j = 1\}\right|$), the posterior distribution is $\beta(a + n, b + m)$. So, the update is efficient and the trust value in turn becomes $E(X|I_1, \ldots, I_k) = \frac{b+m}{a+m+n}$, and remains aligned with our running intuition.

[1] From the perspective of psychology, this is admittedly an oversimplification of "experience" in assuming it to be binary (either "good" or "bad"). Nonetheless, we use this model here as a somewhat representative mechanism widely used in the internet; but without implying any claim on its psychological accuracy.

This procedure can be repeated as many times as we wish, and scales without ever running into issues of numeric integration or even having to represent the involved distributions explicitly at any point. Extensions are possible in various ways, such as:

- Accounts for reliability of updates: suppose that the information is uncertain, say, if the data item "$I_j = 1$" is actually the statement "$\Pr\big(I_j = 1\big) = p$" for some (known) certainty value $0 < p < 1$. That is, whether or not the experience is actually positive cannot be told for sure. How can we condition on such an uncertain event? One solution is *model averaging*, i.e., we create the posterior as a mix of two updates, one taking $I_j = 1$ with probability p and the other one assuming $I_j = 0$ with probability $1 - p$. The (new) posterior is then $p \cdot f_\beta(t|a, b+1) + (1-p) \cdot f_\beta(a+1, b)$.
  Since the updating is a linear operation, the procedure further repeats without essential changes, except for the mix of course to grow over many updates. It can be shown, however, that the growth is $O(n^2)$ for a total of n updates (independently of the values of p per update) [5].
  The confidence value p must be obtained from different sources, and usually is a "quality measure" of the feedback source itself. For machine learning algorithms, p can be a measure of accuracy. If the feedback is coming from a classifier (e.g., regression, support vector machine, or others), the palette of metrics (receiver operating characteristic, confusion matrices, and many more) can be used to compute values for p here.
- More fine-grained scales: as for Amazon or eBay, feedback can be given on a more fine-grained scale from 1 to 5 stars, or similar. This naturally integrates in the above procedure under a proper interpretation of the number of stars: instead of conditioning on a single feedback, say $I_k = 3$, we can condition on 3 feedbacks $I = 1$ instead. Likewise, assigning 5 stars to an experience can correspond to 5 positive feedbacks in the above scheme.
- Alternatively, we may also resort to more general distribution models integrating Binomial distributions (allowing for an integer range for the feedback) instead of the binary (Bernoulli) distributions as we had above. Conjugacy to the Beta distribution and hence efficiency of the updating process remains intact.
- Trust aggregation: in complex, especially technical, systems, trust in a component may not obviously translate into trust in the overall system. In security risk management, the maximum principle looks for the maximum risk among all relevant parts of a system, which becomes the risk assigned to the overall system. This method has a statistical counterpart that can be displayed in the above framework: a celebrated Theorem due to Abe Sklar tells that the joint distribution $F_{X_1,\ldots,X_n}$ of random variables $X_1, \ldots, X_n$ can be written in the form $F_{X_1,\ldots,X_n} = C(F_{X_1}, \ldots, F_{X_n})$, in which:
  - F_{X_i} for $i = 1, 2, \ldots, n$ are the marginal distributions of each r.v. (not necessarily independent of the others), and
  - $C : [0,1]^n \rightarrow [0,1]$ is a *copula function*, which – roughly speaking – is a multivariate distribution with all uniform marginals.

If we let $X_1, \ldots, X_n$ be β-distributed r.v. as constructed above, then the overall trust in the system is again another (not necessarily β-distributed) r.v., whose distribution can be compiled from the trust distributions per component upon knowing the copula function C. This function embodies the mutual dependencies between the components and separates the dependency model from the individual trust models. Its choice is thus usually influenced by domain knowledge, but independently of it, every copula satisfies the upper Fréchet-Hoeffding bound $C(x_1, \ldots, x_n) \leq \min\{x_1, \ldots, x_n\}$, where the min-operator is itself a copula function. This bound is just the maximum principle of IT security management: it just says that the overall trust in the system is determined by the least trust in any of its parts (equivalently, the "chain is only as strong as its weakest element"). Taking $C = \min$ is thus a valid worst-case and hence default choice in absence of better, more detailed, knowledge of the system components interplay towards trust.

The above considerations justify the β-reputation as a model of trust, but it may fail to reliably reflect the human understanding of trust, which is generally asymmetric. In brief, humans may lose trust much faster than they gain it. The model above, however, is symmetric, in the sense that positive and negative feedback go into the model with equal importance. While this is certainly fair, such fairness is not necessarily an accurate approximation of subjective trust perception. In addition, the model bears some "inertia", in the sense that changes to the trust value will eventually become smaller the more updates are done to the model. Equivalently said, the model will eventually become more and more stable, as the updates carry the model to convergence. This is yet another contrast to human trust treatment, since the pessimist may lose trust entirely upon a single negative experience.

On the positive side, this is a whitebox model, designed for ease of understanding. Speaking about usability, a trust measure that appears opaque to people and is as such not itself "trusted" may be less preferable than a simpler model that whose mechanisms are easier to follow (similarly to how open source software is often perceived as trustworthy, because it has no hidden or invisible parts). This puts it in contrast to more sophisticated yet partly black-box methods of aggregation, neural networks being a typical example where flexibility and power is traded for a complex input-output relation that does not necessarily align with human reasoning (and for that reason, however, may be more powerful indeed).

In any case, trust is a subjective measure, and the objectiveness suggested by the above model, despite its statistical underpinning, remains subjective too. Assessments from which trust values are computed may rely on assumptions such as the belief in cryptographic protections [8] (noting that asymmetric cryptography crucially rests on computational intractability, which in many cases has strong empirical support yet lacks mathematical proofs). Many practical difficulties of modern cryptographic security relate to complex matters of key management, and the complexity of such systems themselves. Although powerful and highly sophisticated cryptographic mechanisms could be used, the degree of (subjective) trust in them is a matter outside analytical provability. More importantly, the overwhelming success of cryptography in achieving its goals has moved it mostly outside the focus of contemporary attackers, spending the majority of effort on more "economic" attack strategies like social engineering.

3.1 Data as a Basis of (Dis)Trust

All this adds to intrinsic subjectivity of trust, not the least so since humans often remain the weakest element in any cyber-physical system. Nonetheless, humans (as providers of domain expertise) as well as computer systems (as providers of data and data analytics) remain indispensable sources of information and big data to base security and trust upon.

The challenge is the separation of sense from nonsense in such big data (including among others, say, information exchanged through blogs, personal communication, and other social channels), which usually calls for both, machine and human intelligence. The effectiveness of this mix depends on the aforementioned matters of understandability and technical support, starting data preparation first. Throughout the rest of this article, we stress that our concern is not judging the quality of the data itself, but rather the quality of what we conclude from it. Big data is not only a matter of getting many records; missing data and incomplete data records may severely reduce the "bigness" of the data. Moreover, it is important to know what we are looking for *before* looking into the data (more concretely, hypotheses need to be formulated *before* the data collection; the converse approach of having data and then looking into what can be learned from it can be the first step towards data dredging).

Dealing with missing data is an involved matter, and can be done in three basic ways:

(1) Amputation: simply discard all records that are incomplete; this, however, can severely cut down the available data (making it no longer "big" perhaps).
(2) Imputation: fill the gaps with data inferred from the remaining data. This can be done in several ways again, but filling the gaps with information obtained from the rest of the data set, apparently, cannot add any new information to the data. Thus, the information deficiency remains, yet only "disguised" to some extent.
(3) Treating missing data as a category of its own. This may yield conclusions from the fact that data is absent. However, logical deductions from the absence of facts must be made with care.

There is no general rule on what to do with missing data, and each of the above methods has its areas of success and cases of fail, often strongly dependent on whether or not the gaps occur systematic or at random. Ultimately, it thus remains a matter of domain knowledge and careful model analysis and validation, which of the three basic methods above (or another one) is most suitable. A similar related challenge is outlier elimination, which we leave out of our scope here.

A reasonable trust management will have the bulk of information processed by algorithms (machine intelligence), leaving ultimate decisions and alert handling to a human expert. The system will thus ask the human operator for invention upon certain signals recognized in the pool of available information (*anomaly detection*), and in designing such a system, it is useful to distinguish weak from strong signals, and to understand the meaning of a signal. Table 1 provides a selection of statistical tools with remarks on individual pros and cons. In the following, we confine ourselves to a necessarily non-exhaustive selection of methods, whose main purpose is highlighting potential difficulties as a guidance for selection, which includes:

- Pearson Correlation: this is a popular method of drawing indications of statistical similarity, dependence or other relations. While easy to apply and to interpret, correlation must be treated with care for several reasons:
  - It measures only linear dependencies between variables, ignoring possible nonlinear dependencies. For example, the variable X and $Y = X^2$ are clearly dependent, but have zero correlation if $X \in \{-1, 0, +1\}$. Concluding about independence from low correlation is thus incorrect.
  - "High" correlation may point toward some stochastic dependence, but neither causality nor functional dependence. Most striking examples are found in [9], such as, for example, the apparently high correlation of ≈ 0.9471, between the "per capita cheese consumption" and the "number of deaths by bedsheet tangling" (Fig. 2), whereas an implication or causality between the two seems clearly absurd.

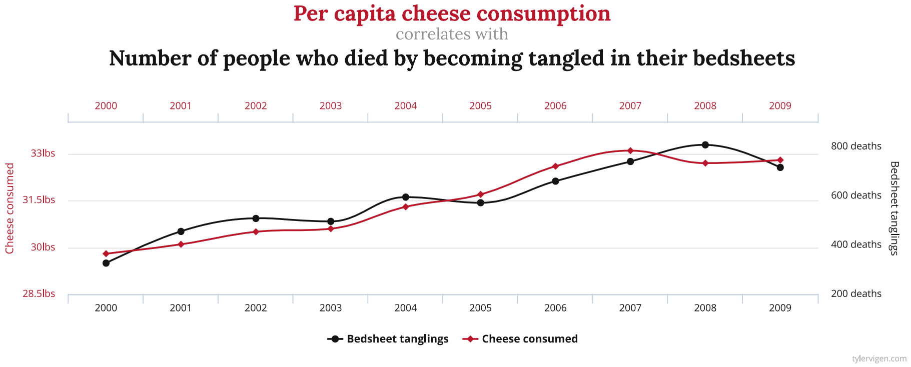


Fig. 2. Apparent dependence absurdly indicated by correlation [9]

- Statistical tests: These empirically refute an a priori hypothesis based on existing data. They cannot prove a hypothesis, nor is it correct to form a posterior hypothesis based on the data at hand. Inserting numbers into some formula to verify its correctness is far from being a mathematical proof. However, if the formula is incorrect on a given set of numbers, those numbers make an valid counterexample. It is the same story with statistical tests: the data can be consistent with the test's hypothesis, but this may be a coincidence. However, when the data is inconsistent with the hypothesis, the data is clearly a counterexample.

  Every (classical) statistical test thus runs along these lines of thinking: suppose that the claim to be verified is a statement A.

a. Formulate a null-hypothesis by negating A; let us – in a slight abuse of notation – call the respective opposite claim $\neg A$. The test will be designed to reject $\neg A$ so that the alternative hypothesis, statement A, will be assumed (based on the data).

b. Define a test statistic as some value that:
(1) Is easy to compute from the data,

(2) And has a known probability distribution $F(\cdot|\neg A)$ under assumption $\neg A$, i.e., your null-hypothesis.

c. Given concrete data D, compute the test statistic t_D, and check whether it falls into a certain range of acceptance for the test. This range is typically set as $(1-\alpha)$-quantile of the distribution of the test statistic's distribution $F(\cdot|\neg A)$. A popular value to tip the scale between acceptance and rejection of the null-hypothesis is based on the p-value, being the $p = \Pr(X > t|\neg A)$, i.e., the area under the curve $F(x|\neg A)$ in the interval (t_D, ∞). The null-hypothesis is rejected if $p < 1-\alpha$, when α is the statistical significance level (usually 95% or something similar).

The dangers of tests applied to big data are thus manifold, and at least include the following sources of error:

- Hypothesis that are formed not a priori, i.e., one seeks to "learn from the data whatever we can learn from it". The simple truth is: whatever you seek to learn, you will most likely be able to learn it from big data (as much as a conspiracy theoretician will always successfully find secret codes in the bible, or recognize alien landing sites on aerial photos of a landscape).
- Incorrect conclusions from the test's results: even if the test rejects the null hypothesis, its statistical significance cannot be taken as an error probability, say, in the sense of the β-reputation as we had above. The usual way of setting up a test is towards controlling the error of first kind, which is the chance of accidentally rejecting the null-hypothesis (although the assumption was correct). This error is complementary to the second type occurring when the null hypothesis is accepted although it is wrong. Controlling the second type of error is much more involved and without deeper considerations, nothing can be said about this other possibility.

Nonetheless, a particularly interesting type of test regards *Benford's Law*, which can indicate potential "unnatural" manipulations in data series. This test has seen applications in tax fraud detection and other areas, and is presented here for the intriguing phenomenon that it points out.

Table 1. Comparison of selected statistical methods in the context of big data

Method	Hints	
	Pros	Attention
Pearson correlation, Blackbox models	• Easy to apply • Often do not require much domain knowledge • Widely understood (or at least thought so by many)	• Indications are generally weak, and provide no reliable signal into either direction ("everything okay" or "anomaly") • Require massive amounts of data • Without fine-tuning, necessarily inaccurate
Rule-based detections and statistical models/tests	• Can be made white-box and often enjoy rich theory • Can be very accurate and potentially adaptive	• Domain expertise inevitable • May provide only asymmetric indications (e.g., reliable upon rejecting hypotheses, but not confirming them)

3.2 Benford's Law: Testing for Artificial Data Manipulations

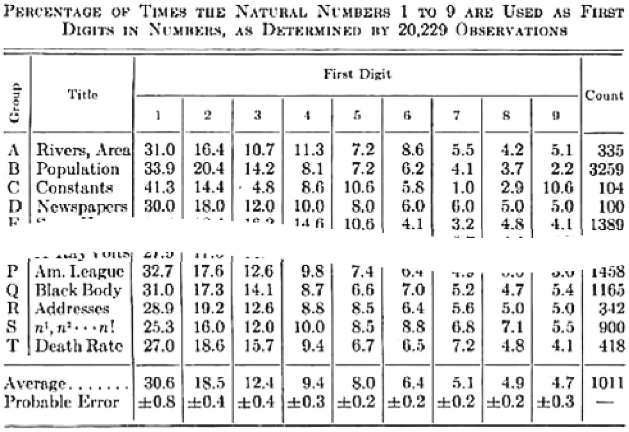
With the battery of statistical tests being widely explored in many branches of computer science, and especially in security, Benford's law is an exceptional example of a test that, without resorting to specific domains or complicated assumptions, manages to point out manipulations in many different datasets.

The key idea is to test not the numbers for any particular distribution, but rather to look at how the leading digit(s) in the numbers are distributed. Independently, Newcomb [10] (around the year 1881) and later Frank Benford (in 1938) [11], observed that in data arising from natural processes, the digit "1" appears substantially more often as the leading digit. The second most frequent digit is "2", followed by "3", with these three making more than 60% of all digits in a dataset.

The *Newcomb-Benford law* (or *Benford law* for short) precisely tells $\Pr(\text{leading digit}(X) = d) \propto \log_{10}\frac{d+1}{d}$ for the digits $d = 1, 2, \ldots, 9$, excluding the case of a leading zero for obvious reasons. This formula is surprisingly simple to derive: if X is an n-digit real number, when would its first digit be d? Obviously only if $d \cdot 10^n \leq X < (d+1) \cdot 10^n$, or by taking the base-10 logarithm, $\log_{10}(d) + n \leq \log_{10}(X) \leq \log_{10}(d+1) + n$. The range for the mantissa of $\log(X)$ to fall within for having d as leading digit has thus the width $\log_{10}(d+1) - \log_{10}(d) = \log_{10}\left(\frac{d+1}{d}\right)$. Assuming a "uniform" scattering of numbers over the real line, the claimed likelihoods are obtained. Benford's originally published material nicely supports the accuracy of this calculation; Fig. 3 shows the empirical values, next to the tabulated values on the right side.

Testing the law is straightforward: first, compute the relative frequency of leading digits in the given dataset, and compare it to what it should be according to the Newcomb-Benford law above. A deviation exceeding some threshold can be taken as an indication to dig deeper and perhaps look for artificial manipulations to the data (an abnormality). In general, the law is applicable whenever there are (i) many influence factors, (ii) the data set is large (big data). The test will, however, most likely fail on data that (i) is artificial or systematic, such as serial or account numbers, credit card numbers, etc., (ii) the data obeys natural limits (minimum or maximum bound), or (iii) if the data base is small. The exclusion of artificial or systematic data may appear restrictive but only mildly so: Many kinds of numbers like serial numbers, packet indices, network card (MAC) addresses, or ISBN numbers follow a precise structure and are thus often logically checkable for consistency (as they carry check-listed prefixes, verification digits, or similar). Thus, such number, unlike those arising from physical processes, usually do not need a statistical checkup.

The test can be generalized to more than the leading digit, with the respective law following in the same way as in our derivation above. For practical purposes, it is conveniently available in the `benford.analysis` [12] and `BenfordTest` [13] packages for the `R` system [14].



Percentage of Times the Natural Numbers 1 to 9 are Used as First Digits in Numbers, as Determined by 20,229 Observations

Group	Title	First Digit 1	2	3	4	5	6	7	8	9	Count
A	Rivers, Area	31.0	16.4	10.7	11.3	7.2	8.6	5.5	4.2	5.1	335
B	Population	33.9	20.4	14.2	8.1	7.2	6.2	4.1	3.7	2.2	3259
C	Constants	41.3	14.4	· 4.8	8.6	10.6	5.8	1.0	2.9	10.6	104
D	Newspapers	30.0	18.0	12.0	10.0	8.0	6.0	6.0	5.0	5.0	100
E	[illegible]	[illegible]	[illegible]	[illegible]	14.6	10.6	4.1	3.2	4.8	4.1	1389
	[illegible]	[illegible]	[illegible]	[illegible]							
P	Am. League	32.7	17.6	12.6	9.8	7.4	6.4	[illegible]	[illegible]	[illegible]	1458
Q	Black Body	31.0	17.3	14.1	8.7	6.6	7.0	5.2	4.7	5.4	1165
R	Addresses	28.9	19.2	12.6	8.8	8.5	6.4	5.6	5.0	5.0	342
S	$n^1, n^2 \cdots n!$	25.3	16.0	12.0	10.0	8.5	8.8	6.8	7.1	5.5	900
T	Death Rate	27.0	18.6	15.7	9.4	6.7	6.5	7.2	4.8	4.1	418
Average.		30.6	18.5	12.4	9.4	8.0	6.4	5.1	4.9	4.7	1011
Probable Error		±0.8	±0.4	±0.4	±0.3	±0.2	±0.2	±0.2	±0.2	±0.3	—

d	$\log\frac{d+1}{d}$
1	0,301
2	0,176
3	0,125
4	0,097
5	0,079
6	0,067
7	0,058
8	0,051
9	0,046

Fig. 3. Empirical evidence of the Newcomb-Benford law [11]

4 Integration Towards Practical Trust Management

Now, let us discuss how the techniques described above lend themselves to application with the signals obtained from technical systems. Anomaly detection and data collection systems can serve as sources for the Bayesian updates and provide data for statistical tests, and Fig. 4 shows how the above trust and manipulation tests would integrate with a system like synERGY: essentially, trust can almost “naturally” be derived from the data that the system generates (note the overlap of Figs. 1 and 4 at the “broker” component), possibly exploiting already existing classification functions that a SIEM or event anomaly detection modules may already offer. The point is here a double use of these features, not only for the system’s primary purpose (e.g., anomaly detection), but also perhaps for trust establishment as an add-on “almost for free” to the existing SIEM. The block “classification” may herein embody not only existing analysis modules from the host system, but also offer its own analyses based on statistics as above.

For the NB test, suitable data would include (but be not limited to): latency times, packets per time unit, packet sizes, but in particular also measurement data inside the packet content; basically, any data arising from physical processes would be suitable. Meta-information and protocol overhead data, such as serial numbers, packet numbers, or similar, would not be suitable for NB testing. This data undergoes more systematic checks in anomaly detection engines (where events are analyzed for logical consistency using rule-based checks and by virtue of sophisticated statistics). Basically, the anomaly detection can deliver two kinds of output useable with the Beta reputation model:

(1) An “everything OK” result upon a test for an anomaly. This would mean that two events in question are being checked for consistency, with a positive outcome, meaning that no indication of suspicious behavior was found. For the component in question, we can compactly represent the trust model as a pair of integers (a, b)

being the parameters of the Beta distribution and defining the trust value $\frac{b}{a+b} \in (0,1)$. A Bayesian update on this distribution upon a positive incident then just increases the parameter $b \leftarrow b+1$ (and accordingly changes the Beta distribution mix if there is uncertainty tied to this update, using model averaging).

(2) An indication of an anomaly: this would usually refer to some specific component, whose Beta reputation model, represented by a pair (a,b) of parameters for the Beta distribution would be updated into $(a,b) \leftarrow (a+1,b)$, i.e., increasing the count of negative experience.

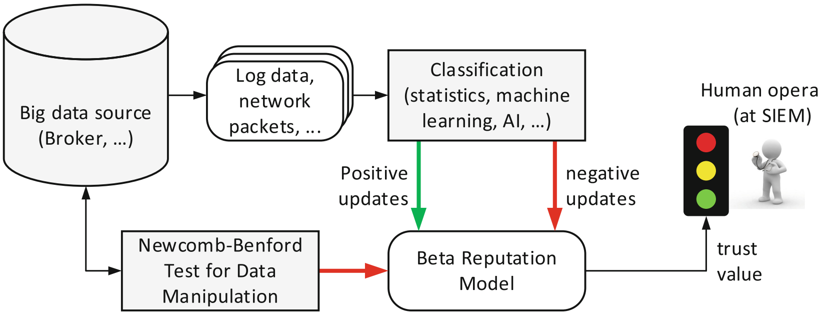

Big data source (Broker, ...)
Log data, network packets, ...
Classification (statistics, machine learning, AI, ...)
Human operator (at SIEM)
Positive updates
negative updates
Newcomb-Benford Test for Data Manipulation
Beta Reputation Model
trust value

Fig. 4. Integration of Trust Models and Manipulation Tests based on the synERGY example

5 Conclusion

Whether the analysis of big data is valuable or produces nonsense highly depends on the proper way of data selection and data analytics. This work discussed one application of big data for trust management, and discussed a few do's and don'ts in the application of some standard and non-standard techniques.

A general word of warning is advisable on the use of black-box models such as some neural networks. Despite the tremendous success of deep learning techniques in a vast variety of applications, the results generally remain confirmed only because they apparently work, but do so without offering any deeper explanations as to the "why". If not only the result is relevant, but also the reason why it is correct, then neural networks can only deliver half of what is needed. Generally referring to trust, transparency is a qualitative and important requirement, simply because understanding the "why" of a result helps fixing errors and improving mechanisms for the future.

The take-home messages of this work are briefly summarized as follows:

1. The strategy on how to fill the gaps in missing data is crucial (you should not infer information that you inserted yourself before).
2. You cannot use big data to tell you something (as it can tell you anything); you have to formulate a question and use the big data to get an answer to it.
3. Trust is always a subjective matter, no matter how "objective" the underlying model may be. That is, complex math or formalism can create the illusion of accuracy or reliability, although neither may hold.

4. Knowing is generally better than not knowing: in a choice between two models, one a white, the other a black box, the more trustworthy model is always white (in security, the trustworthy paradigm is Kerckhoffs' principle, demanding that every detail of a security algorithm should be openly published, with the security only resting on the secrets being processed).
5. Never blindly rely on any machine learning or statistical method: using a black-box model in a default configuration is almost a guarantee of failure. Instead, utilize domain expertise as much as possible, and calibrate/train models as careful as you can. This is the only way of inferring anything decent.

Acknowledgement. This work was partly funded by the Austrian FFG research project synERGY (855457).

References

1. Skopik, F., Wurzenberger, M., Fiedler, R.: synERGY: detecting advanced attacks across multiple layers of cyber-physical systems (2018). https://ercim-news.ercim.eu/en114/r-i/synergy-detecting-advanced-attacks-across-multiple-layers-of-cyber-physical-systems. Accessed 13 Jul 2018
2. SecurityAdvisor. HuemerIT. https://www.huemer-it.com/security-solutions/#SecurityAdvisor. Accessed 4 July 2018
3. Wurzenberger, M., Skopik, F., Settanni, G., et al.: AECID: a self-learning anomaly detection approach based on light-weight log parser models. In: Proceedings of the 4th International Conference on Information Systems Security and Privacy, pp. 386–397. SCITEPRESS - Science and Technology Publications (2018)
4. Friedberg, I., Skopik, F., Settanni, G., et al.: Combating advanced persistent threats: from network event correlation to incident detection. Comput. Secur. **48**, 35–57 (2015). https://doi.org/10.1016/j.cose.2014.09.006
5. Rass, S., Kurowski, S.: On Bayesian trust and risk forecasting for compound systems. In: Proceedings of the 7th International Conference on IT Security Incident Management & IT Forensics (IMF), pp. 69–82. IEEE Computer Society (2013)
6. Robert, C.P.: The Bayesian Choice. Springer, New York (2001)
7. Jøsang, A., Ismail, R.: The beta reputation system. In: Proceedings of the 15th Bled Electronic Commerce Conference (2002)
8. Rass, S., Slamanig, D.: Cryptography for Security and Privacy in Cloud Computing. Artech House, Norwood (2013)
9. Vigen, T.: Spurious Correlations, 1st edn. Hachette Books, New York (2015)
10. Newcomb, S.: Note on the frequency of use of the different digits in natural numbers. Am. J. Math. **4**(1/4), 39 (1881). https://doi.org/10.2307/2369148
11. Benford, F.: The law of anomalous numbers. Proc. Am. Philos. Soc. **78**(4), 551–572 (1938)
12. Cinelli, C.: benford.analysis: Benford analysis for data validation and forensic analytics (2017). https://CRAN.R-project.org/package=benford.analysis
13. Joenssen, D.W.: BenfordTests: statistical tests for evaluating conformity to Benford's Law (2015). https://CRAN.R-project.org/package=BenfordTests
14. R Core Team. R: a language and environment for statistical computing (2018). http://www.R-project.org

GDPR Transparency Requirements and Data Privacy Vocabularies

Eva Schlehahn[1] and Rigo Wenning[2](✉)

[1] Unabhängiges Landeszentrum für Datenschutz (ULD, Independent Centre for Privacy Protection) Schleswig-Holstein, Kiel, Germany
[2] World Wide Web Consortium/European Research Consortium for Informatics and Mathematics (W3C/ERCIM), Sophia Antipolis, France
rigo@w3.org

Abstract. This tutorial introduced participants to the transparency requirements of the General Data Protection Regulation (GDPR) [35]. Therein, it was explored together with the attendees whether technical specifications can be valuable to support transparency in favour of a data subject whose personal information is being processed. In the context of the discussions, past and present international efforts were examined that focus on data privacy vocabularies and taxonomies as basis work to enable effective enforcement of data handling policies. One example of a current undertaking in this area is the W3C Data Privacy Vocabularies and Controls Community Group (DPVCG) which aims at developing a taxonomy of privacy terms aligned to the GDPR, which encompasses personal data categories, processing purposes, events of disclosures, consent, and processing operations. During the tutorial session, the potential of such efforts was discussed among the participants, allowing for conclusions about the need to re-align and update past research in this area to the General Data Protection Regulation.

Keywords: General Data Protection Regulation · EU law · Transparency · Data privacy vocabularies · Technical specifications supporting GDPR compliance

1 Introduction

With the increasing digitization, the growing success of IoT devices on the market, and the incremental deployment of Big Data analysis of customer behaviour, it is apparent that ICT services and systems are by now widely used to link various types of information, recognize patterns and correlations, and to assess risks or chances on the basis of statistical insights. Data subjects whose personal data is being collected and processed are usually not aware of the scope and consequences of such assessments.

This leaves them exposed to the frequently opaque usage and commercialization of their personal information by data-driven companies. There is a significant lack of control by data subjects, since it is very difficult for individuals

E. Kosta et al. (Eds.): Privacy and Identity 2018, IFIP AICT 547, pp. 95–113, 2019.
https://doi.org/10.1007/978-3-030-16744-8_7

as end users of ICT services to obtain a clear picture how much data has been collected about them, from which sources, for which purposes, and with whom it has been shared. This situation is compounded by deficiencies in terms of controller and processor controllability and accountability[1]. With the intention of changing the regime of opaqueness by data-driven businesses using long and for the layman customer incomprehensible privacy policies, the European legislators made an effort to give transparency and intelligible information of the data subject some increased weight with the GDPR. The purpose of this tutorial was to analyse the requirements of the GDPR in terms of transparency and information obligations of the controllers. The following sections reporting on the session content will also foray into the domains of ethics and technology to determine whether those can in addition to the legal demands provide some insight how transparency can be understood and realized.

Based on the three dimensions legal, ethics and technology, a more distinct insight can be gained what transparency actually means and what it needs to encompass to meet the threshold of GDPR compliance. In this context, limits and challenges to its realization are explored, taking into account the upcoming ePrivacy Regulation as well. Past and current approaches are explained on how privacy by design via technical specifications aimed to enhance data protection compliance. In the last section, conclusions are drawn that call for more work and research in this area.

2 GDPR Transparency Requirements

From European data protection law perspective, transparency is a core necessity to empower the data subject. This means knowledge and the means to hold controllers and processors of his or her personal data accountable. For instance, it has been relatively clearly stated in recital 43 of the GDPR by explicitly mentioning transparency as a tool to better *'balance out the power asymmetry between data subjects and organizations'*. In this context, the emphasis on the empowerment of the data subject is reinforced by the explicit requirement of transparent information, communication and modalities for the exercise of the rights of the data subject, Art. 12 (1) GDPR (bold highlights by the authors):

> *'1. The controller shall take appropriate measures to provide* ***any information*** *[...] relating to processing to the data subject in a* ***concise, transparent, intelligible and easily accessible form, using clear and plain language****, in particular for any information addressed specifically to a child. The information shall be provided in writing, or by other means, including, where appropriate, by electronic means. When requested*

[1] Cf. with regard not only to the GDPR, but also to the review of the ePrivacy Directive, see [12]:pages 3, 7, 10, and 11 as well as in [11], pages 4 f. The results of the public consultation and the Eurobarometer survey outcomes strongly indicate a lack of citizen's confidence of being able to control and protect own personal data online.

> *by the data subject, the information may be provided orally, provided that the identity of the data subject is proven by other means.'*

Beyond this obligation for the controller, the GDPR has a multitude of other sources also determining that the perspective of the data subject is the deciding factor whenever it seems doubtful whether transparent information was provided about a processing operation. This is a central difference to the domain of IT security, where the processing organisation, its business secrets and company assets are the paramount subjects of protection. Therefore, in the realm of personal data protection with its fundamental rights underpinning, the following questions present themselves whenever a personal data processing operation is intended:

- Which data shall be collected and processed, and to which extent?
- In which way shall the data be processed, using which means?
- For which purposes shall the data be processed, and by whom?
- Is a transfer to and/or storage at other parties/foreign countries foreseen?

A concise knowledge about the points above is a necessary precondition enabling data subjects to exercise their rights granted by the GDPR. Such rights include the right to be informed, to access, rectify, or erase one's own personal data. Consequently, it is essential to capture the complete life-cycle of personal data. This ranges from the moment of initial collection over all processing operations performed until the deletion of the information.

Yet, transparency is important not only for data subjects, but also for controllers, processors, and data protection supervisory authorities as well. For instance, data controllers usually desire to maintain internal knowledge and controllability of their own processing operations. This does not only benefit business efficiency needs, but is also important with regard to compliance efforts in order to properly guarantee the data subject‘s rights. Moreover, the aforementioned compliance efforts must be demonstrable by the controller (see Article 24 GDPR). By Articles 12–14 GDPR, the controller is obliged to fully comply with comprehensive transparency and information duties, which will in turn require the implementation of correlating technical and organizational measures. Besides the controllers, data processors have the obligation to assist the controller in compliance efforts while being bound to controller instructions and oversight. Furthermore, knowledge about the inner workings of a personal data processing operation is also crucial for data protection supervisory authorities to perform their supervision and audit duties.

In addition to the general transparency requirements in the GDPR, the conditions of valid consent play a significant role. According to Article 4 (11) in combination with Art. 7 GDPR, valid consent must be freely given, specific, informed and unambiguous. The statement of consent must be a clear affirmative action of the data subject, and given for one or more specific purposes. It is notable that the existence of valid consent must also be demonstrable by the controller of the processing operation. Consequently, transparency is a crucial element from many different perspectives. This includes fairness and lawfulness personal data processing. Below, a tabular overview is given. It shows the various articles and recitals of the General Data Protection regulation mentioning and requiring transparency:

Article	Title
5 (1) a.	Principles relation to processing of personal data
12	Transparent information, communication and modalities for the exercise of the rights of the data subject
13	Information to be provided where personal data are collected from the data subject
14	Information to be provided where personal data have not been obtained from the data subject
15	Right of access by the data subject
19	Notification obligation regarding rectification or erasure of personal data or restriction of processing
25	Data protection by design and default
30	Records of processing activities
32	Security of processing
33	Notification of a personal data breach to the supervisory authority
34	Communication of a personal data breach to the data subject
40	Codes of conduct
42	Certification
Transparency mentioned in Recitals:	
32, 39, 42, 58, 60, 61, 63, 74, 78, 84, 85, 86, 87, 90, 91, 100	

From an ethics perspective, transparency is a central requirement as well. Many ethical principles have evolved historically and are recognizable in values that are laid down e.g. in the European Convention on Human Rights [17], or the European Charter of Fundamental Rights [13]. According to the European Group of Ethics in Science and New Technologies, core values are e.g.:

- The dignity of the human being
- Freedom
- Respect for democracy, citizenship, participation and privacy
- Respect of autonomy and informed consent
- Justice
- Solidarity [21]

Already since 1985, ethical experts demand transparency in the context of ICT. Moor has introduced transparency as the crucial element to encounter the so-called '*invisibility factor*', which is inherent when information and communication technologies are being used. This '*invisibility*' has three dimensions:

- **Invisible abuse**, e.g. taking advantage by the use of ICT to adapt the program or to remove or alter confidential information.
- **Invisible programming values**, where a programmer (either consciously or even unconsciously) influences the workings of a software algorithm and embeds his own values or prejudices.
- **Invisible complex calculations**, which are '*beyond human comprehension*', the system as '*black box*' where no one can tell if the results generated are correct. [29]

The general purpose of transparency from an ethics perspective is making the underlying values of coded software (algorithms) recognizable. This aims not only at the processing itself, but also at the results generated by automated decision-making.

From a technical perspective within the ICT sector (esp. in the US domain), transparency was for quite a long time understood as the exact opposite, i.e. '*obfuscating*' all information about systems and processes – and not burden the user with it [20]. However, the GDPR's concept of transparency that wants to give users knowledge and control over the processing of their data and over the workings of the ICT systems, is increasingly recognized outside of Europe as well. Moreover, it gets increasingly recognized that transparency should aim not only at user interface (UI) aspects, but should encompass the whole ICT system including the system architecture and the data flows [4,18,23]. Typical high level examples how transparency can be supported by technical and organisational means are the verification of data sources (keeping track of data), the documentation of IT processes, logging of accesses & changes of the data stock, versioning of different prototypes/systems, documentation of testing, documentation of (related) contracts, or of consent (if applicable: given/refused/withdrawn), consent management possible from a mobile device, and the support of data subject's rights via technology, e.g. easy access to own personal data, possibilities of deletion or rectification of wrong, inaccurate or incomplete information [31][2]. From technical perspective, transparency generally aims at the predictability and auditability of the used IT, an aspect that is often also called provenance. This entails re-tracing and understanding past events, showing the technical and organisational setup, and avoiding or mitigating possible future issues. Usually, three different dimensions of provenance are being differentiated, namely the provenance of data, provenance of processes, and reasoning (or analytical) provenance. Provenance of data means that in all cases, the data flow is documented, while the documentation as well as the system itself can give insight about the source, type, quality and contextual allocation of the data, including the applicable data handling rules [6]. Examples how to realize data provenance

[2] With an exemplary list of transparency-enhancing technical and organizational measures referenced in the handbook of the Standard Data Protection Model recommended for use in Germany.

with technical measures are sticky policies, differentiated data reliability scores, or automated deletion routines implemented. Provenance of processes means a proper documentation of the ICT components and analytic tools that are being used to process the data, which includes a documentation of the used analytical parameters as well to avoid such systems acting as kind of a black box producing non-retraceable results [32]. Finally, reasoning or analytical provenance is strongly related to the results of analytic processes. Here, transparency shows how analytics systems have been used to generate a certain output. In contrast to the process provenance, this aspect includes also the human or organizational factors around the ICT usage. Examples of measures to support reasoning provenance are the technical support of information and access rights in favour of data subjects, or the auditability of the processing operations (e.g. by human readable policies) [22]. All three perspectives – legal, ethical, and technical – have one thing in common: The requirement to involve stakeholders. This includes not only data subjects but also addresses controllers and processors. They all need to have relevant and sufficient information in order to understand the respective data processing operations. Only looking at GDPR-obligations for transparency might fall short ethically, if a holistic approach is the goal. In this case, it may be desirable to extend the requirements and understanding also to the risks involved and the decisions based on the results of the processing. Consequently, transparency could be understood as the property that all data processing – meaning all operations on data including the legal, technical, and organizational setting – and the correlating decisions based on the results can be understood and reconstructed at any time. Such an understanding of transparency entails a full capture of:

- the types of data involved,
- their source and quality/reliability
- processing purposes,
- circumstances of the processing,
- used systems and processing operations,
- the generated results,
- lawfulness and
- the related legal responsibilities (accountability) [28][3].

However, such kind of transparency is hard to formalize and measure so far. Fully comprehensive concepts are not yet state of the art. Current implementation approaches typically do not only concern the IT system and technical means alone. Rather, a comprehensive approach in consideration of legal, ethical, technical, and organizational/business expertise seems advisable to avoid discrepancies and to enable synergy effects. The fact that there is no 'universal' solution available should be recognized. Transparency solutions must therefore always be developed dependent on a careful assessment of context, individual

[3] Meis et al. constructed a set of requirements for a transparency focused ontology on the basis of the ISO/IEC 29100:2011standard, OECD principles, and the US fair information practices (FIPs), and which already entails some of these aspects.

case, and the processing purposes, means and foreseen execution. Only with such an earnest approach, the verifiability of data processing operations can be attempted. This goes beyond, but encompasses all transparency requirements of the GDPR in order to achieve coherently formulated functional requirements for automated processing systems.

The upcoming ePrivacy Regulation (ePR) contains interesting transparency requirements. As of now, the application scope includes electronic communicationsdata, meta- and content data. This extends the application scope of the current ePrivacy Directive that is still in force. Concerned with the new regulation will be all OTT[4], ECS[5] and software providers permitting electronic communications, including the retrieval and presentation of information on the Internet. This includes a lot of different apps such as IoT devices and many other things. In the commission proposal version, Art. 9 of the draft-ePR explicitly refers to the GDPR for consent requirements, which includes the correlating information and transparency obligations of the controller towards the data subject. In terms of transparency, relevant changes were recently hotly debated with regard to Art. 10 of the European Parliament version [34]. This article obliged hardware and software providers to ensure privacy by default, including the possibility of users to set their own privacy settings. However, in the later EU Council version, this article was deleted. Therefore, it is yet highly unclear when the Trilogue process of the ePrivacy Regulation will achieve a compromise result and how it will look like in the end.

3 Data Protection Focus on Technical Specifications

In this section, some examples for basic approaches to GDPR-aligned technical specifications are given. Based on the transparency requirements explained above, the minimum core model for personal data processing policies should usually entail the data categories, processing operation and purpose, storage (retention time and storage location), and the recipients of the data to enable a coherent definition of a data usage policy. Such a core model is visualized below:

Going more into detail for the GDPR-aligned technical specifications, categories of personal data could for example entail differentiations like master record data, location and movement data, call records, communication metadata, and log file data. Moreover, special categories of personal in the sense of Art. 9 GDPR should be taken into account. This concerns personal data related to racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health, and data concerning a natural person's sex life or sexual orientation.

Beyond the categories of the data, technical specifications should support the documentation of processing purpose(s) and the correlating legal ground. If

[4] OTT (Over The Top Services) are communication systems over data networks, e.g. skype.

[5] Electronic Communication Services.

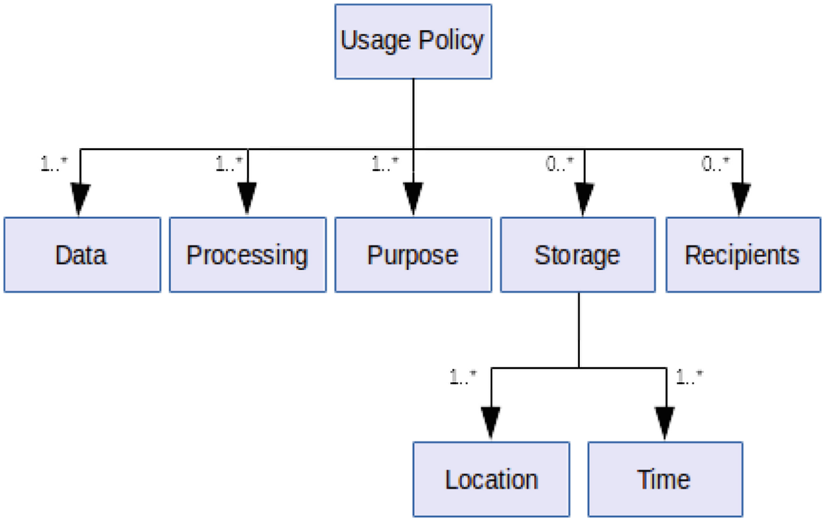


Fig. 1. The SPECIAL Core personal data processing policy model.

consent is determined as the applicable legal basis for processing, a link should be enclosed to the actual consent agreement itself, so the exact wording can be reviewed if needed. Furthermore, a versioning of consent agreements and the capture of the current status could be technically supported, e.g. by attaching pre-designed labels, such as given (if yes, specify whether explicit or implicit), pending/withheld, withdrawn, referring to the personal data of a minor, or referring to the personal data of a disabled person in need of specific accessibility provisions to manage consent. These are of course only initial ideas which could be further developed depending on context and need.

Technical specifications based on GDPR terms should enable a documentation of the involved data controller(s) and processor(s), as well as storage location and cross-border data transfers. For the latter, a specification of the involved foreign countries is also needed since this has significant impact on the legal assessment of the lawfulness of a processing operation. In this context, it is advisable to capture the location of the data centre where the processing and storage occurs, including the location of the controller establishment. For the latter, it is relevant to know whether the data transfer occurs only within the European Union, or to a third country with a legal basis for compliance acc. to Art. 44 et seq. GDPR (treating them as ‘EULike’). Examples for such a legal ground for third country transfer could be an adequacy decision by the European Commission, appropriate safeguards, or existing Binding Corporate Rules (BCR). Where possible, a link should be provided that points towards any source documenting the respectively applicable legal document, e.g. to the Commission’s adequacy decision or the BCR text. However, there might also be cases where data transfers to other third countries are foreseen for which any of the legal basis of Art 44 et seq. GDPR are not, or not anymore applicable. This could for instance, be the case if a court declares a BCR invalid, or the Commission changes an adequacy decision. To this end, it seems advisable to use country codes (e.g. according to TLD, ISO 3166), which allow for an easy later adaption of a policy in case of legal changes. Furthermore, in terms of data handling policies, it deems sensible to also incorporate labels that can express rules excluding data transfers to some jurisdictions (e.g. 'notUS', 'notUK') (Fig. 1).

Regarding the storage periods, specifications for certain deletion times and correlating actions required can be defined. Some rough examples are:

- delete-by_ or delete-x-date_month_after <event>,
- no-retention (no storage beyond using once),
- stated purpose (storage only until purpose has been fulfilled),
- legal-requirement (storage period defined by a law requiring it),
- business practices (requires a deletion concept of controller),
- or indefinitely (e. g. for really anonymized data, and public archives).

Last, but not least, technical specifications should enable enforcing rules how to handle the data. For instance, defining the user or access activity that is being allowed or even determined for the personal data within an ICT system, such as read-only, write, rectify, disclose, deletion, anonymize, pseudonymize, or encrypt. Furthermore, notification obligations of the controller towards the data subject could be captured as well under certain preconditions, eventually with predefined action time, e.g. reminder of consent withdrawal right, or communication in case of a data breach.

4 Privacy Enhancing Technologies: The Evolution of Technical Approaches for Transparency and Controls

In her PhD Thesis, McDonald had found that most privacy policies are written beyond most adults reading comprehension level [26]. In another study, she showed that the time of reading all those natural language privacy policies would take an average person with average viewing habits around an average 201 hours of reading privacy policies per year [27]. For work, this would mean an average of 25 working days only to read privacy policies. Additionally, current privacy policies have a 'take it or leave it' approach. Data subjects can read the privacy policy, but mostly they cannot change any of the data collection. We see this change now with the advent of GDPR. But the interfaces are mostly very crude, sometimes culminating at an opt-in/out of more than hundred trackers. Under such circumstances, data self-determination can be exercised only for a few selected sites. The nominal transparency by privacy policies, carefully crafted by legally savvy people remains nominal. Aleecia McDonald had the merit to scientifically prove an assumption that was made very early on in the development of privacy enhancing technologies (PETs): Because the bandwidth of humans is very limited, the computer should help humans to better understand their situation.

4.1 From PICS to P3P

The quest for more transparency on the Web started very early, already in 1995. One of the very early and very successful adopters of web technologies was the porn industry. They were the first to make money with their content. At some point, the porn content was spilling over to people who did not want to see it. This was also an issue of self-determination. And there was a technical challenge. People wanted to know before the actual content was downloaded and displayed to avoid bad surprises. This included filtering content for children. By labelling the content with a rating, a filter would be able to recognize the labels and react or block based on the rating. This was opposed to several governmental initiatives that wanted to install central filters on the Web to filter out so called 'illegal and harmful content'. The PICS [24] filtering would not have given such huge power to a single node on the Web. Freedom of information, which includes freedom to receive information, would be preserved. Not a central government institution would decide what people could see, but every individual or family or school could define a filter based on the ratings and their cultural preference. The plan to solve issues of self-determination with labelling came back in several iterations over the past two decades. And it became more sophisticated with each iteration. From a technology point of view, PICS was a very simple tool. There was no XML and no RDF yet in 1996. There was a very simple syntax that transported labels via an HTTP response header. The labels applied to the URI that the browser had requested. The labels themselves were not specified by W3C as the organisation did not feel competent in this area. But a real good labelling system was never found. The Internet Content Rating Association (ICRA) created the predominant labelling scheme. The issue with the system was that normal authors did not have an incentive to label their pages unless they were of a certain type. In 1999, all porn sides carried labels because they wanted to be found. Others did not want to label their pages because there was a (justified) fear that governmental nodes on the Web would then filter all content thus stifling freedom of speech. But most pages were not labelled meaning PICS remained of limited usefulness. Because of the criticism, the lack of incentives, and because the labelling was complex and coarse at the same time, PICS did not take off. The browsers finally removed support for the filtering and ICRA went out of business in 2010.

In 1997, when there was still a lot of enthusiasm about PICS, the idea came up to also use such a transparency scheme for privacy and data protection. In fact, people found out that a lot of http protocol chatter was stored in log files and used for profile building. People did not realise that such profiling was happening as the browser was totally opaque about it. The US Senate threatened the advertisement industry with legislative action. Industry feared that such legislation would damage their revenue model. It was better to improve the user experience than to continue to provide a picture perfect example on why legislation was needed. This created a discussion within W3C about technical remedies to the opaqueness of the HTTP protocol. Major vendors like Microsoft joined a W3C Working Group to make a tool that could replace legal provisions. A combination of researchers and industry started to create the Platform for Privacy Preferences (P3P) [14]. The idea was similar to the one in PICS: Servers would announce what data they collect and what they do with it. To express the data collection and usage, P3P created the P3P vocabulary. P3P came before XML, so the file was in a similar, but not quite compliant data format with angle brackets. This allows a service to encode their practices in a machine readable file and store it on the server. For the client side, the Group created a vocabulary to express preferences and match those preferences against the P3P policy document found on the server for the resource requested via HTTP [37]. The idea was that APPEL [36] would enable privacy advocates to write preferences for consumers. But APPEL was finally taken off the standards track as it had technical and mathematical problems [1]. The development concentrated on the policy language and client side preferences. The preference exchange language was downgraded to a secondary goal.

4.2 The Success and Decline of P3P

Everyone in the technical community found the idea behind P3P compelling. The vocabulary was well respected and a lot of sites started to implement P3P on the server side. At the same time, browsers were complaining that it was too complex to implement P3P on the client side. In a last minute chaotic action, the P3P WG accepted a proposal from Microsoft to introduce so called ‘compact policies’. The verbose and complex XML-like file expressing the P3P policy was replaced by a very coarse representation of abbreviated policy tokens transported by a HTTP header.

The power of P3P became shortly visible when Microsoft[6] announced that the new Internet Explorer (IE) 6.0 would only allow third parties to set a cookie if this third party would send P3P compact policy tokens about data they collect and about the associated purposes. Within a few month, P3P compact tokens where present in a majority of HTTP driven exchanges. But they were very different from what the Working Group had expected. Someone somewhere on a developer platform created a ‘make-IE-6 –happy’ string with P3P compact tokens in them. This would fool IE 6 to believe that the service would do the thing it announced and allow the third party cookie to be set. Within short time, a very large proportion of the tokens found were those ‘make-IE-6-happy’ tokens. Of course those were obvious lies using technical means. Critics of P3P had always said that P3P itself had no enforcement. And indeed, P3P just made declarations easier and machine readable to help humans assess the situation, remove the opacity. It was never meant to be an enforcement tool. It removed opacity and relied on the fact that the legal system would sanction lies and deceptive behaviour. And there were cases when this was applied successfully. Especially when people really tried to implement P3P and exposed in P3P what they were really doing. This way someone found out that the US Drug addiction

[6] Microsoft had considerable market power on the Web in 2002.

online service for addicts had a tracking cookie. The cookie was subsequently retired.

The US Senate finally dropped the privacy legislation and the European data protection authorities preferred direct legal action within an orderly administrative procedure. There was no incentive for the industry anymore to invest into privacy enhancing technologies. Paying for lobbyists was more cost effective and allowed proven technology to remain unchanged.

What remained from P3P was its vocabulary, especially the 'STATEMENT' section. This section is still cited and influences the way people express policies in the area of data protection.

4.3 From the Web to the Enterprise World: Prime and PrimeLife

In Europe, research on the topic continued. From 2004–2008, the PRIME project[7] explored new ways. The assumption was that privacy policies can be turned into privacy rules. The system would then automatically enforce those rules. The principles of the project were the following:

1. Design starting from maximum privacy.
2. System usage governed by explicit privacy rules.
3. Privacy rules must be enforced, not just stated.
4. Trustworthy privacy enforcement.
5. Easy and intuitive abstractions of privacy for users.
6. An integrated approach to privacy.
7. Privacy integrated with applications [9].

The system had several use cases that were implemented using the technology created by the project. The project was very research oriented and decided for maximum privacy where ever possible. This included already anonymous credentials [8] and the machine-assisted management of pseudonyms. As it was very research driven, the assumption was that the client would issue preferences and the server would acknowledge receipt of those preferences or rules and act accordingly. But this did not work out in practice. In fact, most systems are designed for a limited variety of options. The user can chose between those options. If the user or data subject comes up with new preferences that are not yet implemented as a workflow on the server side, the system will simply fail. The PRIME project did use RDF [5] on both sides of the equation which cured some of the difficulties P3P had concerning the matching of preferences to policies. But it was too early because the use of RDF or Linked Data or graph data was not yet very widespread. The project decided deliberately against taking into account legacy enterprise databases and went for the pure research by requiring all systems to be RDF in order to work. The main argument was that integrating those legacy systems with their legacy interfaces would allow a malicious actor to circumvent the enforcement engine of the PRIME system. Given those practical constraints,

[7] https://cordis.europa.eu/project/rcn/71383/factsheet/en.

the PRIME project produced good scientific results, but its practical relevance remained marginal. It was not really usable for real world systems unless someone would create a brand new system from scratch. PRIME had improved the understanding of the difficulties with transport anonymity and also advanced considerably the state of the art concerning the sticky policy paradigm.

Drawing conclusions from this initial experience with a privacy-enabled data management system, most of the partners of the PRIME project created the PrimeLife project[8] to now concentrate on the challenges for such a privacy-enabled system in relation to real world systems. When the PrimeLife proposal was created, the predominant data format was XML [3,10] and the predominant software architecture was Web Services[9]. PrimeLife concentrated on those technologies and reduced the dependency on RDF. The attacking model was changed. An assumption was made that a company wanted to do the right thing. In fact, rights management systems are very hard to enforce once the entire computing resources are in the hand of the attacker. Consequently, PrimeLife had a security boundary between the client side and the server side, and a focus on data usage control on the service side. This allowed efforts to push boundaries on both sides. PrimeLife continued integrating ways to achieve results and checks without consuming personally identifiable information while further developing a system to create data value chains across company borders. One special merit of PrimeLife was the invention of the term '*downstream data controller*'. This terminology allowed researchers to better repartition and assess the respective responsibilities between data controllers and responsibilities further down in the data value chain. It allowed clearing the fog in discussions where people partly did not realise that they were talking about the same thing. Or, that they believed they were talking about the same person, while meaning another one. The term '*downstream data controller*' greatly reduced the confusion. For PRIME, the sticky policy paradigm was rather easy to implement as RDF with its URIs [7,16][10] on every triple had a natural way of addressing a specific data packet or data record. In fact, this is like a sentence in English language. To make a policy sticky, it has to be attached to the data record it applies to. Within RDF data, every object has a URI. So it is sufficient to create a new triple with a policy that points to the data record it wants to apply to. The URI used is a world unique identifier. This means wherever the data travels, the relation to the policy remains intact. This allows data value chains across company borders to honour the policies expressed and even impose the respective rules to subsequent downstream data controllers. In PrimeLife, because web services were used, the stickiness of the policy made a rather complex system necessary. All was using standard data formats, namely XACML [30] and SAML [2] to create, manage, transport and execute policy statements along a given data value chain. While this was very pragmatic at the time, technology has moved on and different things have to be used today.

[8] http://primelife.ercim.eu/.

[9] https://www.w3.org/2002/ws/ accessed 2019-01-21.

[10] RDF uses IRIs to identify objects.

PrimeLife had a rather holistic view, taking into account the data life-cycle and providing new ways to manage access control in a way that was much closer to the needs for business, e.g. depending of the role of the person wanting to access certain information. PrimeLife had an entire work package concentrating on user interface issues. Data self-determination, if taken seriously, requires the data subject to understand what is happening. This is not only an issue for transparency and the provision of information from the server/service side. It is also an issue of cognitive limitations by humans confronted with the huge wealth of information transitioning through today's communication systems. PrimeLife experimented with icons and logos and user interfaces and did usability testing in a lab environment [19]. One of the lasting outcomes was a logo: Users were tasked to find out about a privacy tool PrimeLife had developed. It was collecting data about who is collecting what to identify trackers and show them to the user[11]. Users had the task to find out who collects what about them. The challenge was, whether the users would find the tool and click on the right logo. All kinds of logos were tested. But only with an icon with footsteps on it, users found the tool easily. PrimeLife was also very successful in the scientific field testing out role based access control and a new type of group based social networking. It created a new policy language that was able to express all the metadata in standard languages and extended XACML to carry more metadata.

After PrimeLife, there have been further research and development efforts focused on providing better transparency for users of digital services. For example, the research work done in the A4Cloud project[12] built upon the PrimeLife Policy Language (PPL) and extended the identified requirements in order to create a proof of concept for an accountability policy language (A-PPL) for cloud computing contexts. This policy language addresses data handling rules corresponding to cloud provider's privacy policies, data access and usage control rules, retention period and storage location rules, logging, notification, reporting and auditability [33]. However, those efforts need further work since they have been made before the reform of the European data protection framework with its enhanced transparency requirements in favour of data subjects.

4.4 SPECIAL: From Enterprise Systems to Big Data and Knowledge Graphs

Enterprise systems have evolved a lot in the meantime. The European Commission has successfully implemented their public sector information strategy.[13] The aim is to provide public information that can be combined with private sector information to form new innovative services and products. The Big Data Europe project (BDE)[14] created an open source platform ready to use for everyone to

[11] The privacy dashboard was a Firefox extension that stored all data from the HTTP chatter into a local database and was able to show the tracking to the user. See http://primelife.ercim.eu/results/opensource/76-dashboard.

[12] http://a4cloud.eu/.

[13] See https://www.w3.org/2013/share-psi/ for more information and pointers.

[14] https://www.big-data-europe.eu.

provide an easy tool for processing analysis of information for the public sector in all seven societal challenges put forward by the Commission: Health, Food, Energy, Transport, Climate, Social sciences and Security. The project quickly found a first challenge. Most of the data found, personal or not, was of high heterogeneity. Within the same societal challenge a wealth of databases in silos was found. The variety issue was solved by using RDF and Linked data to join data from a high variety of sources. To do so, BDE developed a method to semantify the data streams coming into the big data platform. They called it semantic lifting. With an all semantic data lake, we are back in a situation where the insights of PRIME can be used to make policies sticky, notably by just adding policy information to the knowledge graph created via the semantification and other transformations. For the analysis, the parallelization of the processing was not known so far to the normal inference engines. BDE started work on the SANSA [25] stack now further developed by University of Bonn. It allows accomplishing even more complex policy data processing and inference in an acceptable amount of time. Now this engine was again usable to come up with a privacy enhancing system for big data. The SPECIAL project[15] uses this system to produce a new tool for sticky policy and data usage control within a big data environment. Special, like the initial PRIME project, uses the Linked data properties to annotate data. Through semantification, all data has a URI and can thus be an object of a Linked data statement. For internal purposes and for performance, the Linked data can be transformed into something else to better fit the legacy systems. But the data must retain Linked data properties, especially being linked to a policy statement. If data from a system is given to commercial partners in a data value chain, of course with the consent of the data subject, the RDF Linked data platform plays the role of a transport format. After the technical challenges of the past projects, the SPECIAL project found rather social challenges in context of the technical issues. Of course, a deep understanding of the Linked data world is needed to design policy annotated workflows. While the technology stack has not yet arrived in many production systems, it is a rather mature area with a high potential for new use cases. We are now moving from the question of how processing is organised to the question about the semantics for the actual policy annotations. P3P has given some direction with its statement vocabulary. But this was really advertisement driven and is not sufficient for the very generic privacy transparency and data management tool that will be created by SPECIAL.

This is why W3C organised a Workshop on Privacy and Linked Data in April 2018 [15]. The workshop assumed that most services today, lack the tools to be good citizens of the Web. Which is related, but not limited, to the work on permissions and on tracking protection. Because those permissions and tracking signals carry policy data, the systems have to react upon those signals. To react in a complex distributed system, the signals have to be understood by more than one implementer. The challenge is to identify the areas where such signals are

[15] Scalable and policy-aware linked data architecture for privacy, transparency and compliance (SPECIAL), https://www.specialprivacy.eu.

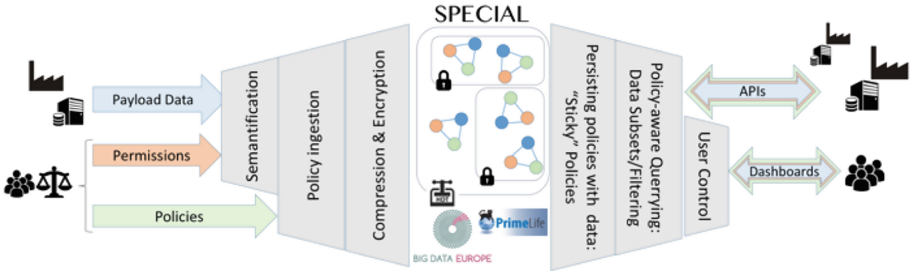


Fig. 2. A birds-eye view on the special data flow

needed for privacy or compliance and to make those signals interoperable. The Workshop concluded that more work on Data Privacy Vocabularies is needed. The Workshop participants decided to initiate a W3C Data Privacy Vocabularies and Controls Community Group (DPVCG).1 The DPVCG will develop a taxonomy of privacy terms, which include in particular terms from the new European General Data Protection Regulation (GDPR), such as a taxonomy of personal data as well as a classification of purposes (i.e., purposes for data collection), and events of disclosures, consent, and processing such personal data. Everybody is welcome to join the effort (Fig. 2).

5 Tutorial Outcomes and Conclusions

The main objective of the tutorial session was to introduce participants to the transparency requirements of the GDPR and to delve deeper into the possibilities of technical means supporting their realization. This was achieved by an introductory presentation, and discussions during and after this presentation that took place among all attendees. These discussions often evolved around the significance of earlier attempts of creating technical specifications and vocabularies for privacy terms, their usefulness for future endeavours in this field, and initial ideas what could be captured in such vocabularies and taxonomies to enable meaningful and enforceable data handling policies. The discussions showed that the work of the SPECIAL project as well as of the W3C Data Privacy Vocabularies and Controls Community Group are very important first steps into the right direction, while recommendations were made by workshop participants to stick very close to the GDPR and to avoid the pitfalls of former efforts in the area of standardisation. However, the discussions showed that especially in the context of big data applications, further research as well as real development work will be needed to inch closer to more GDPR-aligned processing taxonomies that can be adopted by businesses as well.

Acknowledgments. Supported by the European Union's Horizon 2020 research and innovation programme under grant 731601.

References

1. W3C Workshop on the long term Future of P3P and Enterprise Privacy Languages (2003). W3C. https://www.w3.org/2003/p3p-ws/
2. Security assertion markup language (saml) v2.0. Technical report, March 2005. https://www.oasis-open.org/standards#samlv2.0, https://wiki.oasis-open.org/security/FrontPage#SAML_V2.0_Standard
3. Extensible markup language (xml) 1.0 (5. edition). Technical report, November 2008. http://www.w3.org/TR/2008/REC-xml-20081126/
4. Engineering Privacy by Design (2011)
5. Rdf 1.1 primer. Technical report, June 2014. http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/
6. Gupta, A.: Data provenance. In: Liu, L., Özsu, M.T. (eds.) Encyclopedia of Database Systems. Springer, Boston (2009). https://doi.org/10.1007/978-0-387-39940-9
7. Berners-Lee, T., Fielding, R.T., Masinter, L.: Uniform resource identifier (URI): Generic syntax. Technical report (2005). http://www.ietf.org/rfc/rfc3986.txt
8. Camenisch, J., Lysyanskaya, A.: An efficient system for non-transferable anonymous credentials with optional anonymity revocation. In: Pfitzmann, B. (ed.) EUROCRYPT 2001. LNCS, vol. 2045, pp. 93–118. Springer, Heidelberg (2001). https://doi.org/10.1007/3-540-44987-6_7
9. Camenisch, J., Leenes, R., Sommer, D. (eds.): PRIME - Privacy and Identity Management for Europe. Lecture Notes in Computer Science, vol. 6545. Springer, Berlin (2011). https://doi.org/10.1007/978-3-642-19050-6
10. Collins, C.: A brief history of xml, March 2008. https://ccollins.wordpress.com/2008/03/03/a-brief-history-of-xml/
11. European Commission: Flash eurobarometer 443: e-privacy. Technical report, December 2016. http://data.europa.eu/euodp/en/data/dataset/S2124_443_ENG
12. European Commission: Summary report on the public consultation on the evaluation and review of the eprivacy directive. Technical report, August 2016. https://ec.europa.eu/digital-single-market/en/news/summary-report-public-consultation-evaluation-and-review-eprivacy-directive
13. European Council, European Parliament, and European Commission: Charter of Fundamental Rights of the European Union. Number 83 in Official Journal of the European Union C. European Union, pp. 389–403, March 2010. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2010:083:0389:0403:en:PDF
14. Cranor, L.F.: Web Privacy with P3P. O'Reilly & Associates Inc., Newton (2002). ISBN 0-596-00371-4
15. Decker, S., Peristeras, V. (eds.): Data Privacy Controls and Vocabularies: A W3C Workshop on Privacy and Linked Data (2017). W3C. https://www.w3.org/2018/vocabws/
16. Duerst, M., Suignard, M.: Internationalized resource identifiers (iris). Technical report 3987, January 2005. http://www.ietf.org/rfc/rfc3987.txt
17. ECHR2010: Convention for the protection of human rights and fundamental freedoms as amended by protocol no. 11 and no. 14, June 2010. http://conventions.coe.int/treaty/en/Treaties/Html/005.htm
18. Goodman, B., Flaxman, S.: EU regulations on algorithmic decision-making and a "right to explanation". AI Mag. **38**(3) (2017)

19. Holtz, L.-E., Nocun, K., Hansen, M.: Towards displaying privacy information with icons. In: Fischer-Hübner, S., Duquenoy, P., Hansen, M., Leenes, R., Zhang, G. (eds.) Privacy and Identity 2010. IAICT, vol. 352, pp. 338–348. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-20769-3_27
20. Inchauste, F.: The dirtiest word in UX: Complexity, July 2010. http://uxmag.com/articles/the-dirtiest-word-in-ux-complexity
21. Kinderlerer, J., Dabrock, P., Haker, H., Nys, H., Salvi, M.: Opinion 26 - Ethics of information and communication technologies. Publications Office of the European Union, February 2012. ISBN 978-92-79-22734-9. https://doi.org/10.2796/13541, http://bookshop.europa.eu/en/ethics-of-information-and-communication-technologies-pbNJAJ12026/
22. Kodagoda, N.: Using machine learning to infer reasoning provenance from user interaction log data: based on the data/frame theory of sensemaking. JCEDM Spec. Issue **11**(1), 23–47 (2017)
23. Koops, B.-J.: On Decision Transparency, or How to Enhance Data Protection after the Computational Turn, pp. 196–220 (2013)
24. Krauskopf, T., Miller, J., Resnick, P., Treese, W.: Pics label distribution label syntax and communication protocols. Technical report, October 1996. https://www.w3.org/TR/REC-PICS-labels-961031
25. Lehmann, J., et al.: Distributed semantic analytics using the SANSA stack. In: d'Amato, C., et al. (eds.) ISWC 2017. LNCS, vol. 10588, pp. 147–155. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68204-4_15
26. McDonald, A.M.: Footprints Near the Surf: Individual Privacy Decisions in Online Contexts. Ph.D. thesis (2010). https://kilthub.figshare.com/articles/Footprints_Near_the_Surf_Individual_Privacy_Decisions_in_Online_Contexts/6717041
27. McDonald, A.M., Cranor, L.F.: The cost of reading privacy policies. I/S: J. Law Policy Inf. Soc. **4**(3), 543–568 (2008). http://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/isjlpsoc4&section=27, https://kb.osu.edu/dspace/bitstream/handle/1811/72839/ISJLP_V4N3_543.pdf
28. Meis, R., Wirtz, R., Heisel, M.: A taxonomy of requirements for the privacy goal transparency. In: Fischer-Hübner, S., Lambrinoudakis, C., Lopez, J. (eds.) TrustBus 2015. LNCS, vol. 9264, pp. 195–209. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-22906-5_15
29. Moor, J.H.: What is computer ethics? Metaphilosophy **16**(4), 266–275 (1985). ISSN 1467–9973. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9973.1985.tb00173.x, https://web.cs.ucdavis.edu/~rogaway/classes/188/spring06/papers/moor.html
30. Moses, T.: Extensible access control markup language (xacml) v2.0. Technical report (2005). http://docs.oasis-open.org/xacml/2.0/access_control-xacml-2.0-core-spec-os.pdf
31. Conference of the Independent Data Protection of the Authorities. The standard data protection model. Technical report (2016). https://www.datenschutzzentrum.de/uploads/sdm/SDM-Methodology_V1.0.pdf
32. Pandit, H., O'Sullivan, D., Lewis, D.: Queryable provenance metadata for GDPR compliance. Procedia Comput. Sci. **137**, 262–268 (2018)
33. Azraoui, M., Elkhiyaoui, K., Önen, M., Bernsmed, K., De Oliveira, A.S., Sendor, J.: A-PPL: an accountability policy language. In: Garcia-Alfaro, J., et al. (eds.) DPM/QASA/SETOP-2014. LNCS, vol. 8872, pp. 319–326. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-17016-9_21

34. Sippel, B., European Parliament: Report on the proposal for a regulation of the European parliament and of the council concerning the respect for private life and the protection of personal data in electronic communications and repealing directive 2002/58/ec (regulation on privacy and electronic communications), October 2017. http://www.europarl.europa.eu/sides/getDoc.do?type=REPORT&mode=XML&reference=A8-2017-0324&language=EN
35. European Union: Regulation (EU) 2016/679 of the European parliament and of the council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation), May 2016
36. W3C: A P3P preference exchange language 1.0 (APPEL1.0) (2002)
37. W3C: The platform for privacy preferences 1.1 (P3P1.1) specification (2006)

Wider Research Applications of Dynamic Consent

Arianna Schuler Scott[1](✉), Michael Goldsmith[1], and Harriet Teare[2]

[1] Department of Computer Science, University of Oxford, Oxford, UK
arianna.schulerscott@cs.ox.ac.uk
[2] Centre for Health, Law and Emerging Technologies,
Nuffield Department for Population Health, University of Oxford, Oxford, UK

Abstract. As research processes change due to technological developments in how data is collected, stored and used, so must consent methods. Dynamic consent is an online mechanism allowing research participants to revisit consent decisions they have made about how their data is used. Emerging from bio-banking where research data is derived from biological samples, dynamic consent has been designed to address problems with participant engagement and oversight. Through discussion that emerged during a workshop run at the IFIP 2018 Summer School, this paper explores wider research problems could be addressed by dynamic consent. Emergent themes of research design, expectation management and trust suggested overarching research problems which could be addressed with a longer term view of how research data is used, even if that use is unknown at the point of collection. We posit that the existing model of dynamic consent offers a practical research approach outside of biobanking.

Keywords: Consent · Engagement · Revocation · Trust · Digital rights · Research practice · Data use · Cybersecurity

1 Introduction

Medical research relies on human subjects for cutting edge research and world-changing discoveries. Some studies require active human participation while others do not, specifically those using biological samples donated by patients and participants alike. Biobanks are research frameworks that store biological samples and data derived from those samples for research use.

Consent, or permission to use an individual's data is a fundamental tenet of human participatory research - identifiable personal information such as biological data must be protected and responsibility for this protection lies with both research teams and the institutions behind them. Informed consent is given by a research participant in response to information communicated about a research study's associated risk, benefits and procedure. This informed decision is made

E. Kosta et al. (Eds.): Privacy and Identity 2018, IFIP AICT 547, pp. 114–120, 2019.
https://doi.org/10.1007/978-3-030-16744-8_8

with a complete understanding of what is being agreed to - the individual must be aware of their decision to take part in the research process.

Legal provision, ethical oversight and technical controls have been used to provide this protection and as technology has developed, the way these mechanisms interact within research is changing. Dynamic consent is a real-time implementation of consent focusing on engaging participants with the research process and allowing them to revoke consent for data to be used for research purposes.

This paper explores whether there may be wider applications of dynamic consent than the context of bio-banking. By asking those involved in research about participant consent decisions and research expectations, we provide some idea of the problems experienced by researchers around engagement and revocation. This paper considers whether dynamic consent provides a solution to such problems, grounding arguments in discussion arising from the "Exploring Attitudes to Dynamic Consent in Research" workshop delivered at the 2018 IFIP Summer School. We begin with a short review of dynamic consent and how it is situated in biomedical research before describing the workshop, discussing emergent themes and providing direction for further work.

2 Background

A tool that supports lawful sharing and re-use of data [2], dynamic consent promotes the active participation of biomedical research participants to create socially aware and impactful research. A "social licence" [3] for research means that it's value is reciprocal - dynamic consent puts control in the hands of participants [6], allows their input to feed back into research practice [5] and emphasises the importance of individual knowledge [11] and autonomy of choice [10]. Biobanks are research frameworks where data is derived from biological samples. They often implement broad consent, where participants agree to unspecified future use of their sample at the point of donation, biopsy or clinical test.

The argument for a broad form of consent is convincing because a sample's use is rarely known at the point of collection, either by biobank or donor. This presents two problems: that consent is delegated to an oversight committee that is largely invisible outside of the biobank (biobank participants prioritise transparency and the societal interest of research over their own consent [4]), and a lack of choice or control in how data is used [9] or shared. Dynamic consent allows individuals to choose the level of consent they wish to give, it does not have to be competition with broad forms of consent, as "broad" may be a level they feel comfortable with. Underlying technology must focus on enabling a variety of consent options at the point at which a choice is made - individual participants making that decision rather than researchers making it for them.

3 Method

Author A.S. interviewed participants in June/July 2018 who were involved in a study using dynamic consent. These interviews gathered feedback from researchers who had built dynamic consent into their study, and participants who had taken part. The workshop prompts below emerged from initial thematic analysis of a focus group with the research team and 20 participant interviews asking what they thought consent should look like in research, and what their experience had been. Provisional themes of research expectations and consent decisions were put to workshop attendees to prompt discussion around how to practically engage individuals in how their personal data is used for research purposes.

Twelve people attended the IFIP workshop, "An Exploration of Attitudes to Dynamic Consent in Research" in August 2018, a group of postgraduate students from different disciplines, academics, researchers and policy-makers. After an introduction to dynamic consent, attendees split into mixed groups of four and were asked "Do expectations of research differ between researchers and participants?", designed to prompt discussion around expectations of data-use and any other wider research context. Given time to discuss and present this, a second question was put to the group, "What consent decisions surround individuals [as a participant in a research study] and why might they revoke their data?" which aimed to draw out specifc examples of data that consent could be asked for and revoked, as well as the provisions made by those requesting access. Conversation led to debate so the final part of this workshop consisted of open, chaired discussion.

Notes were taken by the chair and collected from group work. These were transcribed into comments and grouped under the prompt they had originally appeared under: "Why do expectations of research differ...?", "What consent decisions surround...?" and "General discussion". Thematic analysis [1] was used to find initial codes that emerged from the comments. These codes were used to search for themes which were then reviewed and defined. The wider narrative for this workshop is to consider the degree to which individuals should be involved in research design, and what that looks like in practice.

4 Results

This workshop was originally designed to validate a study in progress and provided the opportunity for mutual learning amongst participants as to shared difficulties in research practice. Three themes emerged from the discussion: research design, researcher and participant expectation and trust. Table 1 below shows examples of the comments made during the workshop.

4.1 Research Design

Communication was identified as a common problem, with miscommunication caused by differing levels of participant engagement. A point raised here was whether communication informs or misinforms participants - that researchers must be representative in what they are communicating. Further to this, the

Table 1. Example comments from IFIP Summer School workshop for each of the themes drawn out by thematic analysis.

Design	Expectations	Trust
Communication is a common problem	Participant preferences need to have impact	Getting people into a study requires trust
Platforms, or direct contact to engage participants?	Does the participant expect a quick fix?	There is a level of trust as to how data is used
Can (lay)people understand the language used?	People do not assume (research) data will be misused	People hold back if research conflicts with their values

importance of communication method rather than the information being communicated, emerged as a factor in choosing between using platforms and contacting participants directly. In the words of one attendee, "the latter presents more control and the former conveys more complex information". Designing or using a platform also carries recruitment considerations: marketing ("people need to know it exists") and accessibility requirements ("I'd ask whether participants need to be trained to use it?").

Revocation of consent to data sharing must be an option in research whether consent is broad or dynamic, even if participants do not use it. The advantage to using dynamic consent is that it provides an accessible, nuanced mechanism for withdrawal. Research design must account for the withdrawal process. Information needs to be accessible to non-experts and although no specific requirement seemed to be identified, the point was made that there is an obligation to remind individuals of their participation. Communication of research goals in a way that both the team and participants understand is difficult, but designing research process alongside a communication strategy that prioritises accessible, simple and clear participant involvement could start to address this. Two suggestions emerged around the difficulty of mobilising participants at the beginning of a study: being explicit about immediate benefits to encourage participation, and clarifying the project's individual and societal relevance. "Ownership creep" emerged as a problematic phenomenon where the individual or group responsible for a project was (or became) unclear, causing a lack of focus.

4.2 Researcher and Participant Expectations

Long and short term expectations of a study, such as overall time frame and immediacy of feedback, are made explicit at the point of engaging a participant. This lends agency in that any preferences expressed result in action if change is required (one concern raised was that this is not always the case, in the speaker's experience). An example of avoiding a lack of participant agency could be allowing a participant to specify how often they would like feedback from the study. Setting initial expectations and encouraging early engagement mitigates boredom - identified as a reason that participants might revoke consent.

At a fundamental level it is the violation of the individual that presents a problem. Participants' data has value but that value is largely unappreciated by the participants themselves, as are the potential harms of inappropriate data sharing or leakage ("someone's appointment being affected by their employer finding out" or "monetary gain of third parties", two examples of harm). A lack of empirical evidence creates difficulty in setting precedent and managing expectations as to what responsible data practice looks like in research.

4.3 Trust

Recruiting participants to a study relies on trust. The quality of research may rely on the interaction between participant and researcher, so when using technological interventions, to what extent can a platform be trusted? There is a level of trust as to what happens with participants' personal data. Culture and background are important factors in recruiting research participants. Individuals may hold back if the study is supported by or supports a cause that they cannot or will not abide by. An example discussed was Jehovah's Witnesses and studies that might later include or use blood transfusions.

5 Discussion

Table 2 below shows comments categorised by D (research design), E (research expectations) and T (trust) grouped under existing principles of dynamic consent (engagement and revocation over time).

If an individual agrees to "wider use" of their data, this may mean that, as in biobanks, researchers may want to use that data for other purposes. This may mean sharing with third parties or sharing at a later time in some unspecified way. Recent legislative developments such as the European General Data Protection Regulation (GDPR) [8] and Data Protection Act (DPA) [12] in the United Kingdom have highlighted individual data rights to a public audience but do not directly apply to processing data for research purposes, especially if that data is anonymised. The conversation around data-sharing needs to begin

Table 2. Comments grouped under an extended model of dynamic consent.

Engagement	Revocation	Persistence
T1, T2, T3, T4, T5 E1A, E1B, E1C, E1D, E1E, E2B, E3A, E4A, E4B, D1, D2, D3, D5A, D5B, D6A, D7B, D7C, D7D, D8, D10, D11	T4, E3D, E7, E8, D4, D6A, D7A, D11	E1C, E2A, E2B, E4B, E5, D1, D3, D4, DB6, D7D, D7E, D9, D10, D11

at the point of consenting to a study which in the case of biobanks would be the point of consenting to data collection for non-specific research use. As biobanks collect more data and different types (genomic, diagnostic, clinical and lifestyle, for example), oversight will be needed as to who data is shared with and for what purposes.

Communicating risk around data-misuse mitigates threat ahead of time but increased technology use presents cybersecurity challenges in research practice - while data-security concerns are a part of initial research design, there are more obvious priorities such as research goals and methods. Security concerns must be part of the initial stages of research that involves the use of human participants as part of a risk-based approach to data-protection. Knowing where vulnerabilities are likely to be allows research methods to be developed that mitigate them. One purpose for consent procedures is to limit this type of coercion [7], providing evidence of institutional compliance, or "good behaviour".

6 Conclusion

A significant issue in bio-banking is that data about individuals is being collected (or derived from biological samples) for unknown future uses. Not knowing what these uses might be means that researchers may not be able to anticipate the risks associated with their misuse. Dynamic consent is a mechanism that was created to address problems presented by broad forms of consent by focusing on engaging participants with the research process and allowing revocation of consent. From a workshop run at the IFIP Summer School that drew on the experiences of academics, researchers, students and policy-makers themes emerged around research practice, participant expectations, researcher expectations and trust. Many of these issues centre on a need to consider data-use over an extended period of time. For participants, what their involvement looks like and how they or society benefit. For researchers, the relevance of their work and how its impact is communicated to the wider world.

We suggest that the dynamic consent model has wider applications and should be extended to research contexts outside of bio-banking, as its fundamental principles of consent revocation and participant engagement over time provide a longer term view of how data is used and shared by the researchers collecting it. Dynamic consent is a tool that provides evidence of institutional data-protection and accommodates participant autonomy. Examples are needed of projects that use a dynamic form of consent in building research process and procedure that communicates with participants on their own terms, for the greater good.

References

1. Braun, V., Clarke, V.: Using thematic analysis in psychology. Qual. Res. Psychol. **3**(2), 77–101 (2006). https://doi.org/10.1191/1478088706qp063oa
2. Dixon, W., et al.: A dynamic model of patient consent to sharing of medical record data. bmj **348**, g1294 (2014). https://doi.org/10.1136/bmj.g1294
3. Dixon-Woods, M., Ashcroft, R.: Regulation and the social licence for medical research. Med. Health Care Philos. **11**(4), 381–391 (2008). https://doi.org/10.1007/s11019-008-9152-0
4. Hoeyer, K., Olofsson, B.O., Mjörndal, T., Lynöe, T.: Informed consent and biobanks: a population-based study of attitudes towards tissue donation for genetic research. Scand. J. Public Health **32**(3), 224–229 (2004). https://doi.org/10.1080/14034940310019506
5. Javaid, M., et al.: The RUDY study platform – a novel approach to patient driven research in rare musculoskeletal diseases. Orphanet J. Rare Dis. **11**, (2016). https://doi.org/10.1186/s13023-016-0528-6
6. Kaye, J., Whitley, E., Lund, D., Morrison, M., Teare, H., Melham, K.: Dynamic consent: a patient interface for twenty-first century research networks. Eur. J. Hum. Genet. **23**, 141 (2015). https://doi.org/10.1038/ejhg.2014.71
7. O'Neill, O.: Some limits of informed consent. J. Med. Ethics. **29**, 4–7 (2003). https://doi.org/10.1136/jme.29.1.4
8. European Parliament: Regulation (EU) 2016 of the European Parliament and of the Council, on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (2016)
9. Simon, C., et al.: Active choice but not too active: public perspectives on biobank consent models. Genet. Med. **13**(9), 821–831 (2011). https://doi.org/10.1097/GIM.0b013e31821d2f88
10. Teare, H., et al.: The RUDY study: using digital technologies to enable a research partnership. Eur. J. Hum. Genet. **25**(7), 816–822 (2017). https://doi.org/10.1038/ejhg.2017.57
11. Teare, H., Morrison, M., Whitley, E., Kaye, E., Kaye, J.: Towards 'Engagement 2.0': insights from a study of dynamic consent with biobank participants. Digit. Health. **1**, 1–13 (2015). https://doi.org/10.1177/2055207615605644
12. The Stationary Office: Data Protection Act 2018, chapter 12 (2018). http://www.legislation.gov.uk/ukpga/2018/12/pdfs/ukpga20180012en:pdf

Selected Papers

Glycos: The Basis for a Peer-to-Peer, Private Online Social Network

Ruben De Smet(✉), Ann Dooms, An Braeken, and Jo Pierson

Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium
{rubedesm,ann.dooms,an.braeken,jo.pierson}@vub.be

Abstract. Typical Web 2.0 applications are built on abstractions, allowing developers to rapidly and securely develop new features. For decentralised applications, these abstractions are often poor or non-existent.

By proposing a set of abstract but generic building blocks for the development of peer-to-peer (decentralised), private online social networks, we aim to ease the development of user-facing applications. Additionally, an abstract programming system decouples the application from the data model, allowing to alter the front-end independently from the back-end.

The proposed proof-of-concept protocol is based on existing cryptographic building blocks, and its viability is assessed in terms of performance.

Keywords: Online social network · Peer-to-peer · Privacy by design · Privacy

1 Introduction

Privacy on online social media comes in two forms. Platforms generally give plenty of privacy control to platform users, in form of *social privacy*: users can control which friends can access what content. Recently, the Cambridge Analytica scandal [11] proved again the lack of *institutional privacy*: while users can choose with whom of their social connections they share data, the host or institute that takes care of the platform usually has unlimited access to personal data. Privacy enhancing tool (PETs) are developed to counter several privacy issues by technological means.

One category of PETs are privacy-preserving databases, where the database of a service itself takes a responsibility on the data it exchanges. This often relates to P3P, which is a web standard that encodes a service's privacy practices in a machine readable way [22]. An overview of privacy-preserving databases is given in Sect. 3.1.

Another often encountered paradigm in these PETs is moving data away from a central host or institution: decentralisation of services is believed to enhance institutional privacy for its end-users, since the institution itself is taken out of the picture. Several efforts have been made, both academical and community

Published by Springer Nature Switzerland AG 2019
E. Kosta et al. (Eds.): Privacy and Identity 2018, IFIP AICT 547, pp. 123–136, 2019.
https://doi.org/10.1007/978-3-030-16744-8_9

projects, to “re-decentralise” the internet, or parts thereof. In Sect. 3, we enumerate some notable projects that attempt to decentralise online social media.

We argue that at least one problem in this “re-decentralisation” is the lack of abstractions for developers. Where in typical centralised systems developers have tools like SQL (often in combination with object relational mapping (ORM)), or cookies (often as part of an authentication system), decentralised systems are often built “from scratch”, drafting protocols (or extensions thereof) on a per-feature basis.

As an additional consequence, the coupling of the front-end application and the back-end decentralised networking components make it difficult to migrate data, or to fix a security issue in the back-end in a consistent, forward-compatible way.

In Sect. 4 we propose a proof-of-concept protocol for authenticated, confidential data exchange in a peer-to-peer network. This protocol should allow a participant to share with their friends, and stay anonymous for the rest of the network; it offers a form of cryptographically mandatory access control on the data. Since it is based on a peer-to-peer overlay network and therefore has no central processing infrastructure, the system should be lightweight enough to run on constrained devices like smartphones. We evaluate the performance characteristics in Sect. 5.

2 Problem Statement

Many protocols on the internet are federated; examples including email or XMPP. In case of email, Mailchimp (a large email marketing company) notes in 2015 that more than 70 % of their email targets Google’s `gmail.com` domain. Their statistics exclude the “foreign” domains hosted on Google’s and Microsoft’s mail servers, which suggests an even larger market share [13]. Other online resources suggest that both Microsoft and GMail are by far the world most popular email service providers [9,18]. This illustrates that federated networks *may* still lead to centralisation, defeating the decentralisation and privacy-related benefits[1] [17].

The case of email illustrates another drawback of federated systems: the user has to pick a provider. A quick survey on Google, Bing and DuckDuckGo results in `mail.com` and `gmail.com` as top two results, with Microsoft’s `live.com` usually third for the keywords “create email account”.

Another prime example of an attempt at decentralising authentication is OpenID, an open standard and authentication protocol. In practice, users employ a large OpenID provider (often Google or Microsoft), which effectively centralises login history with a few providers.

In the Web 2.0 paradigm, developers employ certain tools (abstractions, SDKs, libraries) that aid the development of their applications. For example,

[1] In case of email, it is enough that just *one* participant in a conversation should be on a malicious server to compromise all communication. PETs such as PGP try to overcome this issue.

SQL (with optional ORM or query builder) is used to store and retrieve data. "Asynchronous Javascript and XML (AJAX)" is used for dynamically changing retrieving information. Cookies (with for example OAuth) are used for authentication and (re-)identification. Similar abstractions can be identified in the mobile app paradigm, building applications for Android or similar.

For decentralised applications, these abstractions are often non-existent, too domain-specific (e.g. Pandora has objects like "Person", "City"), or too low-level: PeerSoN [7] uses files, while RetroShare's GXS [28] uses "groups" containing "messages" as basic building blocks.

The remainder of this paper is concerned with the development of an abstract data model and platform, meant to be at the basis of a peer-to-peer online social network (OSN). It should be portable and efficient enough to run on smartphones, and it should provide a minimum of access-control. These properties fit our interpretation of the privacy definition of Agre and Rotenberg: "The freedom from unreasonable constraints on the construction of one's own identity" [2]. OSNs are important in human social communication; they should facilitate social interaction, and their use and development should not be obstructed by technical difficulties.

3 Related Work

Troncoso et al. have enumerated different properties of decentralised systems. A system can be called decentralised, while in fact certain aspects are still inherently or partially centralised, e.g. trackers and supernodes in BitTorrent or Tor's Directory Authorities [30].

3.1 Privacy-Preserving Databases

The field of privacy-preserving databases is concerned with storing, processing, and releasing data while preserving data. Different database management system (DBMS) have different properties regarding privacy preservation.

Often cited is Platform for Privacy Preferences Project (P3P), a web-based protocol that enables websites to communicate their privacy practices to the browser. The browser interprets and represents this information, and can automatically make decisions based on user preferences [22]. Research is being carried out to develop DBMS that are able to enforce promises encoded in languages such as P3P [5];

Another research area is the development DBMS that allow queries over encrypted data. These systems are typically cloud- or infrastructure-based, as opposed to peer-to-peer. As an example, Cao et al. developed a graph database that supports queries over its encrypted data [8].

3.2 Private Online Social Networks

Efforts for building a decentralised online social network are almost commonplace, with for example Diaspora*, Mastodon, and SecuShare. One notable community project is RetroShare, which is a so-called friend-to-friend network[2] providing file-sharing, fora and other services. In October 2017, Soler published the Generic data eXchange System (GXS) [28], on which they ported RetroShare's fora and newsgroups. The goal of GXS is to make development of new features easier, by providing an abstract layer for developers.

One academic decentralised online social network is called PeerSoN [7]. PeerSoN uses a distributed hash table (DHT) to localise files on a decentralised network. Writing to and reading from those files is subject to mandatory access control (MAC), implemented using cryptography.

A commercial example is MaidSAFE, who are developing a distributed filesystem [12], supported by cryptographic currency [14] based on supernodes.

3.3 Cryptographic Building Blocks

Where centralised applications can rely on the infrastructure granting or denying access, a peer-to-peer system has to rely on cryptography and key management. After all, when data passes through or is stored on unknown or untrusted peers, they should not be able to read it.

The cryptographic currency Monero takes this principle to the extreme: their protocol attempts to hide the sender, receiver and amount of a transaction, while still solving the double-spend problem [25]. Monero relies on a few existing cryptographic building blocks to reach their goal, two of which are at the basis of *Glycos*.

To conceal the sender, Monero uses ring signatures [24]. This allows the real sender to hide himself among a list of potential senders. Additionally, they anonymise the receiver by computing a related but unlinkable receiver key. We use a variant of both schemes in Sect. 4.5, with similar purposes.

4 Solution Design

For both centralised and decentralised applications, providing the developer with abstractions has several benefits. Applications are faster, easier and more secure to develop and to maintain, and the developer does not need knowledge of the underlying systems.

We propose a building block for distributed and private data storage, the equivalent of a DBMS in the classical paradigms, based on graph databases.

4.1 Privacy by Fine-Grained Access Control

Porting privacy definitions to a peer-to-peer setting is anything but trivial, and requires deeper research on its own. An illustration: legislation like the GDPR [23] is concerned with processors and controllers, which both are typically depicted by legal entities that process or control personal data. In a peer-to-peer

[2] Troncoso et al. refer to this as "P2P: Nodes Assist Other Nodes" [30].

setting, where the central authority and institution is taken out of the picture, it becomes difficult to clearly point out who is processor or controller: they all depend on the specifics of the considered peer-to-peer system.

In an overlay network like the one presented, one could say the *whole network* becomes the hosting institution. Institutional privacy thus means privacy with respect to the network's peers.

When we consider the institution as an eavesdropper, some of the properties we want to achieve are:

confidentiality. The overlay network should not learn the semantic meaning of the data it stores.
control. The end-user should control the data he stores on the overlay network.
unlinkability. The overlay network cannot sufficiently distinguish whether two items of interest are related or not [21].

By storing data in a granular way as edges and vertices, and anonymising every data point, we can ensure unlinkability. We will employ well-established cryptographic building blocks to anonymise data, encrypt data, and provide access control.

4.2 Access Controlled Graph Database

A graph database is a database of triplets (s, p, o); a subject s, predicate p, and object o. A triplet (s, p, o) represents the directed edge with label p, from s to o.

We construct a query system wherein vertices and edges are efficiently searchable and traversable for authorised users, while being encrypted, and thus unintelligible for unauthorised users. Data is stored on a DHT based on Kademlia [19]. All vertices have an owner and an (optionally empty) access control list; the owner of a vertex can optionally grant others the right to append additional edges to specific vertices. In Fig. 1, Alice has granted Bob the right to post on her wall.

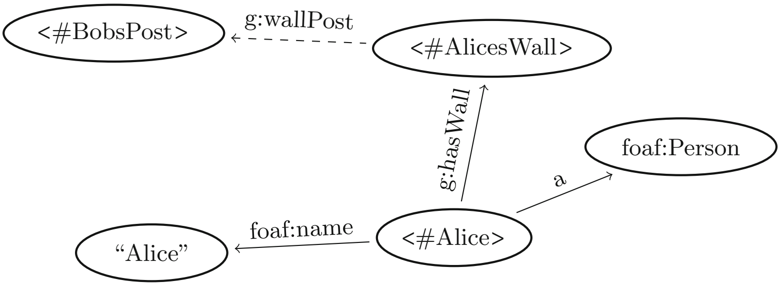


Fig. 1. Bob writes a message on Alice's wall. This is only possible if Alice has granted Bob the rights to do so; otherwise, the network will not accept Bobs post (`<#BobsPost>`). The definition of those access rights are contained within every vertex.

For this paper, we assume a network of trust; users have access to (correct) key information of their peers. Trust models used in PETs include trust-on-first-use (e.g. Signal) and offline key exchange (PGP, OTR).

4.3 Data Model

In a conventional graph database, information does not have access rights; we thus propose a simple[3] access control model that extends an RDF-based model [16]. By splitting up the concept of vertices and edges into two separate objects, it is possible to alter a vertex' content independently from the edge list, and vice versa. It also allows us to add security-related information to both objects.

A vertex s is identified by its owner, and contains an access control list enumerating users that can create edges with s as subject. This allows for typical OSN features like a personal "wall", where Alice's friends can leave posts to be read for her and her friends. Those posts in turn can then contain a comments section, to allow for more interaction.

4.4 Notational Conventions

Since we will be using a few cryptographic concepts, it is necessary to define some notation. The public key system used throughout the design is Curve25519 [4]. Public keys are points on an elliptic curve, and their discrete logarithms are the respective private keys. When ℓ is the size of the underlying field, $r \leftarrow [0, \ell - 1]$ picks a field element uniformly at random. We will assume a known long term public key pk^i_{LT} with corresponding private key sk^i_{LT} for every participant i. Identification can happen in a face-to-face meeting: we assume that two persons that want to use the network together have access to correct key information of each other.

$\mathcal{H}$ is a cryptographically secure hash function (the Keccak-based [6] SHA 3-256 function), and $\mathcal{H}_s$ is a hash function onto the underlying field of the elliptic curve.

4.5 Implementation

We will store every vertex s of the graph database as an object in a Kademlia-based [19] DHT, storing vertices that have s as subject alongside s for easy graph traversal. The DHT thus understands two kinds of `put` operations; one for vertices and one for edges, while the `get` operation returns both the vertex and the associated edges.

Note that since vertices are identified by their owner, we cannot use the long term $pk^{\mathrm{i}}_{\mathrm{LT}}$ public key. Instead, inspired on ByteCoin's "Stealth Addresses"

[3] This model is "simple" in the sense that more complex models are possible, and may be interesting for future research: co-ownership, write-only, or read-only rights can all have useful applications.

[29,31] and Monero's "one-time addresses" [25], we *derive* a random key from the long term key $pk^{\mathrm{i}}_{\mathrm{LT}}$:

Algorithm 1 (Generate an ephemeral public key). Given the public key $A = aG = pk^{\mathrm{alice}}_{\mathrm{LT}}$ of Alice, Bob generates an ephemeral (*o*ne-*t*ime) public key for Alice as follows:

$$\begin{aligned} r &\leftarrow [0, \ell - 1] \\ R &\leftarrow rG, \\ pk^{\mathrm{alice}}_{\mathrm{OT}} &\leftarrow \mathcal{H}_s(rA)G + A, \\ sk^{\mathrm{alice}}_{\mathrm{OT}} &\leftarrow \mathcal{H}_s(aR) + a. \end{aligned}$$

This key clearly belongs to Alice: only Alice knows the integer a required to construct her secret key. She can recognise that this key belongs to her by checking whether A' equals $pk^{\mathrm{alice}}_{\mathrm{OT}}$ in

$$A' = \mathcal{H}_s(aR)G + A.$$

Due to this property, we will call R the "recogniser". Note that the serialisation of the ephemeral public key together with the recogniser only takes the size of two points (R and $pk^{\mathrm{alice}}_{\mathrm{OT}}$), which is 64 bytes when using Curve25519 [4]. Since r is only used in the key derivation, it is a temporary variable. Note that $\mathcal{H}_s(rA) = \mathcal{H}_s(aR)$ is an elliptic-curve Diffie-Hellman key agreement [10,20] with a random key $R = rG$.

The ephemeral public key $pk^{\mathrm{alice}}_{\mathrm{OT}}$ is indistinguishable from random. Formally, the probability distributions of (R = rG, A = aG, $pk^{\mathrm{alice}}_{\mathrm{OT}}$) and ($R = rG, A = aG, C = cG$) are computationally indistinguishable for r, a, c chosen randomly and uniformly from $[0, \ell - 1]$.

Proof. Assume we can distinguish (R = rG, A = aG, $pk^{\mathrm{alice}}_{\mathrm{OT}}$) and ($R = rG, A = aG, C = cG$) using some distinguisher $\mathcal{A}$. This means we can solve the decisional Diffie-Hellman problem: to distinguish ($R, A, K = rA = aR$) and (R, A, C), it suffices to run $\mathcal{A}$ on ($R, A, \mathcal{H}_s(K)G + A$) and ($R, A, C$). □

We can now define a vertex.

Definition 1 (vertex). *A vertex V is a 7-tuple* $(O, R, ACL, R_{ACL}, v, c, S)$.

The key O is an ephemeral, unique public key derived from the private key held by the owner of this vertex, the *owner key*. The point R is the recogniser used to generate key O. The list ACL is the *access control list*, listing all ephemeral public keys that are allowed to link other vertices from this vertex using edges. The point R_{ACL} is the recogniser used to generate all ephemeral public keys in ACL. Optionally, v is the encrypted associated value or content of the vertex. The *clock* c is a positive integer to keep track of the vertex version. The *Schnorr signature* S is a $\mathcal{BNN} - \mathcal{IBS}$ signature [3,26,27] of (R, ACL, R_{ACL}, v, c) generated using O.

In this definition, the access control list ACL contains ephemeral public keys generated with a common r, thus having the common recogniser $R_{ACL} = rG$. This operation effectively anonymises vertex appends, while the assigned users can still recognise (using R_{ACL}) their eligibility to create edges.

By using Algorithm 1 to generate O and the keys in ACL, these public keys are indistinguishable from random and thus *unlinkable to their owners*.

There is still one problem to overcome: imagine we use the above (Schnorr) signature to sign an edge. The signer is always identifiable, and Eve—the eavesdropper—could distinguish edges based on their associated signer. Eve should only learn about the *validity* of the edge. In "How to leak a secret" [24] Rivest, Shamir, and Tauman describe an elegant concept and method to overcome this issue. They propose a so-called "ring-signature", a signature which proves knowledge of one secret key of a set, without revealing which.

A ring-signature scheme based on elliptic curves is documented by Abe, Ohkubo, and Suzuki [1, Appendix A].

Algorithm 2 (Generate ring signature). A signer with secret key x_k signs message m with public-key list $\mathcal{R}_s = Y_0, Y_1, \ldots, Y_{n-1}$

1. Select $\alpha, c_i \leftarrow [0, \ell - 1]$ for $i = 0, \ldots, n-1, i \neq k$, and compute $z = \alpha G + \sum_{i=0, i\neq k}^{n-1} c_i Y_i$
2. Compute

$$c = \mathcal{H}_s(\mathcal{R}_s || m || z)$$

$$c_k = c - \sum_{i=0, i\neq k}^{n-1} c_i \mod q$$

$$s = \alpha - c_k x_k \mod q$$

3. Return $\sigma = (s, c_0, \ldots, c_{n-1})$

Algorithm 3 (Verify ring signature). A verifier verifies signature σ. (L, m, σ) by checking whether

$$\sum_{i=0}^{n-1} c_i \cong \mathcal{H}_s\left(\mathcal{R}_s || m || \left(sG + \sum_{i=0}^{n-1} c_i Y_i\right)\right) \mod q$$

An edge can now be defined as an object with an encrypted value, pointing from a subject to an object, with the value and identifier of the object being encrypted:

Definition 2 (edge). *An edge E between two vertices $V_s = (O_s, R_s, ACL_s, R_{ACL,s}, v_s, c_s, S_s)$ (the subject) and $V_o = (O_o, R_o, ACL_o, R_{ACL,o}, v_o, c_o, S_o)$ (the object) is a 5-tuple*

$$E = (O_s, ACL_E, K_{ACL}, \mathcal{E}_k(l, O_o), \mathcal{R}_s, S).$$

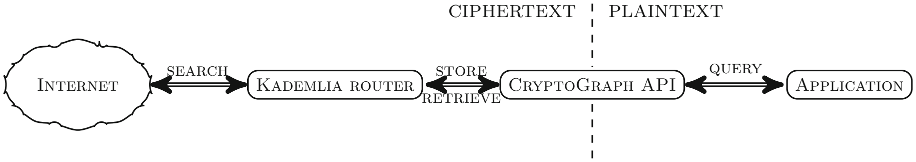


Fig. 2. The two main components implemented are the middleware for the cryptographic graph, and the custom Kademlia router. The Kademlia router is responsible for connection with other peers, storing and retrieving encrypted graph data on the network. The graph API implements the encryption and decryption of the graph, feeding it from and back to the application.

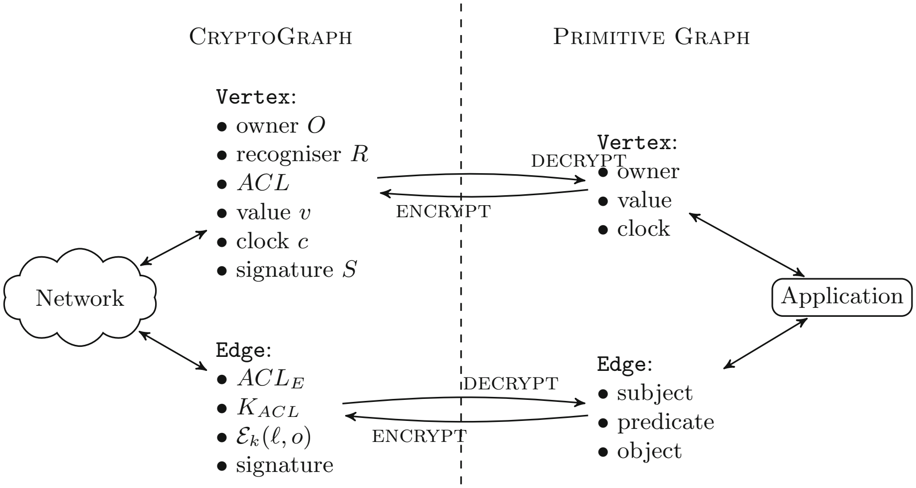


Fig. 3. The graph API. Clear text operations are clearly separated from cipher text domain operations, and conversion between the two domains happens through an explicit encrypt or decrypt method.

The *ring* $\mathcal{R}_s = \{O_1, O_2, \ldots, O_i = O_s, \ldots, O_n\}$ is a set of public keys containing all n public keys in ACL_s. For every key in ACL_s, ACL_E contains the encrypted key k. It is encrypted n times using a standard hybrid encryption, based on a Diffie-Hellman exchange with the random point K_{ACL} using the symmetric cipher $\mathcal{E}$. l is the *label* of the edge. S is a ring signature [1, Appendix A] over the ring $\mathcal{R}_s$ of $(O_s, ACL_E, \mathcal{E}_k(l||O_s))$. The label and object are encrypted using the same symmetric cipher $\mathcal{E}$ with key k.

```java
final Profile p = new Profile();
p.setName("Alice Cryptographer");

// Save `p' on network `connection' with owner `privateKey'
ID profile_id = connection.pushProfile(p, privateKey);

connection.findProfile(profile_id, privateKey,
        new FetchEventListener<Profile>() {
    @Override
    public void onComplete(Profile profile) {
        // Do something with the found and decrypted profile
    }
});
```

Listing 1. The Android-compatible Java library is used in this example to create a `Profile` object, assign a name, store it on the network and then asynchronously fetch it from the network.

5 Performance Evaluation

To validate the technical viability, we have built a demonstrator implementation in Rust [4]. We call this demonstrator *Glycos*, and serves as a middleware providing an interface for traversing the graph with an asynchronous API. It contains the necessary networking and cryptographic components to query the network for, and to create and store vertices and edges. For a graphical overview, refer to Figs. 2 and 3.

Additionally, it contains an object relational mapping ORM interface that maps objects to vertices and edges and vice versa. This allows a developer to think in terms of objects and their relations, like is common when working with relational databases. The ORM-interface contains generated bindings to Java, to demonstrate the viability on the Android platform. Listing 1 contains example code verified testing on both a virtual and a physical Android device.

Since practicality and performance are key in the design, a thorough analysis of both aspects is mandatory. Note that a vertex can be serialised in $156 + |v| + 32|ACL|$ bytes when taking a 64-bit integer for the clock value. We saved 32 bytes by using the $\mathcal{BNN} - \mathcal{IBS}$ *Schnorr signature* [3] scheme, which allows us to omit the owner key from the serialisation.

When an edge E is transmitted together with its accompanying vertex, we can omit the subject O_s and the ring $\mathcal{R}_s$ from Es serialisation. This allows a serialisation in $112 + |l| + 80|R_s|$ bytes.

The maximum transmission unit (MTU) for Ethernet is about 1500 bytes, so for a small (< 15) amount of participants both a vertex or an edge could fit a single Ethernet frame. At 1500 bytes, one megabyte can store around 700 vertices or edges, one gigabyte around 700 000. Since modern smartphones and computers

[4] "Rust is a systems programming language that runs blazingly fast, prevents segfaults, and guarantees thread safety." https://www.rust-lang.org/.

come with plenty of storage, often exceeding 8 GB, this is ought to be compact enough (Table 1).

Table 1. Specifications of the devices used for benchmarks. All benchmarks were ran on the notebook, except where otherwise noted.

	Notebook	Smartphone
Brand	Lenovo Thinkpad X250	Lenovo Moto Z Play
CPU	Intel Core i5-5200U (Broadwell)	ARM Cortex A53 (MSM8953)
Core count	2 cores, 4 threads	8 cores
Clock frequency	2.20 GHz	2.0 GHz
RAM	16 GB DDR 3 at 1600 MT/s	3 GB DDR 3
Operating system	Arch Linux	SailfishOS 2.1.3.7 armv7hl
Rust compiler	1.29.0-nightly (`874dec25e` 2018-07-21)	

We ran a few benchmarks to measure how fast vertices and edges can be generated and decrypted. Timings correspond to a at least few hundreds of encryptions and decryptions per second. Note that at the time of writing, the ARM based smartphone platform takes no advantage of the available NEON instruction set[5] nor from the 64 bit instructions. In other words, the ARM build has still room for optimisation. All notable benchmarks are represented in Table 2.

Table 2. Notable benchmarks. Ring size is taken to be $|\mathcal{R}_s| = 2$ where applicable. A "seal" operation consists of encrypting and signing the vertex or edge (cfr. Fig. 3). An "open" operation is the inverse operation: computing the correct keys and decrypting the vertex or edge. Mean times as reported by the `criterion` library.

	Notebook	Smartphone
verify vertex signature	136.51 µs	2.8427 ms
verify edge signature	157.58 µs	3.0733 ms
"seal" vertex	948.97 µs	13.458 ms
"seal" edge	438.15 µs	8.4213 ms
"open" vertex	391.11 µs	7.6648 ms
"open" edge	129.53 µs	2.5662 ms

[5] NEON support is on the roadmap for `curve25519-dalek`; cfr. https://github.com/dalek-cryptography/curve25519-dalek/issues/147.

6 Conclusion

Decentralisation of a service is believed to lead to more privacy. We noted that today's decentralised online social network (OSNs) come in two forms: at one hand there are federated OSNs, and at the other there are peer-to-peer OSNs. Federated networks have as disadvantage that the end-user has to choose a provider or "pod", which in the case of e-mail has lead to *re*-centralisation of users' data.

Most peer-to-peer networks reinvent the wheel: often on a per-feature basis, these systems mainly design a private and secure protocol. This is in contrast with centralised services, where developers employ abstractions like SQL, ORM, and cookies to build applications, often without having to consider cryptography.

An abstract data model can help to overcome this unbalance. While existing data models such as GXS [28] have also observed this unbalance, proposed solutions are often still application specific. We propose a simple graph database-like service built upon Kademlia, on which application developers can store and query arbitrary data. This data model is encrypted and authenticated and thus only readable and writeable by users with the necessary permissions. Moreover, it has been made relatively easy to use through the ORM layer, and shown to be efficient enough to run on mobile devices.

7 Future Work

In the current model, efficient update and delete operations are still lacking, due to the risk of replay attacks. By introducing a notion of time, or more precisely the notion of *happened-before* [15], these attacks can be countered, and efficient deletion could be implemented. These are important considerations, since these features would increase the user's *control* over their data.

As touched upon in Sect. 4.1, privacy properties and definitions are not well studied in a peer-to-peer context. Formally identifying adversaries and their capabilities in a peer-to-peer OSN, and making provable definitions about them can increase confidence in these applications.

Looking at *Glycos* as a middleware, future research should further enhance the platform in itself, making it more practical to build actual applications and to make peer-to-peer overlay systems simpler to develop.

References

1. Abe, M., Ohkubo, M., Suzuki, K.: 1-out-of-n signatures from a variety of keys. In: Zheng, Y. (ed.) ASIACRYPT 2002. LNCS, vol. 2501, pp. 415–432. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-36178-2_26
2. Agre, P.E., Rotenberg, M.: Technology and Privacy: The New Landscape. MIT Press, Cambridge (1998)
3. Bellare, M., Namprempre, C., Neven, G.: Security proofs for identity-based identification and signature schemes. J. Cryptol. **22**(1), 1–61 (2009)

4. Bernstein, D.J.: Curve25519: new Diffie-Hellman speed records. In: Yung, M., Dodis, Y., Kiayias, A., Malkin, T. (eds.) PKC 2006. LNCS, vol. 3958, pp. 207–228. Springer, Heidelberg (2006). https://doi.org/10.1007/11745853_14. ISBN 978-3-540-33852-9
5. Bertino, E., Byun, J.-W., Li, N.: Privacy-preserving database systems. In: Aldini, A., Gorrieri, R., Martinelli, F. (eds.) FOSAD 2004-2005. LNCS, vol. 3655, pp. 178–206. Springer, Heidelberg (2005). https://doi.org/10.1007/11554578_6
6. Bertoni, G., et al.: Keccak sponge function family main document. In: Submission to NIST (Round 2), vol. 3, p. 30 (2009)
7. Buchegger, S., et al.: PeerSoN: P2P social networking - early experiences and insights. In: Proceedings of the Second ACM Workshop on Social Network Systems Social Network Systems 2009, co-located with Eurosys 2009, Nüurnberg, Germany, March 2009, pp. 46–52 (2009)
8. Cao, N., et al.: Privacy-preserving query over encrypted graph-structured data in cloud computing. In: 2011 31st International Conference on Distributed Computing Systems (ICDCS), pp. 393–402. IEEE (2011)
9. Datanyze: Email Hosting Market Share Report. Datanyze, 12 June 2018. https://www.datanyze.com/market-share/email-hosting. Accessed 14 June 2018
10. Diffie, W., Hellman, M.: New directions in cryptography. IEEE Trans. Inf. Theory **22**(6), 644–654 (1976)
11. Graham-Harrison, E., Cadwalladr, C.: Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. In: The Guardian, March 2018. https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election
12. Irvine, D.: MaidSafe Distributed File System. Technical report (2010)
13. Khan, O.: Major email provider trends in 2015: Gmail's Lead in- creases. Mailchimp, 15 July 2015. https://blog.mailchimp.com/major-email-provider-trends-in-2015-gmail-takes-a-really-big-lead/
14. Lambert, N., Ma, Q., Irvine, D.: The Decentralised Network Token. Technical report MaidSafe, Technical report, Safecoin (2015)
15. Lamport, L.: Time, clocks and the ordering of events in a distributed system. Commun. ACM **21**(7), 558–565 (1978)
16. Lassila, O., Swick, R.R.: Resource Description Framework (RDF): Model and Syntax. W3C Recommendation. W3C (1997). https://www.w3.org/TR/WD-rdf-syntax-971002/. Accessed 20 Oct 2017
17. Lewis, S.J.: On emergent centralization (2018). https://fieldnotes.resistant.tech/defensive-decentralization/. Accessed 31 Oct 2018
18. Lewkowicz, K.: Here's What We Learned After Tracking 17 Billion Email Opens [Infographic], 21 March 2017. https://litmus.com/blog/2016-email-client-market-share-infographic. Accessed 14 June 2018
19. Maymounkov, P., Mazières, D.: Kademlia: a peer-to-peer information system based on the XOR metric. In: Druschel, P., Kaashoek, F., Rowstron, A. (eds.) IPTPS 2002. LNCS, vol. 2429, pp. 53–65. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-45748-8_5
20. Miller, V.S.: Use of elliptic curves in cryptography. In: Williams, H.C. (ed.) CRYPTO 1985. LNCS, vol. 218, pp. 417–426. Springer, Heidelberg (1986). https://doi.org/10.1007/3-540-39799-X_31
21. Pfitzmann, A., Hansen, M.: A terminology for talking about privacy by data minimization: anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management (2010)

22. Platform for Privacy Preferences (P3P) Project. W3C Recommendation. W3C, February 2014. https://www.w3.org/P3P/. Accessed 31 Oct 2018
23. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 27 April 2016
24. Rivest, R.L., Shamir, A., Tauman, Y.: How to leak a secret. In: Boyd, C. (ed.) ASIACRYPT 2001. LNCS, vol. 2248, pp. 552–565. Springer, Heidelberg (2001). https://doi.org/10.1007/3-540-45682-1_32
25. Van Saberhagen, N.: Cryptonote v2.0 (2013)
26. Schnorr, C.-P.: Efficient signature generation by smart cards. J. Cryptol. **4**(3), 161–174 (1991)
27. Schnorr, C.-P.: Method for identifying subscribers and for generating and verifying electronic signatures in a data exchange system. U.S. pat. 4995082, February 1991
28. Soler, C.: A Generic Data Exchange System for Friend-to-Friend Net- works. Technical report, INRIA Grenoble-Rhone-Alpes (2017)
29. Todd, P.: [bitcoin-development] Stealth addresses (2014). https://www.mail-archive.com/bitcoin-development@lists.sourceforge.net/msg03613.html. Accessed 12 Feb 2017
30. Troncoso, C., et al.: Systematizing decentralization and privacy: lessons from 15 years of research and deployments. In: Proceedings on Privacy Enhancing Technologies, vol. 2017, no. 4, pp. 404–426 (2017)
31. user ‘bytecoin’. Untraceable transactions which can contain a secure message are inevitable (2011). https://bitcointalk.org/index.php?topic=5965.0. Accessed 12 Feb 2017

GDPR and the Concept of *Risk*:

The Role of Risk, the Scope of Risk and the Technology Involved

Katerina Demetzou(✉)

Business and Law Research Centre (OO&R),
Radboud University, Nijmegen, Netherlands
K.Demetzou@cs.ru.nl

Abstract. The prominent position of *risk* in the GDPR has raised questions as to the meaning this concept should be given in the field of data protection. This article acknowledges the value of extracting information from the GDPR and using this information as means of interpretation of *risk*. The 'role' that *risk* holds in the GDPR as well as the 'scope' given to the concept, are both examined and provide the reader with valuable insight as to the legislature's intentions with regard to the concept of *risk*. The article also underlines the importance of taking into account new technologies used in personal data processing operations. Technologies such as IoT, AI, algorithms, present characteristics (e.g. complexity, autonomy in behavior, processing and generation of vast amounts of personal data) that influence our understanding of *risk* in data protection in various ways.

Keywords: Risk · Concept · Data protection · Accountability · Compliance · Role · Scope · Fundamental rights · New technologies

1 Introduction

The GDPR[1] is the new EU legal framework for the fundamental right of personal data protection and for the free flow of personal data, which repeals the preceding Directive 95/46.[2] While the GDPR preserves the key concepts and the basic data protection principles, it introduces several novelties which aim to achieve an effective and high level protection of personal data. One such novelty is the prominent position[3] of the concept of *risk*, both in terms of forming and in terms of triggering legal obligations. *Risk* forms legal obligations in the sense that it has become one of the criteria that the data controller should take into account when deciding on the most appropriate technical and organizational measures;[4] *risk* triggers a legal obligation in the sense that if

[1] See [2].

[2] See [1].

[3] This prominent position of the concept of *risk* has led legal scholars to talk about a 'riskification' of the EU data protection legislation. See [49]. Also [43].

[4] See the general legal obligation in Article 24(1) GDPR.

Published by Springer Nature Switzerland AG 2019
E. Kosta et al. (Eds.): Privacy and Identity 2018, IFIP AICT 547, pp. 137–154, 2019.
https://doi.org/10.1007/978-3-030-16744-8_10

there is no risk then the obligation need not be fulfilled.[5] According to the WP29, "A risk is a scenario describing an event and its consequences, estimated in terms of severity and likelihood".[6] The fact that consideration should be given to both 'likelihood' and 'severity' is also mentioned in Recitals 75 and 76 of the GDPR.[7] The use of the concept of *risk* is part of the approach adopted by the European legislature in the GDPR, towards a more proactive, scalable and effective data protection.

Despite its importance, *risk* lacks a legal qualification under the EU general legal framework on personal data protection (GDPR). The legislature provides us with examples of *risks* (e.g. in Recital 75, 91 GDPR) but does not give the tools for assessing other (new) types of risks, their severity and their likelihood, in an objective and consistent way. The legal qualification of *risk* in relation to data protection and the provision of objective legal criteria against which 'likelihood' and 'severity' will be measured, will allow data controllers to examine each processing activity and reach reliable and contestable conclusions as to the (high) risk(s) presented.

The GDPR, the legal framework in which the concept of *risk* is introduced, should constitute a major source of extraction of legal criteria that will be used as means of interpretation of the concept of *risk* in data protection. Understanding the meaning of *risk* in relation to the particular characteristics of the GDPR presents two important benefits; firstly, it adds objectivity to the assessment of *risk(s)*, in that it allows for the use of language and means, that those involved in the field of data protection share and understand. The legislature requires such an objective assessment of *risk* in Recital 76 GDPR.[8] Secondly, it provides for an interpretation which "guarantee[s] that there is no conflict between it and the general scheme of which it is part".[9] This approach has also been encouraged by the WP29 in its Opinion[10] on the concepts of 'controller' and 'processor'. The WP29 thereby highlighted the need to deal with the concept of 'data controller' as an autonomous concept, meaning that it has "its own independent meaning in Community law, not varying because of -possibly divergent- provisions of national law" (see footnote 10) and that "although external legal sources can help identifying who is a controller, it should be interpreted mainly according to data protection law".[11] This "uniform and therefore autonomous interpretation of such a key concept" contributes to an effective application of data protection rules and to a high level of protection (see footnote 10).

[5] See for example the legal obligation in Article 35(1) GDPR to perform DPIAs where there is 'high risk'.

[6] See [18], 6, [16], 7: "severity and likelihood of this risk should be assessed".

[7] Recital 75 GDPR "The risk to the rights and freedoms of natural persons, of varying *likelihood and severity* […]",
Recital 76 GDPR "The *likelihood and severity* of the risk to the rights and freedoms of the data subject […]".

[8] Recital 76 GDPR "Risk should be evaluated on the basis of an objective assessment, by which it is established whether data processing operations involve a risk or a high risk".

[9] See [41], 14.

[10] See [14], 8.

[11] See [14], 9.

The purpose of this article is to identify the elements in the GDPR that should be taken into account and explain the way they inform the interpretation of the concept of *risk* in data protection. The research question of this article is, thus, formed as follows:

> "How do the role and the scope of *risk* in the GDPR as well as the technology involved in data processing operations inform the meaning of *risk* in the field of data protection?"

To answer this research question, I will first look into the role that the European legislature has attributed to *risk*, by relating it to the principle of accountability and the approach that this principle brings in the GDPR [Sect. 2]. The discussion on the role of the concept of *risk* leads to the conclusion that *risk* in the GDPR should not be understood as a '(non) compliance risk'. *Risk* should, on the contrary, be understood as referring to 'the rights and freedoms of natural persons' as explicitly suggested by the legislature's wording. In Sect. 3, I will examine the (broad) scope of the concept of *risk*. In Sect. 4, I will discuss the technology involved in data processing operations. My purpose is to show that the concept of *risk* in data protection is highly influenced by the technology used. New technologies and the particularities they present should be taken into consideration when interpreting the concept of *risk* in data protection. In the Conclusion [Sect. 5] I answer the Research Question and summarize the findings of this article.

2 The Role of *Risk* in the GDPR: Accountability and Risk-Based Approach

As mentioned in the Introduction, the concept of *risk* has been given a prominent position in the GDPR. Having acknowledged this legislative choice, the following question is raised: What is the role of *risk* in the GDPR? What should always be kept in mind when interpreting and applying data protection rules, is that the ultimate purpose of these rules is "to protect fundamental rights and freedoms of natural persons and in particular their right to privacy, with regard to the processing of personal data".[12] On top of that, what should also be kept in mind is the exact role of a concept in the given legal framework.

The importance of clarifying the role of a concept in order to interpret it in a way aligned with its role, could become apparent via the example of 'personal data'.[13]

The Example of 'Personal Data'

'Personal data' is a concept that relates to the material scope of the data protection legal framework. The material scope determines the conditions under which a case falls under the legal framework and natural persons benefit from the legal protection it offers. In the *Google Spain* case, the CJEU said that 'the provisions of Directive 95/46

[12] See [13], 4.

[13] Article 4(1) GDPR: 'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;".

[…] must necessarily be interpreted in the light of fundamental rights'.[14] In its *Ryneš case,* the CJEU said that "derogations and limitations in relation to the protection of personal data must apply in so far as is strictly necessary"[15] which, *a contrario,* confirms the intention of giving a broad meaning to concepts that relate to the material scope of the GDPR.

Based on the role that this concept holds, being actually the 'gateway' for a natural person to get the protection offered by the data protection legislation, the European legislature has provided for a definition of 'personal data' that is 'extensive',[16] flexible enough so that it adapts to the new technological context[17] and has a general wording "so as to include all information concerning an identifiable individual".[18] When interpreting the concept of 'personal data', the approach is objective[19], meaning that the data controller's intention (or knowledge of whether the data that are processed qualify as personal data)[20] does not matter; as long as the data can likely result in the identification of a natural person, data protection law applies.[21] Also, the approach is factual, in the sense that in order that data qualify as personal data, the specifics and the context of each case is what should be examined ("the extent to which certain identifiers are sufficient to achieve identification is something dependent on the context of the particular situation").[22] An example is provided by the CJEU in the case of *Nikolaou v. Commission,* whereby even if the applicant was not named, the information published in the press release was personal data given that the applicant was easily identifiable "under the circumstances".[23] What we acknowledge from this example is that the particular role that the concept of 'personal data' plays in the system of protection of the fundamental right to data protection, is of tantamount importance to the way that this concept is interpreted.

The Role of Risk

Coming back to the concept of *risk*, I will discuss its role in data protection by referring the concept to the principle of accountability. The principle of accountability is found in Article 5(2) which reads: "The controller shall be responsible for, and be able to demonstrate compliance with, paragraph 1 ("accountability"). The controller shall be responsible for *ensuring compliance* and shall be able to *demonstrate compliance* with data protection principles in practice. The 'data controller'[24] is the actor that bears the

[14] See [3], para 68 (Also see, [5], para 37, and [6], para 68).

[15] See [7], para 28 (Also see, [11], para 39, and [8], para 52).

[16] See [21], 10.

[17] See [46].

[18] See [21], 9.

[19] Which, as mentioned by Purtova [46], was also followed in the Breyer case [9].

[20] Check [4], para 72 ("The fact that their character as personal data would remain "unknown" to internet search engine provider, whose search engine works without any human interaction with the data gathered, indexed and displayed for search purposes, does not change this finding").

[21] See [16], 10.

[22] See [13], 13.

[23] See [10], para 222.

[24] Article 4(7) GDPR.

responsibility to process personal data in accordance with the principles established in EU data protection law. It is a concept that plays a crucial role in data protection, since it "determines who shall be responsible for compliance with the data protection rules, and how data subjects can exercise their rights in practice".[25] In other words it helps in allocating responsibility. The legislature's intention is to "stimulate controllers to put into place proactive measures in order to be able to comply with all the elements of data protection law".[26]

While accountability is not a novel concept in data protection[27], the shift[28] that we acknowledge in the GDPR is about the way that the legislature has chosen to set up a modified compliance scheme by, *inter alia*, materializing the accountability principle via a general obligation in Article 24[29] and via more specific obligations (e.g. DPOs, DPIAs, Privacy by design) all of which have a common characteristic: they all suggest specific measures and mechanisms which establish a proactive approach,[30] facilitate the implementation of accountability and therefore enable compliance and its demonstration thereof. They do not add any new principles; instead, they serve as mechanisms for the effective implementation of the already existing data protection principles.[31] Accountability and the compliance scheme as shaped by the legislature could be considered as a legal strategy "for defending what has been formally recognized".[32] And that is the data protection principles in Article 5. It has been argued that these principles constitute the "essence" of the fundamental right to data protection, in the meaning of Article 52(1) of the Charter of Fundamental Rights.[33] If this is the case, then the principle of accountability, whose main goal is a more effective data protection in practice, along with all the mechanisms and measures that enable it, relates to the scope of the fundamental right to data protection in that it requires that all data protection principles should be respected. In that way it actually enhances the rights-based approach and the fundamental rights character of the data protection legal framework.

Article 24, which constitutes the general obligation that materializes the principle of accountability, stipulates that the technical and organizational measures taken by the controller should be dependent on the "nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for the rights and

[25] See [14], 2.

[26] See [28], 22.

[27] It first appeared as a basic data protection principle in the OECD Guidelines. See [29].

[28] Alessandro Spina has also talked about a transformation in the GDPR, which is about an "enforced self-regulation model for managing technological innovation in uncertain scenarios", in Spina [49].

[29] As has been pointed out by the EDPS [27]. "Article 24 refers to the implementation of all data protection principles and the compliance with the whole of the GDPR", para 25.

[30] See [30].

[31] See [20], pp. 2,6; [30], 27.

[32] See [38], 3.

[33] Joined cases *Digital Rights Ireland and Seitlinger and Others* [8], para 40. See also [42]; EDPS [27], para 30.

freedoms of natural persons".[34] The legislature introduces *risk* as a criterion for "the determination of the concrete measures to be applied".[35] This is a choice that adds scalability when it comes to compliance, in the sense that the scope of the legal duties of data controllers depends on the risk posed by their processing operations,[36] and more specifically the *likelihood and severity* of that risk. Scalability is inextricably linked to the principle of accountability,[37,38] in that the latter is "implemented through scalable obligations".[39] Putting *risk* in the spearhead of the compliance scheme as a way to implement the principle of accountability, does not and should not in any way alter the scope of the fundamental right. What it alters is the scope of the legal duties of data controllers, since these become dependent on the risks presented by the specific processing operations. Therefore, *risk* is a concept used in order to enable accountability, in that, by adding scalability to legal obligations, it allows for a more effective protection of personal data. It should be thus understood as a major criterion that belongs to the accountability principle and the proactive approach it establishes. Accountability has been characterised as a "fundamental principle of compliance".[40] That leads us to the conclusion that risk is a concept inextricably linked to the principle of accountability, and thus to compliance. *Risk* and compliance are in this way deeply interconnected.[41]

The 'Risk-Based' Approach

Because of the prominence of *risk* in the GDPR as well as the acknowledgment of a shift of approach with *risk* being the point of reference, the so called "risk based approach"[42] has captured the attention of legal scholars.[43] The debate on the relationship between the rights-based and the risk-based approach, was resolved by the WP29 which stated that the risk-based approach is a "scalable and proportionate approach to compliance" instead of an "alternative to well-established data protection rights and principles".[44] This statement makes us think twice on whether a debate between the rights-based and the risk-based approach, actually exists. Based on the line of arguments in the previous Subsection,[45] not only should we not consider these two approaches as opposing, but, on the contrary, we should understand the risk-based approach as a strategy for the enhancement of the rights-based character of the legal

[34] Also Recital 74 GDRP, mentions that measures of controllers should take into account the risk to the rights and freedoms of natural persons.

[35] See [15], 2.

[36] See [17].

[37] See [15].

[38] This has also been upheld by the EDPS [28], para 104.

[39] See [23], 3.

[40] See [30], 19.

[41] See [35].

[42] The WP29 itself has also published a Statement on the risk-based approach of the GDPR: See [17].

[43] See [48]; [43]; [34].

[44] See [17].

[45] Subsection on The role of *risk*.

framework. The risk-based approach is integrated into the rights-based nature of the GDPR. They are not strictly separated and thus we should not follow a linear scheme whereby, first, full legal compliance takes place (in line with the rights-based character of the framework) and on top of that risk calculations are done (in line with the risk-based approach).[46]

By understanding *risk* as a concept inextricably linked to the principle of accountability and thus to compliance, we come to some valuable conclusions. First of all, relating the risk-based approach to the principle of accountability provides for a firm legal justification of *risk* as a criterion for the scalable and proportional approach to compliance. Secondly, we explain the modified compliance scheme introduced in the GDPR and we understand it as a way of enhancing the fundamental rights nature of data protection. This leads us to the third point which is the clarification of the relationship between the risk-based approach and the rights-based approach. The risk-based approach is not stand-alone. On the contrary, it is an expression of the principle of accountability[47] and is integrated into the rights-based approach, which is supposed to enhance.

Following the previous line of argumentation and based on the conclusions made, we can come to a further conclusion; we can draft away from the position that *risk* in the GDPR should be understood as a compliance (with the GDPR) risk.[48] The legislature seems to be trying to mitigate such a 'non-compliance risk', by adopting a different approach (which I discussed in the Subsection on 'The role of *risk*') that introduces *risk* as a major criterion for more effective compliance. If we understood, under this scheme, risk as non-compliance risk, we would then run into a circle. Furthermore, if compliance with the GDPR meant that risks to rights and freedoms are reduced to an acceptable level, then the question that is raised is why then having *risk* as a concept inextricably linked to compliance, in the first place?

3 The Scope of *Risk* in the GDPR: Risks to the "Rights and Freedoms of Natural Persons"

In the previous Section, we saw that the role, the meaning and the scope of each concept in the GDPR are highly interlinked. This interpretation has relied primarily on the legislature's broad wording which has been upheld and further elaborated by the WP29 and the CJEU. However, in the case of *risk* we do not have such clear guidance yet. Therefore, we need to carefully examine the legislature's wording which in combination with the role that *risk* plays, and the overall purpose of the GDPR, will provide us with important information as to the scope of the concept. Article 1(2) stipulates that:

[46] See [35].

[47] This is something acknowledged also by the WP29, which stated that the "risk based approach […] has been introduced recently as a core element of the accountability principle itself", [17], 2.

[48] See [35]; [34].

> "This Regulation protects fundamental rights and freedoms of natural persons and in particular their right to the protection of personal data".

Article 24 should also be taken into consideration. This is because of the role of *risk* and its connection to the principle of accountability, as discussed in Sect. 2. According to Article 24, for the implementation of appropriate technical and organizational measures, the data controller should take into account "the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for the rights and freedoms of natural persons". Likewise, Article 35(1) requires an assessment of "high risk to the rights and freedoms of natural persons".[49] According to the WP29, the DPIA "primarily concerns the rights to data protection and privacy but may also involve other fundamental rights […]".[50]

We observe that the EU legislature gives an additional (double) dimension to the risks that might occur when personal data are processed. These risks should not be identified and assessed solely in relation to the right to data protection but they transcend its boundaries and have to be examined also in relation to other rights and freedoms that might be interfered with because of the processing operations that take place. Additionally, the risks should not be identified and assessed solely in relation to the data subject (the actor the rights of whom are protected under the data protection legal framework) but they transcend its boundaries and have to be examined also in relation to natural persons.

'Risk to the Rights and Freedoms…'

With regard to the "risk to the rights and freedoms", the legislature provides us with examples of the most relevant rights and freedoms, mainly in the Recitals of the GDPR. Recital 4, talks about

> "all fundamental rights […] the freedoms and principles recognized in the Charter as enshrined in the Treaties, in particular the respect for private and family life, home and communications, the protection of personal data, freedom of thought, conscience and religion, freedom of expression and information, freedom to conduct a business, the right to an effective remedy and to a fair trial, and cultural, religious and linguistic diversity."

The WP29 has enriched the list by referring also to the freedom of speech, freedom of movement, prohibition of discrimination, right to liberty.[51] In Recital 75 the legislature enumerates risks in a non-exhaustive way ("*in particular*"): risk to discrimination, identity theft or fraud, financial loss, damage to the reputation, loss of confidentiality of personal data protected by professional secrecy, unauthorized reversal of pseudonimisation or any other significant economic or social disadvantage, the deprival of rights and freedoms or the prevention from exercising control over their personal data.

To give an example, let us consider the case where a data controller applies anonymization techniques on a set of personal data. Let us say that the personal data

[49] Article 33 GDPR, also talks about "risk to the rights and freedoms of natural persons" in the case of a data breach.

[50] See [18], 6. The same position was upheld by the Article 29 WP [17], 4.

[51] See [18]; [17].

have been properly anonymized. In that case, they are not qualified as "personal" anymore, and therefore do not fall under the scope of the GDPR. As pointed out by the WP29, anonymization is a type of processing activity performed on personal data. While this processing operation will not result in a risk to the data subject's right to data protection (since its purpose is the anonymization of this data), it might well result in a risk to the individual's right to privacy.[52] For example, a dataset, although anonymized, may be given/sold to a third party which will take decisions (e.g. calculation of credit risk) that will produce effects for the natural persons in that dataset. This risk is raised by the specific processing operation which while it is done as a technical measure for the mitigation of data protection risks, the purpose and use of this anonymized data set could raise a privacy risk.

An important point to be made is that *risks* are to be identified and assessed exclusively from the perspective of natural persons.[53] Whereas this is a point made in relation to the legal obligation of DPIAs, it should be understood as a general rule in all cases where data controllers are required to take into account the risks to rights and freedoms.

'…of Natural Persons'

Data controllers should not limit the risk assessment to data subjects but they have to assess whether and in what way the processing operations could negatively impact also non data subjects (i.e. natural persons whose personal data are not being processed). This is also acknowledged by the WP29 which stated that what should be taken into account is "every potential as well as actual adverse effect, assessed on a very wide scale ranging from an impact on the person concerned by the processing in question to a general societal impact".[54] This broad approach confirms also the quest for processing operations that are "designed to serve mankind".[55] Additionally, this approach is in line with the sophisticated technical context,[56] whereby the outcomes of processing operations more often than not "refer to other or more people than those involved in the input data".[57]

The recent "Facebook - Cambridge Analytica"[58] case illustrates the way in which processing operations can raise risks that transcend the boundaries of data subjects and expand to natural persons and society at large. What is worth noting in this case is that the processing operations on personal data of Facebook users that downloaded the app, as well as personal data of their "friends" (who had not downloaded the app), had a broader societal impact, that is the "undermining of democratic legitimacy" via an unlawful and opaque interference with the "opinion formation process" for the elections.[59] These processing operations did not only create risks to the fundamental rights

[52] See [16], 11.

[53] See [18], 17.

[54] See [17], 4.

[55] Recital 4 GDPR "The processing of personal data should be designed to serve mankind".

[56] See also Sect. 4 "The technology involved: IoT, AI, algorithms".

[57] See [51], 207.

[58] See [32].

[59] See [25].

and freedoms of the data subjects but had a serious impact on the core values of our society. The European legislature intends to capture this wide spectrum of impacts, through the wording of both Article 24 and Article 35 GDPR.

4 The Technology Involved: IoT, AI, Algorithms

It has been claimed and analysed why data protection "was born out of a need to protect fundamental rights from the risks created by the computer".[60] Technological developments have increased the risks to privacy and data protection, which need to be counterbalanced by the legal framework.[61] Therefore, when interpreting the concept of *risk* in data protection, one cannot disregard the technologies involved in data processing operations. As already mentioned,[62] *risk* is a scenario describing an event and its consequences. The technology involved in data processing operations relates to the "event" in this definition.

When talking about the data protection legal framework, one should keep in mind that 'technological-neutrality' should not be compromised but upheld. The quest for a technologically neutral protection appears in Recital 15 GDPR[63] and has also been encouraged by both the EDPS[64] and the WP29.[65] Case law from the CJEU has shown that the use of open and broad terms is a strategy that allows for personal data protection to be adaptable to new technologies. For example, in the *Google Spain* case,[66] the broad, functional and in light of the fundamental rights interpretation of the term 'data controller', allowed for the identification of the responsible actor (allocation of responsibility); an actor who is new[67] in terms of functionality (internet search engines) within a complicated digital environment. The European Commission recently highlighted that the "EU's sustainable approach to technologies creates a competitive edge, by embracing change on the basis of the Union's values".[68]

[60] See [36]. Also see, Ware report, pp. 37–38; [27]. " […] the birth of this legal concept is linked to the development and popularization of the computers first, and, more recently, of the Internet", 2.

[61] See [20], para 43.

[62] In the Introduction, where I refer to the definition of *risk* given by the WP29: "A risk is a scenario describing an event and its consequences, estimated in terms of severity and likelihood".

[63] Recital 15 GDPR: "In order to prevent creating a serious risk of circumvention, the protection of natural persons should be technologically neutral and should not depend on the techniques used. […]".

[64] See [28], para 38.

[65] See [20], 12.

[66] Case *Google Spain SL,* [3].

[67] Case *Google Spain SL,* Opinion of AG JÄÄSKINEN [4], para 10: "the present preliminary reference is affected by the fact that when the Commission proposal for the Directive was made in 1990 the internet in the present sense of the www did not exist and nor where there any search engines. […] nobody could foresee how it would revolutionise the world".

[68] See [22] accessed 10 November 2018.

4.1 New Technologies: Characteristics and Challenges

In the following paragraphs I will present some of the main characteristics that new technologies (IoT, AI, algorithms) share. In line with the need to sustain the technologically-neutral character of the GDPR, I will group these characteristics in a way so that it becomes apparent that the legal challenges they present are similar and such should be the response to them.[69] Having done that, I intend to examine if and how these challenges could influence the concept of *risk* and potentially its role.

The Internet of Things (IoT) is a technology sector whereby a network of physical devices, sensors, software etc. is created. This network relies on “large data collection from diverse sources and data exchange with various devices to provide seamless, linked-up and personalized services”.[70] Massive collection and linkage of user data as well as the creation of new information and inferences, are definitive characteristics of IoT, so that a more personalized experience is provided to the users.[71] Artificial intelligence (AI) “refers to systems that display intelligent behavior by analyzing their environment and taking actions –with some degree of autonomy- to achieve specific goals”.[72] AI needs vast amounts of data to be developed, to learn about and to interact with the environment. Algorithms are encoded procedures through which input data are being transformed into a usable, and therefore desirable, output.[73] “By following a logical and mathematical sequence, they can structure and find additional meaning in a big data environment”.[74] Many of these systems operate as “black boxes”;[75] they are opaque software tools working outside the scope of meaningful scrutiny and accountability.

The afore mentioned technologies present a high degree of complexity[76] due to the interdependency between the different components and layers[77]. Each system is part of a larger structure and forms part of a sequence of outputs.[78] Complexity also results from the multiple actors involved in these ecosystems.[79] For their optimal functionality, the systems require the processing of vast amounts of data. Additionally, they also

[69] See also, [53]: “Designing imprecise regulation that treats decision-making algorithms, AI and robotics separately is dangerous. It misinterprets their legal and ethical challenges as unrelated. Concerns about fairness, transparency, interpretability and accountability are equivalent, have the same genesis, and must be addressed together, regardless of the mix of hardware, software, and data involved”.

[70] See [52].

[71] See [50].

[72] See [22].

[73] See [37].

[74] See [51].

[75] See [47].

[76] Complexity is both on a technical and a contextual level. For a more extensive analysis of “technical and contextual complexity of algorithms” check Vedder and Naudts [51].

[77] See [24], 9 “[…] (i) the tangible parts/devices (Sensors, actuators, hardware), (ii) the different software components and applications, to (iii) the data itself, (iv) the data services (ie. collection, processing, curating, analysing), and (v) the connectivity features.”.

[78] See [51].

[79] See [24].

generate a huge amount of data (see footnote 79). Last but not least, these systems present autonomy in their behavior[80] which derives from the self-learning process and leads to the interpretation of the environment and to the execution of actions without human intervention.[81]

The complexity of these systems together with their autonomous behavior lead to the issue of ‘unpredictability’ of both their behavior and their outputs. It is quite possible that methods and usage patterns developed by these systems were not considered, not even imagined by the entity that collects the data nor the data subject at the time of collection.[82] At the same time, the black-box phenomenon, along with the complexity of these systems’ functionality, raise issues with regard to the ‘explainability and interpretability’ of both the systems per se and their outputs.

4.2 New Technologies and the Concept of Risk

According to Rodotà, new technologies and their characteristics create “a reality that becomes estranged from the fundamental rights’ framework” in the sense that “some of the principles underlying the system of personal data protection are being slowly eroded”.[83,84] For example, the principle of transparency highly relates to the issue of interpretability of these systems. The inherent opacity and complexity of algorithmic systems challenges the right to information. However, transparency of processing operations is a fundamental requirement of the GDPR[85] since it constitutes the basis for the data subject to exercise all their rights. As Kaminski notes, “information asymmetries render underlying rights effectively void”.[86] If transparency is difficult (or even impossible) to achieve, then a major risk is the data subject being deprived of having control over their personal data. Against this reality, we must ensure that the regulatory frameworks for developing and using of AI technologies are in line with these values and fundamental rights (see footnote 72). In this line, when talking about *risk* in data protection, it should be understood as a major criterion for enhancing the fundamental rights character of the data protection framework.

As mentioned already, an important characteristic is the complexity presented by these technologies. It is a double-pronged complexity, in terms of systems’ functionality and in terms of actors involved and the network created. The consequence is a distribution of control over multiple actors, which in turn has major implications for the allocation of responsibility among them. For example, “the developer of algorithmic tools may not know their precise future use and implementation [while] [t]he person(s) implementing the algorithmic tools for applications may, in turn, not fully understand

[80] See [24]. “AI software can reason, gather knowledge, plan intelligently, learn, communicate. Perceive and manipulate objects”.

[81] These are characteristics identified and grouped by the Commission [24].

[82] Purtova [46].

[83] Rodotà [38].

[84] For a more extensive overview of how data protection principles are influenced by the advancement of new technologies, check [52]. Also see [40]. And [43], 6.

[85] See [19], 9.

[86] See [39], 21.

how the algorithmic tools operate".[87] Each actor has knowledge limited to their role, and their ability to mitigate risks is again dependent on their role in this chain of actors. The allocation of accountability for algorithmic decision-making becomes therefore complicated.[88] As mentioned in an earlier section,[89] *risk* relates to accountability. Allocation of accountability and responsibility is a prerequisite for compliance and effective data protection. If this is not done correctly, then risk assessment and management will be incomplete and incorrect. This is the reason why, both risk assessment and risk management should be a multi-actor exercise. That points towards the role of developers, who are considered to be the ones with the expert knowledge when it comes to the functionality of new technologies. The GDPR does not impose any legal obligations on developers. However, their involvement is highly recommended. Recital 78,[90,91] the WP29[92] and the EDPS[93] encourage the more active involvement of developers in the identification, assessment and management of risks.

New technologies "may create new types of risks or accentuate existing risks" (see footnote 79). An example of the creation of new types of risks can be found in profiling and invasive inferential analytics.[94] They both involve processing operations that raise new types of risks due to the functionality of the new technologies used (additional collection and sharing of personal data). A similar new type of risks, are cybersecurity risks. An example of the accentuation of existing risks can be found in the case of discrimination. Discrimination is an already existing risk which could, however, be accentuated because of the increased possibility of bias in algorithms. Additionally, new technologies may also bring out other dimensions of the rights and freedoms as we know them. This is, for example, the case of "interdependent privacy"[95] or "group privacy rights".[96] What we realize is that *risk* is not a static concept. In this technical context it is more dynamic than ever before and is subject to transformations/additions/changes. It is a concept highly dependent on the advancement of technology. We can therefore acknowledge the importance of having in place a broad scope of the

[87] See [26], 39.

[88] idem, 39.

[89] Section 2 'The role of *risk* in the GDPR: Accountability & Risk-based approach'.

[90] Recital 78 GDPR: "[…] producers of the products, services and applications should be encouraged to take into account the right to data protection when developing and designing such products, services and applications and, with due regard to the state of the art, to make sure that controllers and processors are able to fulfil their data protection obligations.".

[91] Recitals are not legally binding as are the substantial provisions of the legal framework. However, they are supposed to "cast light on the interpretation to be given to a legal rule", [12], para 31.

[92] See [18], 8: "A DPIA can also be useful for assessing the data protection impact of a technology product, for example a piece of hardware or software, where this is likely to be used by different data controllers to carry out different processing operations. Of course, the data controller deploying the product remains obliged to carry out its own DPIA with regard to the specific implementation, but this can be informed by a DPIA prepared by the product provider, if appropriate.".

[93] See [27], para 37.

[94] See [52].

[95] See [31].

[96] See [45]; [44]; [33].

concept of *risk* as already suggested by the legislature. It will also challenge the process of identifying *risks*, in the sense of being able to foresee them.[97] This points towards the need of having a flexible and dynamic concept of *risk.*

Until now, I have discussed the challenges that new technologies bring and what they tell us about the concept of *risk.* However, what needs to be acknowledged is that new technologies do not only raise risks but can and should be used to also address them (e.g. an AI system will be trained and then used to spot cyberattacks on the basis of data from the concerned network or system) (see footnote 72). Technology is both a friend and a foe for fundamental rights and freedoms. This is also apparent from the fact that "data protection by design and by default" is introduced in Article 25.

5 Conclusion

In this article I answered the following research question: "How do the role and the scope of *risk* in the GDPR as well as the technology involved in data processing operations inform the meaning of *risk* in the field of data protection?"

To answer this research question, I firstly examined the 'role' that has been attributed to *risk* in the GDPR. I argued that *risk* is inextricably linked to the principle of accountability and thus to compliance. Its role is to contribute to the effective protection of the "essence" of the fundamental right to data protection and therefore enhance the rights-based approach of the GDPR. By furthermore explaining that the GDPR introduces a modified compliance scheme through an enhanced accountability principle, I argued that *risk* should not be understood as a '(non)compliance risk'. I then turned to the examination of the 'scope' of *risk.* By looking at the legislature's wording and by identifying the broad scope and meaning they assign to *risk* ("risk to the rights and freedoms of natural persons"*)* I argued in favor of the legislature's intention to place *risk* in a central position in the broader system of protection of fundamental rights and freedoms.

This broad scope is in line with the current data processing reality, whereby data operations are largely and increasingly performed via the use of new technologies (e.g. IoT, AI) thereby raising risks to all fundamental rights and freedoms. Due to the complexity, the autonomy in behavior and the vast amounts of data they process but also generate, new technologies require that *risk* is given a flexible and dynamic meaning, so that it captures new risks that are raised but also novel dimensions of the fundamental rights as we currently know them. Apart from the broad scope in terms of subject matter, *risk* should have an equally broad scope in terms of the actors involved in its assessment and management. Allocation of responsibility in complex ecosystems is difficult given that control is distributed among the multiple actors involved. Adopting a functional approach towards *risk* by examining the factual influence of each actor, is highly suggested and renders the assessment and management of *risk,* a multi-actor exercise, whereby technology developers and providers have an important role to

[97] Wachter [52] "the uncertain value of personal data generated and processed by IoT devices and services necessarily limits the scope of risks that can be foreseen, and thus the protection offered by DPIAs".

play. Apart from the fact that this enhances the role that the risk is called to play, it is also in line with the principle of proportionality that should apply to the legal duties of data controllers. Aiming for a broad, open and flexible concept of risk additionally enhances the technologically-neutral character of the data protection legal framework.

In this article, I extracted information from the GDPR and explained the reason why they should be used as means of interpretation of the concept of *risk* in data protection. This is one important message of this article. The second message of the article is that we shall not disregard the new technologies involved in data processing operations and the way in which their characteristics can influence our understanding of *risk*. This is an element also extracted from the GDPR. The understanding of the role and the scope attributed to *risk* in the GDPR, and of the new technologies involved in data processing operations, all contribute to the development of a theoretical framework against which the concept of *risk* can be approached in an objective and consistent way in the field of data protection. The findings of this article should be understood as a firm starting point for further steps to be taken towards the legal qualification of *risk* as well as of its constitutive elements (likelihood and severity) in data protection.

References

1. Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data OJ L 281, 23 November 1995, pp. 31–50 (1995)
2. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) OJ L 119, 4 May 2016, pp. 1–88 (2016)
3. Case C-131/12 Google Spain SL, Google Inc. v AEPD, Mario Costeja González ECLI:EU:C:2014:317 (2014)
4. Case C-131/12 Google Spain SL, Google Inc. v AEPD, Mario Costeja González ECLI:EU:C:2013:424, Opinion of AG JÄÄSKINEN (2013)
5. Case C-274/99 P Connolly v Commission ECLI:EU:C:2001:127 (2001)
6. Case C-465/00 Österreichischer Rundfunk and Others ECLI:EU:C:2003:294 (2003)
7. Case C-212/13 Ryneš ECLI:EU:C:2014:2428 (2014)
8. Joined Cases C-293/12 and C-594/12, Digital Rights Ireland and Seitlinger and Others ECLI:EU:C: 2014:238 (2014)
9. Case C-582/14 Breyer ECLI:EU:C:2016:779 (2016)
10. Case T-259/03 Nikolaou v Commission ECLI:EU:T:2007:254 (2007)
11. Case C-473/12, IPI EU:C:2013:715 (2013)
12. Case 215/88 Casa Fleischhandels-GmbH v Bundesanstalt für landwirtschaftliche Marktordnung ECLI:EU:C:1989:331 (1989)
13. Article 29 Data Protection Working Party, 'Opinion 4/2007 on the Concept of Personal Data'
14. Article 29 Data Protection Working Party, 'Opinion 1/2010 on the Concepts of "Controller" and "Processor"'
15. Article 29 Data Protection Working Party 'Opinion 3/2010 on the Principle of Accountability'

16. Article 29 Data Protection Working Party ‘Opinion 05/2014 on Anonymisation Techniques’
17. Article 29 Data Protection Working Party ‘Statement on the role of a risk-based approach in data protection legal frameworks.’ Technical report WP 218, 30 May 2014
18. Article 29 Data Protection Working Party ‘Guidelines on Data Protection Impact Assessment (DPIA) and Determining Whether Processing Is “Likely to Result in a High Risk” for the Purposes of Regulation 2016/679’ WP 248 rev 0.1, 4 April 2017. Accessed 4 October
19. Article 29 Data Protection Working Party ‘Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679 (WP251rev.01)’
20. Article 29 Data Protection Working Party WP 168 The Future of Privacy: Joint Contribution to the Consultation of the European Commission on the Legal Framework for the Fundamental Right to Protection of Personal Data. 28
21. Commission of the European Communities, ‘Amended Proposal for a COUNCIL DIRECTIVE on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data’
22. Commission, Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions, Artificial Intelligence for Europe COM(2018) 237 Final. https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe
23. Commission, Communication from the Commission to the European Parliament and the Council, Stronger protection, new opportunities - Commission guidance on the direct application of the General Data Protection Regulation as of 25 May 2018, COM(2018) 43 final. https://ec.europa.eu/commission/sites/beta-political/files/data-protection-communication-com.2018.43.3_en.pdf
24. Commission, Commission Staff Working Document: Liability for Emerging Digital Technologies *Accompanying the document* Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions, Artificial Intelligence for Europe SWD (2018) 137 Final (2018). https://ec.europa.eu/digital-single-market/en/news/european-commission-staff-working-document-liability-emerging-digital-technologies
25. Council of Europe, Internet and Electoral Campaigns – Study on the use of Internet in electoral campaigns, DGI(2017)11. https://rm.coe.int/use-of-internet-in-electoral-campaigns-/16807c0e24
26. Council of Europe, Algorithms and Human Rights Study on the Human Rights dimensions of automated data processing techniques (in particular algorithms) and possible regulatory implications, DGI(2017)12. https://edoc.coe.int/en/internet/7589-algorithms-and-human-rights-study-on-the-human-rights-dimensions-of-automated-data-processing-techniques-and-possible-regulatory-implications.html
27. EDPS (European Data Protection Supervisor), ‘Opinion 5/2018, Preliminary Opinion on Privacy by Design’, 31 May 2018
28. EDPS (European Data Protection Supervisor), Opinion of the EDPS on the Communication from the Commission to the European Parliament, the Council, the Economic and Social Committee and the Committee of Regions – “A comprehensive approach on personal data protection in the European Union”, Brussels, 14 January 2011. https://edps.europa.eu/sites/edp/files/publication/11-01-14_personal_data_protection_en.pdf
29. OECD Guidelines on the Protection of privacy and transborder flows of personal data (1980). http://www.oecd.org/sti/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm

30. Alhadeff, J., Van Alsenoy, B., Dumortier, J.: The accountability principle in data protection regulation: origin, development and future directions. In: Guagnin, D., Hempel, L., Ilten, C., Kroener, I., Neyland, D., Postigo, H. (eds.) Managing Privacy through Accountability, pp. 49–82. Palgrave Macmillan UK, London (2012). https://doi.org/10.1057/9781137032225_4
31. Biczók, G., Chia, P.H.: Interdependent privacy: let me share your data. In: Sadeghi, A.-R. (ed.) FC 2013. LNCS, vol. 7859, pp. 338–353. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39884-1_29
32. Cadwalladr, C., Graham-Harrison, E.: Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. The Guardian, 17 March 2018
33. Floridi, L.: Group privacy: a defence and an interpretation. In: Taylor, L., Floridi, L., van der Sloot, B. (eds.) Group Privacy. Philosophical Studies Series, vol. 126, pp. 83–100. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-46608-8_5
34. Gellert, R.: Why the GDPR risk-based approach is about compliance risk, and why it's not a bad thing. In: Schweighofer, E., Kummer, F., Sorge, C. (eds.) Trends und Communities der Rechtsinformatik - Trends and Communities of legal informatics: Tagungsband des 20. Internationalen Rechtsinformatik Symposions - IRIS 2017 - Proceedings of the 20th International Legal Informatics Symposium. Austrian Computer Society, pp. 527–532 (2017)
35. Gellert, R.: Understanding the notion of risk in the general data protection regulation. Comput. Law Secur. Rev. **34**(2), 279–288 (2018). https://doi.org/10.1016/j.clsr.2017.12.003
36. Gellert, R.: Understanding data protection as risk regulation. Internet J. Law **18**(11), 3–15 (2015)
37. Gillespie, T.: The relevance of algorithms. In: Gillespie, T., Boczkowski, P., Foot, K. (eds.) Media Technologies: Essays on Communication, Materiality and Society. MIT Press, Cambridge (2012). https://doi.org/10.7551/mitpress/9780262525374.003.0009
38. Rodotà, S.: Data protection as a fundamental right. In: Gutwirth, S., Poullet, Y., De Hert, P., de Terwangne, C., Nouwt, S. (eds.) Reinventing Data Protection?. Springer, Dordrecht (2009). https://doi.org/10.1007/978-1-4020-9498-9_3
39. Kaminski, M.: The Right to Explanation, Explained, 19 June 2018. https://doi.org/10.31228/osf.io/rgeus
40. Kuner, C., et al.: The Challenge of "Big Data" for Data Protection' 2 International Data Privacy Law 47 (2012)
41. Lenaerts, K., Gutiérrez-Fons, J.A.: To Say What the Law of the EU Is : Methods of Interpretation and the European Court of Justice' EUI Working Papers AEL 2013/9 (2013). http://cadmus.eui.eu//handle/1814/28339. Accessed 16 June 2018
42. Lynskey, O.: The Foundations of EU Data Protection Law. Oxford University Press (2015). ISBN 9780198718239
43. Macenaite, M.: The riskification of European data protection law through a two-fold shift. Eur. J. Risk Regul. **8**(3), 506–540 (2017). https://doi.org/10.1017/err.2017.40
44. Mantelero, A.: From group privacy to collective privacy: towards a new dimension of privacy and data protection in the big data era. In: Taylor, L., Floridi, L., van der Sloot, B. (eds.) Group Privacy: New Challenges of Data Technologies. PSS, vol. 126, pp. 139–158. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-46608-8_8
45. Mittelstadt, B.: From individual to group privacy in big data analytics. Philos. Technol. **30**, 475 (2017). https://doi.org/10.1007/s13347-017-0253-7
46. Purtova, N.: The law of everything. Broad concept of personal data and future of EU data protection law. Law, Innov. Technol. **10**(1), 40–81 (2018). https://doi.org/10.1080/17579961.2018.1452176

47. Pasquale, F.: The Black Box Society The Secret Algorithms That Control Money and Information. Harvard University Press (2015). ISBN 9780674368279
48. Quelle, C.: 'The "risk revolution" in EU data protection law: we can't have our cake and eat it, too. In: Leenes, R., van Brakel, R., Gutwirth, S., De Hert, P. (eds.) Data Protection and Privacy: The Age of Intelligent Machines, 1st edn, vol. 10. Hart Publishing (2017)
49. Spina, A.: A regulatory mariage de figaro: risk regulation, data protection, and data ethics. Eur. J. Risk Regul. **8**(1), 88–94 (2017). https://doi.org/10.1017/err.2016.15
50. Tene, O., Polonetsky, J.: Big data for all: privacy and user control in the age of analytics. Nw. J. Tech. Intell. Prop. **11**, 239 (2013)
51. Vedder, A., Naudts, L.: Accountability for the use of algorithms in a big data environment. Int. Rev. Law Comput. Technol. - Justice Algorithmic Robes **31**(2), 206–224 (2017)
52. Wachter, S.: The GDPR and the Internet of Things: a three-step transparency model. Law Innov. Technol. **10**(2), 266–294 (2018). https://doi.org/10.1080/17579961.2018.1527479
53. Wachter, S., Mittelstadt, B., Floridi, L.: Transparent, explainable, and accountable AI for robotics. Sci. Robot. **2**(6), eaan6080 (2017)

Privacy Patterns for Pseudonymity

Alexander Gabel(✉) and Ina Schiering

Ostfalia University of Applied Sciences, Wolfenbüttel, Germany
{ale.gabel,i.schiering}@ostfalia.de

Abstract. To implement the principle of Privacy by Design mentioned in the European General Data Protection Regulation one important measurement stated there is pseudonymisation. Pseudonymous data is widely used in medical applications and is investigated e.g. for vehicular ad-hoc networks and Smart Grid. The concepts used there address a broad range of important aspects and are therefore often specific and complex. Some privacy patterns are already addressing pseudonymity, but they are mostly abstract or rather very specific. This paper proposes privacy patterns for the development of pseudonymity concepts based on the analysis of pseudonymity solutions in use cases.

Keywords: Privacy by Design · Privacy patterns · Pseudonymity · Anonymity

1 Introduction

The use of pseudonymisation is proposed in the European General Data Protection Regulation (GDPR) [1] as an important measurement for implementing Privacy by Design and to enhance the security of processing. It would be preferable to render data anonymous such that the data subject is no longer identifiable, but this has been proven hard in some applications by Naranyan without considerable data utility loss [23]. Pseudonymisation of data is already widely used for the processing of patient data in medical studies or in the context of e-health applications [14]. Other application areas where pseudonymity concepts are investigated include Smart Grid applications [32], vehicular ad-hoc networks (VANETs) [20] where location privacy is in the focus, billing [8] and RFID applications [13].

Compared to this considerable amount of pseudonymity approaches for specific use cases, privacy patterns collections [6] and a review of privacy pattern research by Lenhard et al. [18] mention relatively few pseudonymity patterns in their spreadsheet. These patterns are mainly very abstract as e.g. *Pseudonymous Identity, Pseudonymous Messaging* and few are rather complex e.g. *Attribute-based Credentials* [6] or *Pseudonym Broker Pattern* proposed by Hillen [15].

The aim of this paper is to analyse pseudonymity solutions for use cases in various domains, identify important elements of these solutions and propose additional pseudonymity patterns based on these elements. These patterns are integrated with existing patterns in the context of a pattern language.

E. Kosta et al. (Eds.): Privacy and Identity 2018, IFIP AICT 547, pp. 155–172, 2019.
https://doi.org/10.1007/978-3-030-16744-8_11

2 Related Work

Pseudonymity patterns were already proposed by Hafiz [12]. In his pattern language for privacy enhancing technologies he integrated the rather general pattern *Pseudonymous Identity*. This pattern is described as "hid[ing] anonymity targets under a pseudonym" [12]. It is recommended to hide an identity using a "random pseudonym that does not relate to the original". Hafiz lists important use cases and related work regarding pseudonymisation technologies. The pattern itself is however very generic. Important issues (Insider attacks, Reversibility of the pseudonym mapping, ephemeral pseudonyms etc.) are already mentioned but not addressed in detail.

The pattern *Pseudonymous Messaging* [6] has a focus on a specific use case. The idea is to exchange the communication partners' addresses with pseudonyms known by a trusted third party, which preserves the pseudonymity of users but itself is able to re-identify the pseudonyms. This pattern is also known as *Pseudonymous E-Mail* proposed by Schumacher in 2003 [33]. Pseudonymity may also be implemented using *Attribute-based Credentials* [6], which provide a rather complex but full-fledged identity management solution. Privacy Enhancing Technologies (PETs), such as IBM Identity Mixer allow a user to generate unlinkable pseudonyms, while allowing zero-knowledge attribute verifications [17]. Pseudonyms may also be bound to a certain context (domain pseudonyms), to allow linking multiple visits of the same person. Attribute-based Credentials also provide an Inspector Authority for identity recovery. The *Pseudonym Broker Pattern* was proposed by Hillen [15] and is based on a Trusted Third Party (TTP), which generates pseudonyms from the combination of a subject ID, a partner cloud ID and a time-frame. The pseudonyms are therefore relationship-based and time-limited.

Beside single patterns and pattern languages there are several privacy pattern catalogues. The website privacypatterns.eu consists of 26 patterns, with additional eight dark patterns on the subdomain dark.privacypatterns.eu. Another catalogue is privacypatterns.org covering 50 patterns (duplicates removed) without any dark patterns. In general this catalogue is a superset of privacypatterns.eu. Drozd suggests a catalogue where 38 patterns are classified according to ISO/IEC 29100:2011 (E), to integrate privacy patterns into the software development process [7]. Furthermore Lenhard et al. collected and categorized a large list of 148 (not necessarily unique) patterns from different publications as part of their literature study [18]. Caiza et al. created a taxonomy of types of relationships of patterns [3]. These relationships are also employed here to investigate the connections between pseudonym patterns.

3 Background

As a basis for the following analysis pseudonymisation approaches from different use cases are reviewed. Also the comprehensive investigation of general aspects of pseudonymity in the terminology paper by Pfitzmann et al. [28] are considered.

Pseudonyms are mostly prevalent in the health sector and are particularly used for the pseudonymisation of patients data used in medical research. The data usage can be divided into primary or secondary usage. Often data for medical research projects is derived from data collected during the treatment of diseases etc. and as such described as secondary use. Primary usage however may be present, for example when a new medication is tested, without actual treatment in the first place.

Other uses of pseudonyms may occur, for example when only partial records are transmitted to a third party for evaluation (such as blood samples being sent to a laboratory). Furthermore e-health (Electronic Health Record (EHR), German Electronic Health Card (eGK)) approaches often employ pseudonyms to prevent linkability between multiple health organizations, and as such follow the principle of data separation. Modern approaches for pseudonym-based privacy in e-health are usually data owner centric and protect against attackers from the inside (e.g. administrators).

Riedl et al. created PIPE [25], a privacy-preserving EHR system, which employs layer-based security in combination with pseudonymised data fragments to provide unlinkability between a patient's data and their identity, as well as unlinkability between different health record fragments of the same patient. They furthermore employ a thresholded secret sharing scheme as a mechanism to recover access keys in case of destroyed or lost smart cards [31]. However their approach is patented and therefore there are usage restrictions. Heurix et al. also proposed PERiMETER, which extends their previous work to also include privacy-preserving metadata queries [14].

Caumanns describes an architecture developed for the German electronic health insurance card [5], which was developed at the Fraunhofer Institute for Software and Systems Engineering. The approach uses a ticket-based (challenge-response) method to authenticate users, while keeping links between data fragments hidden using pseudonyms. Stingl and Slamanig also proposed an approach based on unlinkable data fragments in 2007 [35]. Other pseudonymisation systems, which are not data owner centric, often employ trusted third parties (TTPs) [26,29] for organizational separated pseudonymisation (often required by law). The TTPs store pseudonym tables or cryptographic secrets necessary to perform pseudonymisation, and often furthermore allow the inverted mapping: re-identification of pseudonyms. Neubauer and Kolb compare different pseudonymisation methods for medical data with a focus on legal aspects [24].

Another area where pseudonyms are investigated are Smart Grid solutions. Data owners in this scenario are typically inhabitants. Detailed data about their energy consumption is collected by smart meters. Low-frequency data is collected for billing purposes, while high-frequency data may be used for fast demand response and to improve the grid efficiency. Furthermore there may be advanced use cases, such as incentive-based demand response schemes [10]. While low-frequency data was more or less collected previously in combination with the customers identity, high-frequency data may have an high impact on the privacy of inhabitants. Therefore to prevent misuse of the data, many approaches use pseudonyms to establish unlinkability between the customers identity and the collected power consumption data. Furthermore temporal unlinkability (established through changing pseudonyms) between sequentially recorded profiles is used to reduce the traceability and therefore the risk of re-identification. Rottondi et al. [32] deploy so-called privacy-preserving nodes (PPNs) together with a secret sharing scheme, to separate the pseudonymisation process from the assigned data and to unlink the network address of the smart meter from the pseudonymised data. Finster and Baumgart combine blind signatures, a lightweight one-way peer-to-peer anonymisation network and a bloom filter to realize pseudonymised data collection without a trusted third party, while preserving the unlinkability between network addresses and customer data [9].

Another interesting area to consider is the incorporation of electric vehicles into the smart grid via vehicle-to-grid (V2G) networks. There challenges arise, such as location privacy, when the vehicle is authenticating to the grid in many different places, as needed for online electric vehicles [16], as well as information about the battery level which may be used for further tracking [19].

In the field of vehicular ad-hoc networks (VANETs) privacy-preserving solutions are investigated mainly with a focus on location privacy [20,37], often by using changing pseudonyms. Mano et al. express the need for pseudonymisation of datasets of location trajectories for analysis of mobility patterns. They claim that anonymised datasets (e.g. using k-anonymity) typically do not provide enough information about those patterns, when compared against pseudonymised per-user trajectories [21]. To protect them concerning re-identification, they propose to exchange the pseudonym at hub locations and introduce metrics and a verification algorithm to check whether the pseudonym exchange can be effective for all users based on plausible paths.

In the area of billing, pseudonyms are used to separate the process of payment (which typically but not always [22] requires the identity of the user) from the actual usage of a particular service [8,38]. Furthermore pseudonyms may be applied to create transactions, which are not linkable in different contexts [34]. Gudymenko proposes a privacy-preserving e-ticketing system for fine-granular billing, by separating pseudonymised tracing of travel records and end user billing using a trusted third party [11]. Falletta et al. propose a distributed billing system, which requires the interaction of multiple entities to disclose the user's identity, therefore avoiding a single trusted third party [8].

In RFID systems, regularly changing pseudonyms (often based on cryptographic algorithms) are used to prevent tracking of RFID tags for unauthenticated readers. Henrici et al. apply the concept of onion routing in an RFID tag pseudonymisation infrastructure to prevent unwanted tracking of RFID tags [13].

Biskup and Flegel use transaction-based pseudonyms and apply a thresholded secret sharing scheme in an intrusion detection system to allow re-identification of a particular user only when a certain threshold of policy violations has been exceeded [2].

4 Analysis of Pseudonymity Approaches

To dissect the different pseudonym systems in the use cases summarized in Sect. 3 as a starting point for the analysis, the following central areas are investigated to identify the basic building blocks of pseudonym systems. First pseudonym generation is investigated and second additional functionality is considered which is necessary for the pseudonym system to fulfil its purpose.

When a pseudonym is used to protect privacy, its purpose is usually to foster the unlinkability between an individual and its pseudonyms. Therefore an important question is the scope a pseudonym. As described by Pfitzmann and Hansen [28], there are different types of pseudonyms, depending on the scope/context[1] of their usage (e.g. role pseudonym, relationship pseudonym, transaction pseudonym etc.). We extend this concept to a general scope, which may be defined by a combination of many factors that limit the usage/validity of a pseudonym. For example a pseudonym may be time-limited (i.e. only valid for one week), as well as relationship-based (i.e. differs for each party interacting with the pseudonym). The idea of the *Minimal Pseudonym Scope* pattern we propose, is to limit this scope to the smallest possible one for the purpose of data processing.

Another component of pseudonym generation is how the actual pseudonym is created. Hafiz suggests in the *Pseudonymous Identity* pattern, that "a random pseudonym [should be adopted], that does not relate to the original" [12]. However, it is not always the case that a pseudonym is really random, since typically pseudonyms are generated. For example cryptographic techniques may be used to generate a pseudonym, e.g. by encrypting or hashing certain information. This is especially useful when pseudonyms should be re-identifiable by a trusted third party. Furthermore, techniques such as *Attribute-based Credentials* also allow the creation of pseudonyms, which may need to fulfil certain cryptographic properties, e.g. in the case of a domain pseudonym. The actual method of generation often depends on other properties of the pseudonym system, therefore no additional pattern is proposed in this area.

In many cases pseudonyms are generated by a trusted third party, which in most cases also allows this trusted third party to re-identify pseudonyms, i.e. to link them back to the original (hidden) identity. However, particularly research regarding pseudonymous e-health systems noticed an inherent risk of re-identification by insiders (e.g. administrators) or due to database leaks. Therefore the abstract pattern *Data-owner based Pseudonymisation* is proposed, which switches roles and allows the data owner to create pseudonyms. The idea is to decrease the risk of re-identification and unwanted linkability in comparison to a trusted third party for pseudonym creation. This however does not mean that re-identification (e.g. in the case of misuse) is always completely impossible. For example in the case of *Attribute-based Credentials* with the presence of an inspector authority, it is still possible to recover the identity behind a pseudonym,

[1] In this paper the notion "scope" is used in order to prevent confusion with the context of patterns.

while the separation of entities (issuer, verifier, inspector authority and user) separates powers. Furthermore in some cases it may be sufficient to let the user prove that she/he is or is not the holder of a pseudonym, e.g. in legal disputes, without a trusted party being able to recover an identity behind a pseudonym.

For the second category regarding additional functionality in pseudonym systems, two main strategies were identified: Protection against re-identification and its counterpart Selective Linkability if needed for a specific service. To protect against re-identification, especially the use of *Anonymisation Networks* is common. The existing pattern *Onion Routing* is not always used, instead proxies or lightweight anonymisation networks are employed, especially in Internet of Things use cases such as Smart Grid or RFID systems. This may indicate, that a more abstract pattern *Anonymisation networks* is necessary to capture the diverse requirements and approaches of such systems.

Furthermore with additional data which may being linked to a pseudonym, the risk of re-identification due to inference attacks increases. To cope with this risk, strategies such as data de-identification/de-sensitization can be used. However, this may also decrease the utility of the data and therefore the quality of the service. Another approach, especially present in e-health use cases is to separate the data into small fragments, which are unlinkable by default but may be linked by the data owner. This prevents trivial linkability for insiders as well as in the case of a database breach, while keeping utility to authorized parties. Hence the pattern *Data fragments* is proposed especially for the use in the e-health context, were sensitive (i.e. medical) data is processed.

To allow users to share the information which pseudonyms for data fragments are connected to the same individual in a selective way, the *Encrypted Link* pattern as a kind of data owner based authorization system is proposed, in contrast to standard access control systems. While the *Data fragments* pattern can be seen as primarily establishing unlinkability and preventing re-identification, the *Encrypted Link* pattern selectively establishes linkability in a secure way without leaking unnecessary information to unauthorized entities.

When the pattern *Minimal Pseudonym Scope* is applied, but selective linkability is necessary to exchange data across different scopes, the *Pseudonym Converter* pattern can be applied. When party A wants to send data regarding a pseudonymous subject S to party B, but A and B have their own distinct pseudonyms referring to S, a pseudonym converter may translate between the parties without directly establishing the link between two pseudonyms. On the other side, given a pseudonym system based on a trusted third party responsible for pseudonym generation, a good practice for data minimization is to apply the *Data hidden from Pseudonymiser* pattern, such that the pseudonymiser is only responsible for translating between identities and pseudonyms without access to related data, not necessary for that sole purpose. We found different methods for hiding the data, such as de-identification, encryption and secret splitting.

Finally it may be necessary to recover the identity behind a pseudonym for reasons such as handling of misuse. This functionality was required in many systems and handled in different ways, therefore the pattern *Recoverable Identity* is proposed to capture this important concept and ways to implement it.

5 Pseudonymity Patterns

In this section we present our patterns for pseudonymity, as well as relations between those pattern and existing patterns (see Fig. 1). We created eight patterns, however due to page limitations, we present only a subset of them[2]. An overview of the remaining can be seen in Table 1.

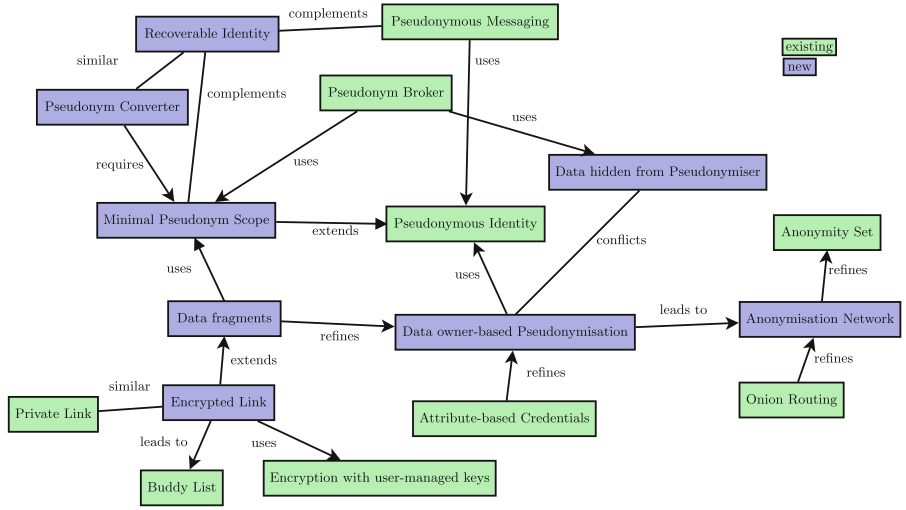


Fig. 1. Pattern language for pseudonymity patterns

5.1 Minimal Pseudonym Scope

Summary: Restrict the linkability of a pseudonym by limiting the usage to the smallest possible scope for the purpose of data processing (data minimization).

Context: It is often not necessary for a pseudonym to have a very broad scope in the general case. Even if linkability across different scopes is necessary, usually not every party (e.g. an attacker) should be able to link pseudonyms trivially.

[2] The full pattern catalogue can be retrieved from https://github.com/a-gabel/pseudonym-privacy-patterns.

Problem: Pseudonyms are usually used to protect an identity from being disclosed. However when using only a single unique pseudonym for an identity, it becomes increasingly traceable and it may be linked across several databases and scopes. With more information about an identity, re-identification of a pseudonym becomes increasingly likely. Also in case of a data breach, datasets with potentially different information about an identity, which refer to the same pseudonym become linkable for attackers.

Forces/Concerns: Controllers may want linkability across different scopes for some services. Users may prefer not to be tracked across multiple scopes.

Solution: To prevent linkability across different scopes using a pseudonym, one may limit the use of a pseudonym to a small scope. For different scopes, different pseudonyms are used, which cannot be linked without additional information. A scope may be depend on a role (e.g. shopping or video on demand), relationship (Company A or B), location, time frame or transaction (one-time use). Furthermore combinations may be useful (e.g. role-relationship), depending on the use case. The controller needs to balance the purpose of the service and privacy of users. The scope has to be chosen according to the principle of data minimization. Selective Linkability can also be established via *Recoverable Identity* or *Pseudonym Converter*, which might decrease the risk in case of a data breach.

Benefits: In case of a data breach, pseudonyms across different scopes may not be linked trivially. Pseudonyms only refer to (small) partial identities, which cannot be linked trivially.

Liabilities: Additional complexity may be necessary, if linkability of pseudonyms in different scopes is necessary under certain conditions (e.g. by applying a *Pseudonym Converter*).

Examples: A user may use different **relationship-pseudonyms** [28], to limit linkability across different organizations. For example a user may want to use a different pseudonym for a dating website and for their business profile. Furthermore the pseudonym of a car in a car-to-x network may change depending on **location** and **time-frame**.

[Known Uses]: Pommerening and Reng use a different pseudonym for each secondary use project of electronic health record (EHR) data [29]. Mano et. al exchange pseudonyms of users when they meet at the same hub and propose a privacy verification algorithm [21]. Rottondi et al. use a time-limited pseudonym to prevent linkability of smart meters over a longer time window in a smart grid system [32]. Industrial uses include the GSM standard with the Temporary Mobile Subscriber Identity (TMSI; time- and location-limited scope), or tokenization which is recommended by the Payment Card Industry Data Security Standard (PCI DSS), resulting in different pseudonyms per party (relationship-based) and time-frame (validity of the tokenization key) [27].

[Related Patterns]: Extends *Pseudonymous Identity*, as it improves the existing solution of protecting identities behind a pseudonym by giving it a small scope, thus making it more difficult to re-identify. Complements *Recoverable Identity*, as the small scope leads to less data being linked to the real identity in case of re-identification. It is complemented by *Recoverable Identity*, as it may help to prevent misuse when many pseudonyms make it hard to track/block a user. Used by *Pseudonym Broker*, as pseudonyms are different for each organization and time frame, as well as by *Data Fragments* and *Data owner-based Pseudonymisation*. Required by *Pseudonym Converter*.

5.2 Recoverable Identity

Summary: The identity behind a pseudonym is recoverable under certain conditions.

Context: Pseudonym system are usually designed such that the re-identification of a pseudonym (i.e. determining the identity behind a pseudonym) is reasonably hard. In many cases it is sufficient to be able to link different transactions via pseudonyms. However in some cases, it might be necessary to recover the identity behind a pseudonym, for example in case of misuse of the system. Then e.g. only a trusted party/combination of multiple trusted parties should be able to recover the identity behind a pseudonym.

Problem: If identity recovery is necessary, it should usually only be possible in very specific and constrained cases.

Forces/Concerns: Users may fear, that their identity is recovered in cases where it is not necessary (e.g. the user did not misuse the system), resulting in compromise of their privacy. Therefore the trusted party which is able to recover pseudonyms should transparently show and enforce their policies. The party should be trusted by both the controller and the users. The controller may want to identify users, e.g. for legal or payment purposes.

Solution: Restrict the ability of identity recovery via organizational and technical constraints.

[Implementation]: One option can be to use a Trusted Third Party for Identity Recovery. The pseudonym mapping may be stored in a table or encrypted inside the pseudonym. Another option is to use secret sharing to allow identity recovery only with $n > t$ operators, or with enough evidence (in case of misuse). Furthermore anonymous credentials/*Attribute-based Credentials* with a trusted inspector authority may be used.

Benefits: The identity behind a pseudonym is only recoverable in very specific, constrained cases. Misuse of the system by pseudonymous users may be limited, as users are informed about the possibility of identity recovery in such cases.

Liabilities: Users may have less trust in the system, if the policy for identity recovery or the technological barriers are too lax.

Examples: In a Pseudonymous Messaging system where users are communicating via email, pseudonymous users may be re-identified by a trusted third party, if they abuse the system, e.g. for illegal purposes. The pseudonymiser (the entity which translates real email addresses to pseudonymous ones) encrypts the original identity inside the pseudonymous e-mail address and is therefore the only entity which is able to recover an identity from a pseudonym only. Another example: In a smart grid system, each smart meter uses a pseudonym, which is generated by encrypting identifiable information (e.g. an ID known to the grid operator) using the public key of a trusted third party (TTP). The TTP may recover identities behind pseudonyms in case of misuse using its private key.

[**Known Uses**]**:** Hussain et al. use a secret sharing scheme to allow only the combination of all revocation authorities to recover the identity behind the pseudonym of an online electric vehicle (OLEV) in case of a legal need, such as refusing to pay after electricity consumption [16]. Rottondi et al. allow the Configurator, a trusted party of a Smart Grid system, the recovery of identities by decrypting the identity as part of the pseudonym using its private key [32]. Biskup and Flegel use a secret sharing scheme to allow re-identification of pseudonyms in an intrusion detection system only when there is enough evidence (i.e. enough events from a certain identity within a time-frame) [2]. Attribute-based Credential Systems allow re-identification of users via a separate Inspector authority.

[**Related Patterns**]**:** Complements *Pseudonymous Messaging*, as it may help to prevent misuse of the messaging service. Complements *Minimal Pseudonym Scope*, as it helps to re-identify users in case of misuse. Similar to *Pseudonym Converter*, as both patterns allow a trusted third party (TTP) to selectively link a pseudonym. In case of the *Pseudonym Converter*, a TTP can link pseudonyms, while in *Recoverable Identity* the TTP can link a pseudonym to an identity.

5.3 Data Hidden from Pseudonymiser

Summary: Data being pseudonymised is not readable by the Pseudonymiser (entity which assigns pseudonyms).

Context: The pseudonymiser (i.e. the entity which creates pseudonyms and assigns them to identities) is usually only responsible for assigning pseudonyms, but does not need to have access to additional data. For example a pseudonymisation entity for medical data may not need access to the assigned medical reports etc. Additionally, pseudonyms may be generated based on unique IDs instead of identifiable information (e.g. name).

Problem: When assigning a pseudonym to an identity the pseudonymiser might learn additional information, which may be unwanted and unnecessary.

Forces/Concerns: The pseudonymiser needs some kind of reference to the original identity. However, information about the person (such as the name or further information) may not be necessary. A secure channel between a data source and the party which receives pseudonymised data might be needed.

Solution: Hide data assigned to an identity by e.g. applying cryptographic measures before pseudonymisation.

[Implementation]: Encryption of the data: Before sending an identity and data to a pseudonymiser, encrypt the assigned data using public key cryptography. The pseudonymiser will receive a tuple $(ID, Enc(data))$ from the data source as pseudonymisation request and will send a tuple $(Pseudonym, Enc(data))$ to a party from which the real identity should be hidden. The receiving party is able to decrypt the hidden data using its private key. Secret Sharing: Use a secret sharing scheme to split the assigned data into parts, which are then pseudonymised by multiple distinct pseudonymisers. The receiving party is able to reconstruct the data if all parts are received, but each pseudonymiser on its own is unable to do so. De-identification: If the pseudonymiser for a specific reason needs to have access to the assigned data, the additional use of de-identification methods to remove identifiable data (e.g. name, ID card number, birth data, ...) is strongly recommended.

Benefits: The pseudonymiser does not learn additional information about an identity. Identities may be referred to as unique random identifiers, such that other identifiable data (such as a person's name) is also not available to the pseudonymiser.

Liabilities: Additional complexity of the system may arise depending on how the hiding mechanism is implemented.

Examples: A medical clinic may need to pseudonymise patients' medical data to be used in a research project. Instead of sending complete patient records with identifiable data (name, birth date etc.) to a pseudonymiser, only a list of randomly generated unique IDs is sent to the pseudonymiser. The pseudonymiser then converts each ID to a unique pseudonym and sends the resulting list (with the same order as the original list) to the research organization. Furthermore the clinic sends de-identified medical records (same order) to the research organization. The research organisation may then refer to a patient using the pseudonym from the list, while the pseudonymiser does not have any access to the medical data. Instead of sending the medical data separately, a clinic may also encrypt it for the research party and send it encrypted to the pseudonymiser, who is unable to read the encrypted data.

[Known Uses]: Pommerening and Reng hide associated medical data for the pseudonymiser by encrypting it for the receiving research organization [29]. Noumeir et al. perform de-identification of radiology data before sending it to a pseudonymisation system to reduce the risk of identification [26]. Rottondi et al. use a secret splitting scheme in a smart meter system to let several pseudonymisation nodes pseudonymise shares of a smart meter (producer) reading, ensuring that these nodes cannot read the data, while the receiving node (consumer) can do so, when receiving all secret shares [32]. Rahim et al. perform pre-pseudonymisation of patient identifiers in addition to encryption of the assigned medical data to completely hide identifiable information from the pseudonymisation server [30].

[Related Patterns]: Used by *Pseudonym Broker*, as the data assigned to a pseudonym is sent to a database or to a portal without any interaction with the Trusted Third Party, which acts as the pseudonymiser. Conflicts with *Data owner-based Pseudonymisation*, because the data owner (i.e. the pseudonymiser) already has knowledge of the data and it is not useful to hide that data. Complements *Pseudonymous Messaging*, as it hides the message content from the party which performs the pseudonymisation of the messages, providing additional privacy.

5.4 Data Fragments

Summary: Split data of a single identity into small fragments and assign each fragment its own pseudonym. Only authorized entities are given the knowledge of which pseudonyms belong together.

Context: Whenever a collection of pseudonymised data records are under risk of re-identification by inference attacks due to the informative value of combined fields.

Problem: A record of data about an identity may contain enough information to re-identify it, even if primary identifiers are removed from the record. For example the combination of the attributes gender, ZIP code and birth date may uniquely identify 87% of the US-American population [36]. Furthermore it may be unwanted in a system to enable anyone with access to the dataset (e.g. also insiders like administrators) to be able to link sensitive data.

Forces/Concerns: Using server-side encryption may not help, if insiders such as administrators have access to the encryption keys. Encrypting the data using end-to-end encryption (i.e. unauthorized entities do not have access to the keys) might help, however when the dataset is large the performance penalty may be unacceptable/impractical. De-identification of the data using techniques from the area of Statistical Disclosure Control may work for some scenarios. However, such techniques may remove data needed for the use case.

Solution: Instead of storing related data with a single pseudonym, split the data into small fragments, which are hard to re-identify by themselves, and assign each fragment its own unique pseudonym. Only authorized persons or systems get the knowledge of which pseudonyms (i.e. which data fragments) belong to the same identity. It is also possible to reveal only partial information about which fragments belong together, to limit access to certain parts of data records. The pattern may furthermore be combined with de-identification to desensitize potentially identifiable data such as birth dates (e.g. mask day and month of birth) before transmitting the data.

Benefits: Enables unlinkability of data fragments by default, while authorized entities are able to link subsets of fragments. May significantly reduce the risk of insider attacks, as insiders are unable to link fragments or establish a relation to an identity. In case of a data breach, data fragments remain unlinkable for attackers without additional knowledge. Computationally efficient, as data fragments do not necessarily have to be encrypted.

Liabilities: Increases complexity of the system, as knowledge about pseudonyms needs to be managed.

Examples: In an e-health system where health records or metadata of records from patients are stored centrally, instead of storing data referring to the same person in a linkable way, data fragments may be used to split health records into small fragments. For example each medical result is stored as a separate fragment. Only the data owner (i.e. the patient) has the knowledge which pseudonyms/data fragments belong to her. When the data owner wants to share fragments with a doctor, new pseudonyms pointing to the fragments can be generated and shared with the doctor.

[Known Uses]: PIPE (Pseudonymisation of Information for Privacy in E-Health) uses data fragments for electronic health records. By default only the patient (data owner) is able to access her health records. Access to the pseudonyms is managed through a central metadata storage which is encrypted with the user's keys. The data owner may decide to give access to some records to selected medical personnel by creating additional pseudonyms referring to fragments [25]. Identifiable and non-identifiable data is also unlinkable by default, so the system may be employed for secondary use, e.g. in research. Stingl and Slamanig describe a concept for an e-health portal, which uses unlinkable and undetectable partial identities of a patient to keep separate health records for participating parties (i.e. dentist and general practitioner access different partial identities) [35]. The Fraunhofer ISST designed a concept for the German electronic health card (eGK), which uses ticket-based authorization and challenge-based authentication to allow fine-granular access control to data fragments, which are unlinkable by default [5]. Biskup and Flegel use a secret sharing scheme to assign each event in an intrusion detection system a unique pseudonym, which keeps events unlinkable until enough evidence for re-identification is available [2]. Camenisch and Lehmann propose the use of "data snippets", which are stored with unlinkable pseudonyms. A central entity is able to link those snippets and may provide de-identified subsets of the original record to authorized parties. They suggest the use a central *Pseudonym Converter*, which is able to convert pseudonyms in a blind way while providing auditability for users [4].

[Related Patterns]: Uses *Minimal Pseudonym Scope*, as every data fragment gets its own pseudonym, therefore the scope of a pseudonym is very limited. Refines *Data owner-based pseudonymisation*, as it allows the data owner in a more specific context (i.e. shared repository of data) to perform the pseudonymisation and therefore provide a more privacy-preserving solution in comparison to a trusted third party solution. Extended by *Encrypted Link*.

6 Discussion and Final Remarks

In this paper privacy patterns and a pattern language for pseudonymity is proposed, which try to close the gap between very abstract and very complex pseudonymity patterns and to ease the development of pseudonym systems. For data minimization and unlinkability *Minimal Pseudonym Scope*, *Data fragments*, and *Data hidden from pseudonymiser* may be applied. To establish selective linkability, while staying restricted to a small audience, *Recoverable Identity*, *Pseudonym Converter* and *Encrypted Link* can assist. Furthermore to shift the asymmetry of power to the data owner side, the patterns *Data owner-based Pseudonymisation* and *Anonymisation Networks* are useful concepts.

To foster the adoption of privacy patterns, the applicability of such patterns in system development processes needs to be evaluated to derive guidelines for developers in the context of privacy engineering processes.

Table 1. Further patterns for pseudonymity (summary)

Pattern	Description
Pseudonym converter	A separate entity, the Converter, is able to translate a pseudonym from one scope to a pseudonym in another scope
Encrypted link	To authorize access to *data fragments* in a way that is not detectable by third parties, encrypt pseudonyms, pointing to *data fragments*
Anonymisation network	Hide the network identity of a communication partner by adding anonymisation nodes between communication partners
Data owner-based pseudonymisation	Generate and assign pseudonyms on the data owner side instead of using a third party, to keep the link between pseudonym and the data owner hidden from other parties

The taxonomy of relations between patterns by Caiza et al. [3] is a promising approach to provide an overview of privacy patterns, because of the visual structure and the relations between patterns. Some of the pattern relations proposed there were not fully applicable in the context of privacy patterns. E.g. *leads to* specifies that a pattern is necessary, as to not leave unsolved problems. However, in our experience the existence of problems may depend on the use case (e.g. *Recoverable Identity*). Also the relation *complements* is defined as symmetric, which was not always the case here.

The importance of a relationship itself is a topic which may be discussed further. Some relationships may be redundant or not helpful (i.e. referencing *Pseudonymous Identity* from every pattern regarding pseudonymity), while others may give helpful insights.

Regarding the patterns for pseudonymity it has to be shown, whether this catalogue is complete or if there may be more patterns, yet to be discovered. An interesting question is, whether it is actually possible to check that a pattern language is exhaustive or to at least get hints where something may be missing. Another observation is the difference in the level of abstraction/complexity between the patterns. Developing a hierarchy of patterns, or clusters of complexity/abstractions could be a useful concept.

Acknowledgement. This work was supported by the Ministry for Science and Culture of Lower Saxony as part of SecuRIn (VWZN3224).

References

1. Regulation (EU) 2016/679 of the european parliament and of the council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:2016:119:TOC
2. Biskup, J., Flegel, U.: On pseudonymization of audit data for intrusion detection. In: Federrath, H. (ed.) Designing Privacy Enhancing Technologies. LNCS, vol. 2009, pp. 161–180. Springer, Heidelberg (2001). https://doi.org/10.1007/3-540-44702-4_10
3. Caiza, J.C., Martín, Y.S., Del Alamo, J.M., Guamán, D.S.: Organizing design patterns for privacy: a taxonomy of types of relationships. In: Proceedings of the 22nd European Conference on Pattern Languages of Programs, EuroPLoP 2017, pp. 32:1–32:11. ACM, New York (2017)
4. Camenisch, J., Lehmann, A.: Privacy-preserving user-auditable pseudonym systems. In: 2017 IEEE European Symposium on Security and Privacy (EuroS&P), pp. 269–284, April 2017
5. Caumanns, J.: Der Patient bleibt Herr seiner Daten Realisierung des eGK-Berechtigungskonzepts über ein ticketbasiertes, virtuelles Dateisystem. Informatik-Spektrum **29**(5), 323–331 (2006)
6. Colesky, M., et al.: Privacy patterns. https://privacypatterns.org/. Accessed 1 Aug 2018
7. Drozd, O.: Privacy pattern catalogue: a tool for integrating privacy principles of ISO/IEC 29100 into the software development process. In: Aspinall, D., Camenisch, J., Hansen, M., Fischer-Hübner, S., Raab, C. (eds.) Privacy and Identity 2015. IAICT, vol. 476, pp. 129–140. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-41763-9_9
8. Falletta, V., Teofili, S., Proto, S., Bianchi, G.: P-DIBS: Pseudonymised DIstributed billing system for improved privacy protection. In: 2007 16th IST Mobile and Wireless Communications Summit, pp. 1–5, July 2007
9. Finster, S., Baumgart, I.: Pseudonymous smart metering without a trusted third party. In: 2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, pp. 1723–1728, July 2013
10. Gong, Y., Cai, Y., Guo, Y., Fang, Y.: A privacy-preserving scheme for incentive-based demand response in the smart grid. IEEE Trans. Smart Grid **7**(3), 1304–1313 (2016)
11. Gudymenko, I.: A privacy-preserving e-ticketing system for public transportation supporting fine-granular billing and local validation. In: Proceedings of the 7th International Conference on Security of Information and Networks, SIN 2014, pp. 101:101–101:108. ACM, New York (2014)
12. Hafiz, M.: A pattern language for developing privacy enhancing technologies. Softw.: Pract. Exp. **43**(7), 769–787 (2013)
13. Henrici, D., Gotze, J., Muller, P.: A hash-based pseudonymization infrastructure for RFID systems. In: Second International Workshop on Security, Privacy and Trust in Pervasive and Ubiquitous Computing (SecPerU 2006), pp. 6-27, June 2006
14. Heurix, J., Karlinger, M., Neubauer, T.: Pseudonymization with metadata encryption for privacy-preserving searchable documents. In: 2012 45th Hawaii International Conference on System Sciences, pp. 3011–3020, January 2012

15. Hillen, C.: The pseudonym broker privacy pattern in medical data collection. In: 2015 IEEE Trustcom/BigDataSE/ISPA, vol. 1, pp. 999–1005, August 2015
16. Hussain, R., Son, J., Kim, D., Nogueira, M., Oh, H., Tokuta, A.O., Seo, J.: PBF: a new privacy-aware billing framework for online electric vehicles with bidirectional auditability. Wirel. Commun. Mob. Comput. **2017** (2017)
17. IBM Research - Zürich: Specification of the identity mixer cryptographic library version 2.4.43. https://abc4trust.eu/index.php?option=com_content&view=article&id=187. Accessed 1st Aug 2018
18. Lenhard, J., Fritsch, L., Herold, S.: A literature study on privacy patterns research. In: 2017 43rd Euromicro Conference on Software Engineering and Advanced Applications (SEAA), pp. 194–201. IEEE (2017)
19. Liu, H., Ning, H., Zhang, Y., Guizani, M.: Battery status-aware authentication scheme for V2G networks in smart grid. IEEE Trans. Smart Grid **4**(1), 99–110 (2013)
20. Lu, R., Lin, X., Luan, T.H., Liang, X., Shen, X.: Pseudonym changing at social spots: an effective strategy for location privacy in VANETs. IEEE Trans. Veh. Technol. **61**(1), 86–96 (2012)
21. Mano, K., Minami, K., Maruyama, H.: Privacy-preserving publishing of pseudonym-based trajectory location data set. In: 2013 International Conference on Availability, Reliability and Security, pp. 615–624, September 2013
22. Martinez-Pelaez, R., Rico-Novella, F., Satizabal, C.: Mobile payment protocol for micropayments: withdrawal and payment anonymous. In: 2008 New Technologies, Mobility and Security, pp. 1–5, November 2008
23. Narayanan, A., Shmatikov, V.: Robust De-anonymization of Large Sparse Datasets. In: Proceedings of the 2008 IEEE Symposium on Security and Privacy, SP 2008, pp. 111–125. IEEE Computer Society, Washington, DC, USA (2008)
24. Neubauer, T., Kolb, M.: Technologies for the pseudonymization of medical data: a legal evaluation. In: 2009 Fourth International Conference on Systems, pp. 7–12, March 2009
25. Neubauer, T., Heurix, J.: A methodology for the pseudonymization of medical data. Int. J. Med. Inform. **80**(3), 190–204 (2011)
26. Noumeir, R., Lemay, A., Lina, J.M.: Pseudonymization of radiology data for research purposes. J. Digit. Imaging **20**(3), 284–295 (2007)
27. PCI Security Standards Council: Tokenization product security guidelines. Technical report 1.0, PCI Security Standards Council, April 2015. https://www.pcisecuritystandards.org/documents/Tokenization_Product_Security_Guidelines.pdf
28. Pfitzmann, A., Hansen, M.: A terminology for talking about privacy by data minimization: anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management (2010)
29. Pommerening, K., Reng, M.: Secondary use of the EHR via pseudonymisation. Stud. Health Technol. Inform. **103**, 441–446 (2004)
30. Rahim, Y.A., Sahib, S., Ghani, M.K.A.: Pseudonmization techniques for clinical data: Privacy study in Sultan Ismail Hospital Johor Bahru. In: 7th International Conference on Networked Computing, pp. 74–77, September 2011
31. Riedl, B., Grascher, V., Neubauer, T.: Applying a threshold scheme to the pseudonymization of health data. In: 13th Pacific Rim International Symposium on Dependable Computing (PRDC 2007), pp. 397–400, December 2007
32. Rottondi, C., Mauri, G., Verticale, G.: A data pseudonymization protocol for smart grids. In: 2012 IEEE Online Conference on Green Communications (GreenCom), pp. 68–73, September 2012

33. Schumacher, M.: Security patterns and security standards - with selected security patterns for anonymity and privacy. In: Privacy, European Conference on Pattern Languages of Programs (EuroPLoP 2003) (2003)
34. Seigneur, J.M., Jensen, C.D.: Trust enhanced ubiquitous payment without too much privacy loss. In: Proceedings of the 2004 ACM Symposium on Applied Computing, SAC 2004, pp. 1593–1599. ACM, New York (2004)
35. Stingl, C., Slamanig, D.: Berechtigungskonzept für ein ehealth-portal. na (2007)
36. Sweeney, L.: Simple demographics often identify people uniquely. Health (San Franc.) **671**, 1–34 (2000)
37. Thenmozhi, T., Somasundaram, R.M.: Pseudonyms based blind signature approach for an improved secured communication at social spots in VANETs. Wirel. Pers. Commun. **82**(1), 643–658 (2015)
38. Zhao, X., Li, H.: Privacy preserving authenticating and billing scheme for video streaming service. In: Wen, S., Wu, W., Castiglione, A. (eds.) CSS 2017. LNCS, vol. 10581, pp. 396–410. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-69471-9_29

Implementing GDPR in the Charity Sector: A Case Study

Jane Henriksen-Bulmer(✉), Shamal Faily, and Sheridan Jeary

Bournemouth University, Poole, UK
{jhenriksenbulmer,sfaily,sjeary}@bournemouth.ac.uk

Abstract. Due to their organisational characteristics, many charities are poorly prepared for the General Data Protection Regulation (GDPR). We present an exemplar process for implementing GDPR and the DPIA Data Wheel, a DPIA framework devised as part of the case study, that accounts for these characteristics. We validate this process and framework by conducting a GDPR implementation with a charity that works with vulnerable adults. This charity processes both special category (*sensitive*) and personally identifiable data. This GDPR implementation was conducted and devised for the charity sector, but can be equally applied in any organisation that need to implement GDPR or conduct DPIAs.

Keywords: Privacy · Case study · General Data Protection Regulation · GDPR · Contextual Integrity · Privacy risk · Data Protection Impact Assessment · DPIA

1 Introduction

The General Data Protection Regulation (GDPR) is the European Union's (EU) new Data Protection Regulation that came into effect on 25th May 2018 [9]. While GDPR affects all organisations, it has particular implications for small to medium enterprises (SMEs) and charities, who, like many other organisations, collect and process personal and/or "special category" (*sensitive*) data, as these organisations often work within financial and resource restraints and therefore, may lack the expertise to fully understand how best to interpret and implement the changes brought in by GDPR. In the UK, the Data Protection Act 1998 (DPA) has been incorporated into UK law through the Data Protection Act 2018 [31], which is in line with GDPR.

GDPR imposes several new obligations on organisations; these include extending the scope and breadth of what data is classed as personal, more rights for individuals in relation to their data; a requirement for organisations to understand and document their data holdings; justify why they collect each piece of data and record the lawful basis for processing data. GDPR also introduces data protection by design and default (DPbDD) and a requirement for organisations to demonstrate compliance to the relevant authorities if challenged.

E. Kosta et al. (Eds.): Privacy and Identity 2018, IFIP AICT 547, pp. 173–188, 2019.
https://doi.org/10.1007/978-3-030-16744-8_12

Privacy protection in practice must be meaningful to be effective [1]. Privacy has to be implemented to not only account for legal requirements but also the context within which privacy protection is required, including looking at the specific sector or industry an organisation works within. Thus, while GDPR may not necessarily require expert knowledge to implement, the requirements and obligations still require interpretation. Charities, like many organisations, find it difficult to fully understand when and how best to implement GDPR.

We present a case study that illustrates how charities and small & medium-sized enterprises (SMEs) can implement GDPR in an organised, step by step approach. As part of this we also present the DPIA Data Wheel: a Data Protection Impact Assessment (DPIA) framework for assessing what the privacy implications of processing data within the organisation are. There are no current solutions for implementing GDPR or carrying out DPIAs in this context. This work will, therefore, benefit any charity or SME dealing with vulnerable clients. Our approach builds on previous work using Nissenbaum's Contextual Integrity (CI) framework [21] to create a decision framework for assessing privacy risks in Open Data [12], and expands on this to support the GDPR implementation and the DPIA framework.

The rest of the paper is structured as follows. We begin by providing an overview of the changes brought in by GDPR in Sect. 2. This is followed by a brief review of risk assessment (Sect. 2.1), before discussing privacy, data privacy and how Contextual Integrity can assist in assessing privacy risks in Sect. 3. This is followed by details of the case study in Sect. 4, outlining the action intervention for implementing GDPR and creating the DPIA framework aimed at SMEs and the charity sector. Finally, we conclude and outline directions for future work in Sect. 5.

2 General Data Protection Regulation (GDPR)

GDPR Article 5 sets out 6 Principles (P): (P1) *Lawfulness* i.e. determining and defining the lawful basis for processing the data; *Fairness* i.e. processing the data fairly with data subjects interest in mind; and *Transparency* i.e. specifying the data to be collected and why, while keeping the data subject(s) informed of how their data will be used. (P2); *Purpose Limitation* i.e. collecting only relevant and necessary data, and processing such data fairly with data subjects interest in mind. (P3) *Data Minimisation* i.e. collecting minimum data, and only collecting data necessary for the specified purpose. (P4) *Accuracy* i.e. keeping the data up to date and correct. (P5) *Storage Limitation* i.e. retaining the data no longer than necessary, and (P6) *Integrity and Confidentiality* i.e. protecting, processing and storing the data securely, and ensuring data is protected from harm, unauthorised or unlawful access.

Data Protection Officers (DPOs) and/or the Data Controller ensure organisations implement appropriate technical or procedural measures to ensure and demonstrate compliance (GDPR, Article 24). To this end organisations must adopt a privacy first policy (DPbDD, GDPR, Article 25), maintain a record of processing activities (GDPR, Article 30), and implement appropriate security measures to protect the data (GDPR, Article 32). Under GDPR Article 35, any processing likely to pose a high risk to the rights and freedoms of the data subject must be assessed. This obliges organisations to assess risks, not from an organisational perspective, but from the perspective of the data subject (the individual). This is the area that this study seeks to address.

2.1 Risk

GDPR asks that organisations must conduct DPIAs for any *high risk* processing activities. High risk processing refers to any large scale processing of personal data. This includes tracking, monitoring, profiling, implementing new technologies, or processing genetic, biometric or special category data (e.g. data relating to health or criminal records) on a large scale (GDPR, Article 35). Processing refers to: *"any operation or set of operations which is performed on personal data or on sets of personal data ... such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available"* (GDPR, Article 4(2)). The Data Subject is the person whose data is being processed, while the Data Controller is the legal entity responsible for making decisions about how the data is processed, this includes any: *"natural or legal person, public authority, agency or other body"* (GDPR, Article 4(7)). Where a third-party processes the data on behalf of the Data Controller, they are referred to as the Data Processor (GDPR, Article 4(8)) or, if a partner organisation jointly manages the data with the Data Controller, they may be the Joint Data Controller.

Conducting a DPIA involves assessing privacy risk. Assessing risk is an integral part of business processes, and helps organisations make informed decisions. However, for privacy, this is usually an extension of assessing security risk. Organisations can use several internationally recognised frameworks for conducting structured risk assessments such as the National Institute of Standards and Technology (NIST) risk framework [22,23], and the International Office of Standardisation's (ISO) [4,11] and [16]. However, these frameworks focus on organisational risk, and don't satisfy GDPR's requirement for assessing risks to the data subject (the individual).

3 Privacy and Contextual Integrity

Our right to privacy as a concept is not a new idea, as early as 1890, Warren and Brandeis discussed *the right to be let alone* [29], while Westin framed privacy from the perspective of the right to control personal information and, in doing so, recognised the context dependent value of information [30]. This idea has since been elaborated and expanded upon. Some refer to privacy as a fluid concept with blurred boundaries [24] or, a contested concept with many facets [18], depending on the context within which it is viewed [28]. Thus, privacy is subjective; every individual has their own view of what privacy is and 'tolerance' (values) or norms of what they consider 'normal' or 'acceptable' when it comes to their privacy [21].

Context has been previously considered as part of a privacy assessment. For example, Solove [28] divides privacy into four broad groups: *Invasions*, *Information collection*, *Information processing*, and *Dissemination*, while Mulligan et al. [18] divide privacy into "five meta-dimensions of theory, protection, harm, provision and scope", sub-divided into 14 sub-dimensions that consider privacy in terms of risk or potential harm. This is akin to security threat modelling, which could assist software design teams in aligning threat modelling with privacy. Ultimately, privacy must be integrated into organisational decision making and thus built into corporate practice [1], which GDPR seeks to achieve through the introduction of DPbDD.

These frameworks consider context but, context is not just about how organisations perceive data privacy. Perceptions and behaviours help people shape what privacy is to them and therefore, when it comes to information and data privacy, their values and norms influence how they perceive data privacy [20]. This can be observed by the choices individuals make about whether or not to share information, how they share information, and why. Some are comfortable sharing very personal details on social media, others are more selective about what they share, and others avoid sharing any information at all. Therefore, there is a difference between what a person chooses to share about themselves and what is shared by others about them i.e. WHO is doing the sharing [20]. Someone may accept their friend sharing their photo within a social circle on Facebook, but not so, if that same photo was shared with the government or their employer given the possible unintended consequences [26,28].

Contextual Integrity (CI) [21] accounts for these previously described contextual nuances. CI considers privacy in terms of data flows, proposing that data privacy should be concerned with how data flows between stakeholders ("transmission principles"), combined with the context within which the data is transmitted. This means that, when it comes to data privacy and how personal data is processed by government departments and organisations, they should primarily be concerned with the individual's "right to an appropriate flow of information" [21]. Thus, CI encompasses all the aspects discussed by Solove [28] and Mulligan et al. [18] but frames these nicely within a theoretical framework devised for decision making and assessing privacy risks in data.

CI assesses privacy risks through three key elements: *Explanation* looks at the current status quo, what the prevailing context is, and how data is used, transmitted, and by whom, *Evaluation* assesses how the data will be transmitted in the proposed new flow, by whom and how this changes the context, and *Prescription* decides if a decision can be made about whether or not the changed flow increase or decrease the privacy risks. Within each of these key elements, the risks are evaluated by looking at privacy from four perspectives: *Actors* (the data- subject(s), sender(s) and receiver(s)), *Attributes* (the individual data items), the *Transmission Principles* (how data is distributed and shared), and the *Context*, i.e. by considering the established norms and values of the actors and society and how these might influence or affect the information flows. For example, actors should be evaluated in relation to their social and job role, the activities of each role, and the values and norms expected of that role. There may also be contrasting duties, prerogatives or obligations associated with one of those roles that could undermine the relationship between the data subject and the person processing the data. Thus, like Mulligan et al. [18], CI views privacy through a risk lens, but focuses on decision making rather than threats and protection.

CI has been used in theoretical discussion about its applicability to a particular scenario or situation [6], although there have been some attempts to consider how CI might be applied in practice. For example, CI has been used to consider appropriate access controls for information flows in system design [2], how attaching tags in message headers can preserve privacy [17], and whether particular practices or sites provide sufficient privacy protection [10,27].

CI has also been used to inform decision making around high level privacy goals for a user community [7], and assessing privacy risks associated with publishing open data [12], which found that organisations consider the data and the attributes when assessing privacy, but fail to take account of the context within which the data is processed. However, by applying CI and also considering the context, more informed decisions could help facilitate the publication decisions. We extended on this work in a case study where we sought to incorporate CI into a DPIA as part of a GDPR implementation process, this is discussed in the next section.

4 Case Study

In this section, we present a case study of an exemplar approach to GDPR implementation in the Charity Sector. The implementation of GDPR will also incorporate the design and creation of a DPIA framework, the DPIA Data Wheel, aimed at this sector and SMEs.

4.1 Background

Most charities rely on public generosity for funding and in-kind support from volunteers to function, with many struggling to raise enough funding to meet all the objectives for their cause. Much work is conducted by volunteers meaning that, even though a charity may collect and manage personal data, they often lack the resources and expertise to assess themselves against legal regulations.

The UK Information Commissioners Office (ICO) has issued some guidance on GDPR to help organisations implement the regulation, but this is so general as to be applicable to all types of organisations [13]. No sector specific guidance is available for the charitable sector, despite requests from the sector for more specific guidelines to be produced [15]. We, therefore, decided this sector would benefit from some assistance and chose to work with a local charity (‘the Charity’) to provide an exemplar approach to GDPR implementation.

The Charity supports those suffering from addiction and substance misuse. It collects personal data from clients to provide them with the care and assistance for dealing with or overcoming their problems. The Charity also needs to ensure data collection and processing satisfies data protection laws, and they rely on several procedures to ensure all processes comply with requirements laid down by legislation such as GDPR and the Care Act 2014. The Charity shares some of the data collected from and about clients with external stakeholders. These may be clinicians and professionals who work with the charity and their clients in providing treatment and advice, or Governing bodies they are legally obliged to share data with, e.g. the Care Quality Commission (CQC) that regulate health and social care in England [5] or the National Drug Evidence Centre (NDEC) that collates statistics on adult addiction users and their treatment [19].

4.2 Approach

We worked with two managers and 29 staff and volunteers who work for the Charity. The case study was conducted over three months and incorporated three staff training sessions and a workshop for a group of 40 other local charities to disseminate the results and evaluate the DPIA Data Wheel. Ethics approval for this case study was sought and granted from the University Ethics Committee.

The research questions (RQ) we asked were: *what data holdings does the Charity have, where and how are these handled currently and to what extent do these comply with GDPR standards?* (RQ1); *what processes does the organisation need to put in place for effective GDPR implementation to demonstrate GDPR compliance?* (RQ2); and *how can the organisation ensure they have in place appropriate processes conducting DPIAs going forward?* (RQ3). The hypothesis supporting these questions and the full methodology can be found at[1].

This project was conducted as an action intervention case study [32], with the unit of analysis being the Charity as GDPR affects all aspects of the organisational processing of data. The first step was to make a detailed GDPR implementation plan evaluating the Charity's readiness, and its ability to achieve DPbDD and demonstrate GDPR compliance to the ICO. The case study was conducted in four phases, each will be described in more detail in the below sub-sections.

4.3 Phase 1 - Data Holdings

We decided that a draft data register would answer RQ1 and help the Charity achieve DPbDD. Therefore, the first step entailed understanding what data the Charity held and how this was processed. This would establish a baseline of what data is collected and how this data is processed within the Charity.

Two parallel pieces of work were carried out: establishing what forms were used within the Charity to collect data, and collecting staff stories. Storytelling as a research method involves collecting narratives or stories to understand people,

[1] https://github.com/JaneHB/DPIA-CS-Protocol.

their actions and ideas. For this study, this would entail staff recounting how they process data as part of their working day using a user story methodology [3,25].

To determine what forms were used within the Charity, the project started with a meeting to discover more about where the Charity were with their GDPR implementation and establish what data was collected and processed within the Charity. This was a very informative meeting which formed the basis upon which the rest of the project was based. It also became evident that the majority of data collection and processing was paper based, securely stored in locked cabinets when not in use. As part of this meeting, copies of the various forms in use as part of the daily operations of the Charity were provided to the research team.

These forms were used as the basis for creating a draft data register containing details of each attribute (individual data item) collected and categorising these based on data sensitivity. This draft register was then further elaborated upon with the information obtained from the parallel piece of work, collecting staff stories.

To collect the staff stories, a spreadsheet was created with 9 columns to capture details of who staff communicate with, what data is communicated, how the communication takes place (e.g. paper or electronic), the regularity of the communication, how long each communication takes, how demanding the staff member finds each communication to be, and whether the communication interferes with or interrupts other duties. The questions asked can be found in the protocol (see[2]). These spreadsheets were circulated to all staff, with 21 staff members respondents, working in eleven different roles. The gender of the respondents was well balanced with approximately half of the respondents being male (9) and half female (10), two respondents chose not to provide gender details.

Some staff completed their stories with few words using one line sentences while others were more descriptive in their stories. Therefore, once the staff stories had been collated, the completed staff stories were returned with an additional column containing questions that sought clarification on different aspects of the staff stories. For example, different staff members referred to different forms using alternate names than those initially collected making it necessary to clarify terminology or confirm which terminology related to each form.

4.4 Phase 2 - Analysis of Data Holdings

The staff stories were analysed to update the draft data register, confirm the list of forms used within the Charity, and gain an overview of how the data travels (the data flows) both internally and externally. From this, it became clear that there were more forms used than originally collected as part of the initial meeting. Consequently, a second meeting was scheduled to update the list of forms, and seek clarification on some of the terminology used; e.g. the various forms used were referred to in different ways by different staff members. At this

[2] Ibid 1.

second meeting, the research team was granted access to the form templates used within the Charity.

Life of the form. Comparing the template forms collected and the staff stories showed the client assessment and care plan forms were the forms containing personal data that was processed most regularly. Both were living documents that included detailed personal information. These included a full medical history (mental and physical), details of a client's social, personal and cultural background, and a list of historic and current professionals responsible or involved with their care. Both documents form part of the contract between the client and the Charity.

These two forms were chosen to capture the data journey through the "life of the form" exercise. This data collection involved another spreadsheet (the "life of the form") devised to investigate in more detail how the data travels, i.e. how these forms are used, transmitted or shared both internally within the Charity and externally with other stakeholders. This spreadsheet asked a series of questions about the journeys the data might take such as where the form that collects the data was born (see[3] for full list).

A column was created within the spreadsheet for each sub-form. Creating multiple columns would allow separate elements to go on different journeys. For example, a page or sub-form may be removed or shared for specific purposes such as faxing to external key professional staff involved in the care of the client, could be captured as part of one of the journeys. The spreadsheet supported up to 10 journeys for each sub-form. The life of the form spreadsheets for the Care Plan and Client Assessment were sent to the CEO and the manager of one of the Charity's houses to be completed with details of how the data travels during its lifecycle.

Analysis. The staff stories were compared with the forms to identify any missing forms, and establish patterns of data flow. This revealed that records pertaining to the client's medication and the client register were the most frequently referred to documents. Moreover, various methods of communication between staff and other stakeholders were mentioned as part of the staff stories. This information was then compared to the completed life of the form spreadsheets to provide a more detailed overview of the data, how it was used and the "journey" each form went on during its life cycle.

This analysis showed that the Charity collect a variety of personal or special category data from their clients that require a legal basis for processing under both Article 6 and Article 9 of GDPR. This included details relating to health, religion and beliefs. The analysis also highlighted a number of common processes and procedures undertaken by staff as part of their daily work, or the form's journey. These were broken down into data relating to clients and data relating to staff and data processing processes.

[3] Ibid 1.

Master Data Register. This information was then used to turn the draft data register into a Master Data Register (MDR) providing details of all the Charity's data holdings. The data included within the MDR was informed by the draft register, the information gleaned from the staff stories and the list of forms downloaded from the second meeting, listing all the individual attributes from each form and categorising these based on level of sensitivity of the data. Personally identifiable data was classed as personal data in accordance with GDPR Article 6, while most of the data collected and classed as sensitive was classed as "special category" data in accordance with GDPR Article 9. The initial data categorisation was based on "best guess" from the information available. These categories were then evaluated by the Charity CEO who verified or changed each of the categories according to the Charity's perspective. The MDR also sought to include several other pieces of information including details of the Data Controller, a justification for collecting each piece of data, details of the processing being carried out, how the data will be stored, and the storage period etc.[4]

The final MDR contained 997 individual data items, categorised according to data sensitivity and justified based on relevant legislation or contractual obligations to facilitate the clients' (*data subjects*) treatment needs. Creating this MDR answered RQ1 and provided the Charity with their starting point towards demonstrating compliance with the obligation to keep "records of processing activities" (GDPR, Article 30).

4.5 Phase 3 - GDPR Process Guidance

Phase three sought to answer RQ2, and involved assessing existing processes and practices to determine how these could be revised to ensure GDPR compliance. The work in this phase centred on reviewing policies and protocols and preparing the supporting documentation and processes necessary for the Charity to demonstrate GDPR compliance. The Charity's privacy policies were reviewed and revised, together with the process for obtaining consent from Clients; and a process for responding to data subject requests for access, erasure and data portability was also created.

Privacy Policies and Data Subjects' Rights. The existing privacy policy given to clients by the Charity was in line with Data Protection Act 1998, but failed to meet the requirement of expressing clearly, in plain language, what data the Charity collect from clients and how this is used. Therefore, a new policy was devised to meet these requirements. This was presented in plain language and includes details of what data is collected, how the data is collected and used, who the data is shared with, how the data is safeguarded, the timeframe for storing the data and details of the data subjects rights in relation to their data. To compliment the new policies, the Charity agreed to create a protocol for dealing with and responding to clients seeking to invoke their rights (e.g. requests for access, erasure and data portability etc.). This ensured a thorough, repeatable

[4] Ibid 1.

procedure was in place to deal with a data subject (client) invoking their rights under GDPR. The privacy policy for staff was also updated to ensure staff are fully aware of their obligations under GDPR.

Consent. The issue of consent is a potential problem for the Charity. It works with vulnerable adults who may initially give consent to processing, but later withdraw their consent or even claim consent was not freely given. For example, a client may claim that they were not capable of giving informed consent at the time or claim they lacked sufficient mental capacity to freely give consent. Thus, there is potential that the Charity's clients (or someone else on their behalf) may argue that consent was not freely given, because there is a power imbalance between the client ("the data subject") and the Charity (as "the data controller") providing the client with treatment and thus, exercising a level of control over the clients and their actions while under their care (GDPR, Article 7). To address this, a meeting was convened to understand precisely what consent was collected from clients (data subjects), how this was collected and used, and what procedures allowed clients to withdraw their consent. Following this meeting and careful study of the legislation, the solution was providing more clarity for the legal basis for processing the data in the first place.

The Charity only processes data to provide effective treatment to their clients as required under the Care Act 2014 (CA). Although the Charity ensures informed consent is obtained from all clients, this is not the main legal grounds for processing the data. When enrolling for treatment, clients complete a "Client Assessment" and "Care Plan". Both documents subsequently form part of the contract between clients and the Charity. The data collected is gathered to satisfy a legal requirement to assess the clients needs prior to and, as part of, providing treatment to vulnerable adults (CA, s. 9), making it necessary for the Charity to provide effective treatment; they cannot help clients without the full history of their addiction and the surrounding circumstances.

The Charity can therefore argue that the processing is necessary for compliance with a legal obligation (GDPR, Article 6(1c)), or the primary legal basis for processing the date is contractual (GDPR, Article 6(1b)) because they cannot perform their work without this information. However, this does not mean that consent is not still required; some aspects of sharing the data may not be required to perform the contract. For example, family members may wish to be kept informed of how the client responds to treatment, which is not a prerequisite requirement for providing treatment. Therefore, for those aspects, informed consent remains required from the client for this type of secondary sharing of the data (GDPR, Article 7(2)). This means the Charity remains compliant provided clear instructions are given that are "clearly distinguishable" from other types of data processing, and provide an easy means for amending or withdrawing consent settings. To this end, the Charity, as part of the contract, would obtain granular informed consent for who they may or may not divulge information to from the client. In addition, a granular "withdraw consent" section was added to the consent form, allowing clients to withdraw easily.

In addition staff training was arranged to inform staff about GDPR, consent and the new protocols, e.g. what these mean for the organisation, for them as staff, and as individuals. The training sessions were designed to make staff think about how they process data as part of their daily work. The training was positively received with one participant commenting; *I will be mindful and start to prompt colleagues around having data around* (P23), suggesting this exercise is likely to positively impact on their behaviour in dealing with data in future.

4.6 Phase 4 - The DPIA Data Wheel

The final phase of the project sought to answer RQ3 by creating a DPIA process for assessing privacy risks. The DPIA was devised based on previous work on assessing privacy risk for open data [12], GDPR, and guidance provided by the ICO on how to conduct DPIAs [14]. The resulting DPIA framework, named "the DPIA Data Wheel", is a step-by-step guide that takes assessors through the process of conducting a DPIA (see Fig. 1, and[5]). The DPIA Data Wheel incorporates questions about the *prevailing* and *surrounding* context to ensure the wider implications of data processing are considered. Moreover, by including the Data Register and the life of the form questionnaire devised and used in Phases 1 and 2, we have facilitated the gathering of comprehensive background information about the data, the actors and the transmission principles (*data flows*) to inform the risk assessment in the Data Wheel. This provides a mechanism that any organisation can use for establishing their own data register and detailed data journeys, thereby acting as a starting point for their own GDPR implementation.

The DPIA Data Wheel asks a series of questions relating to the data devised to provide a full overview of the system, process or project being assessed. The full DPIA Data Wheel is presented in a spreadsheet consisting of 5 tabs, the last containing the various drop-down lists within the spreadsheet. For information purposes, this has been left so practitioners can view this information. These are:

Tab 1: Need for a DPIA. This is the starting point for conducting the DPIA, and consists of a short assessment to help the practitioner determine whether or not a DPIA is required for the system, project or process under review.

Tab 2: Data Wheel. Where a DPIA is required, the Data Wheel is the privacy risk assessment for the process, system or project. As part of this, the DATA part of the Wheel provides the *explanation* [21], while the WHEEL forms the beginning of the risk assessment (the *evaluation*). Practitioners are asked to consider different aspects of the process, system or project including what data they plan to capture (*the "data"*), the people who will process the data (*the "actors"*) and the context within which data is processed (thereby embedding the *context* element of *"CI"*);

Tab 3: Data Register. This was derived from the draft data register created as part of Phase 1. Practitioners are asked to provide more specific and granular

[5] Ibid 1.

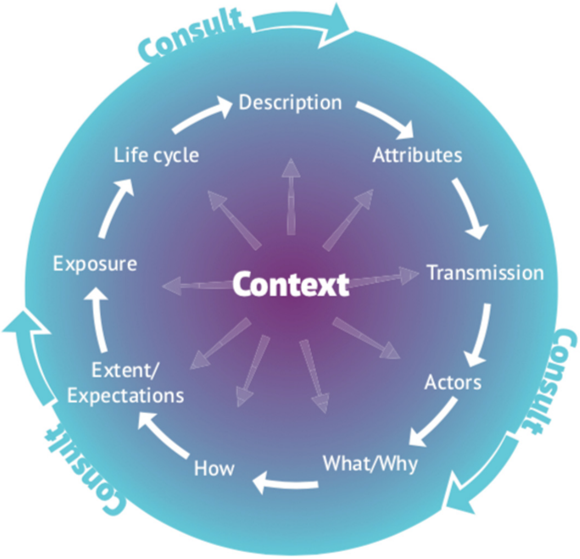


Fig. 1. DPIA Data Wheel

details about the data attributes (individual data items) that they plan to process (*the "data"*). The information gathered here is intended for use to compliment and help inform the risk assessment on Tab 2. It was designed to form part of the organisation's Master Data Register, thereby helping them maintain an accurate overview of the organisation's data holdings;

Tab 4: Life of the form. The questions here were derived from the "life of the form" part of the project. It was included to make practitioners think about how the data travels within their organisation. (*the "transmission principles"*). By considering the 'journey' the data within the system, project or process is likely to take during its lifetime, practitioners will be able to glean valuable insight into where there may be potential risks that will need to be mitigated against;

Tab 5: List. This contains list of all the drop-down menus that form part of the assessment on the other tabs.

The final aspect of the DPIA framework is the "consult" element. This element is not present in the DPIA Data Wheel spreadsheet as this involves ensuring that all relevant stakeholders with a potential interest or input into the process, system or project are consulted on the privacy risks as far as is possible. In the case study, this element was completed through the staff training sessions where the risks identified by management as part of completing the DPIA framework. This served two purposes. First, it allowed the research team to evaluate the effectiveness of the DPIA Data Wheel. Second, it helped avoid "resistance to change", which is a common reaction of staff when any form of change is introduced within an organisation [8]. This is discussed in the next section.

Evaluating the DPIA Framework. To evaluate the DPIA Framework, a DPIA Data Wheel spreadsheet was created for the "Care Plan" and "Client Assessment". On each DPIA, the Data Register and the Life of the Form tabs were pre-populated with the information provided as part of Phase 1 and 2, i.e. the list of attributes collated from the staff stories and the forms provided that the Charity use and the completed life of the form answers that the CEO and House Manager had provided.

The completed DPIA was evaluated in three ways First, by the CEO and the House Manager, who reviewed the DPIA Data Wheel, the Data Register and the Risk Register. Second, as part of the staff training session, the Risk Register was reviewed and evaluated with further risks added. Third, the DPIA Data Wheel was reviewed as part of a workshop where delegates from 40 local charities reviewed the Data Wheel and the Risk Register at an interactive workshop. As well as enabling the evaluation of the DPIA Data Wheel, it also enabled the final "consult" element of the DPIA framework to be achieved by taking the evaluation to stakeholders.

Following the first evaluation, several changes were made to the DPIA framework. Some questions were reworded slightly in the Data Wheel; one question was removed as a duplicate, and another added. In the Data Register, more columns were added for inputting justifications as one was not always sufficient. There can be more than one reason for why a particular attribute is collected, and the Charity wanted to capture these to strengthen their case for justification.

The second evaluation took place during three staff training sessions. These informed staff of the changes introduced by GDPR and provided consultation on the risks identified by senior staff when completing the DPIA Data Wheel. These sessions served as part of the DPIA consultation accounting for internal stakeholders, and resulted in several additional risks being identified that had not been included in the initial completion of the DPIA framework.

In the third evaluation, a group of industry or sector peers served as an external body of stakeholders in reviewing the DPIA Data Wheel. At the workshop, delegates were divided into four groups with each group reviewing the same DPIA Data Wheel. This produced a series of additional risks that had not been included in previous evaluations. Some were generic threats that were relevant to the Charity, such as failure to lock storage cabinets holding data. Others, such as the risk of not informing trustees of a breach, failure to delete data, or insider threat by staff, could be applied more generally across the industry sector.

These sessions resulted in 88 different risks being identified and suggested mitigation strategies for each of these recorded. The staff participants particularly appreciated *the application to our work and potential risks* (P9), while one workshop participant commented that the workshop provided *thought provoking and practical information* (P33). Interestingly, in all of the evaluation sessions, all of the threats identified were related to the organisation and how they should safeguard data rather than to the data subjects themselves, despite the Risk Register specifically having separate columns for risks to be identified for the data subject as well as the organisation. In hindsight, this was to be expected as all previous work and guidelines has concentrated on security and how to safeguard systems and processes. What it did show, however, is that more work is needed to educate practitioners on the need to separate privacy from the perspective of the data subject when assessing privacy risks. Future work will look at this element and how this can best be achieved.

5 Conclusion

In this paper we have answered RQ1 by creating a Master Data Register for the Charity and established how the data is transmitted through the life of the form exercise that recorded how the data travels during its lifecycle. For RQ2 we reviewed and revised the Charity's privacy policies, provided staff training on GDPR and the DPIA risk assessment and arranged for new protocols to be devised to facilitate dealing with data subjects invoking their rights under GDPR. Finally, in creating the DPIA Data Wheel, a standardised DPIA process based on CI, we answered RQ3. The main findings are that CI can be successfully applied to DPIAs and GDPR implementations, although more emphasis needs to be placed on the fact that risks should be assessed from the data subject's perspective rather than the organisation. However, our results do demonstrate how CI can be embedded into DPbDD and the DPIA process to provide the means for other charities and SMEs to be able to use the DPIA Data Wheel to assist in their own GDPR implementation and conduct comprehensive and repeatable DPIAs going forward.

This paper has provided three contributions. First, an exemplar model was presented to illustrate how SMEs and charitable organisations can implement GDPR. Second, we presented the DPIA Data Wheel: a repeatable DPIA framework that facilitates repeatable, consistent privacy risk assessments within an organisation. Finally, we demonstrated how CI can be used to facilitate practical decision making by incorporating the CI concepts into DPIAs.

Future work will examine how these concepts can be developed and strengthened to better guide SME and charity practitioners in assessing privacy risks from the individual's perspective. This in turn will help both SMEs and charities to better safeguard the data subject's privacy from the organisational viewpoint.

Acknowledgments. This work was funded by the Bournemouth University Charity Impact Funding scheme at Bournemouth University.

References

1. Bamberger, K.A., Mulligan, D.K.: Privacy on the Ground: Driving Corporate Behaviour in the United States and Europe. The MIT Press/Massachusetts Institute of Technology, London (2015)
2. Barth, A., Anupam, D., Mitchell, J.C., Nissenbaum, H.F.: Privacy and contextual integrity: framework and applications. In: 2006 Symposium on Security and Privacy [Serial Online], vol. 2006, pp. 184–198. IEEE Xplore Digital Library, Ipswich (2006). https://doi.org/10.1109/SP.2006.32. Cited by 0

3. Bruner, J.S.: Actual Minds. Possible Worlds. Harvard University Press, Cambridge (1986). [Electronic resource]
4. BS ISO 31000:2009: British standards document BS ISO 31000:2009: Risk management. Principles and guidelines. Technical report, British Standard and the International Organization for Standardization (ISO) (2009)
5. Care Quality Commission (CQC): Care Quality Commission (2018). https://www.cqc.org.uk/
6. Conley, A., Datta, A., Helen, N., Sharma, D.: Sustaining privacy and open justice in the transition to online court records: a multidisciplinary inquiry. Maryland Law Rev. **71**(3), 772–847 (2012)
7. Darakhshan, J., Shvartzshnaider, Y., Latonero, M.: It takes a village: a community based participatory framework for privacy design. In: 2018 IEEE European Symposium on Security and Privacy Workshops, EUROSPW, pp. 112–115 (2018)
8. Demirci, A.E.: Change-specific cynicism as a determinant of employee resistance to change. Is, Guc: J. Ind. Relat. Hum. Resour. **18**(4), 1–20 (2016)
9. European Parliament and the Council of Europe: General data protection regulation (GDPR). Regulation (EU) 2016/679 5419/1/16. European Parliament and the Council of Europe, Brussels, April 2016
10. Grodzinsky, F.S., Tavani, H.T.: Privacy in "the cloud": applying Nissenbaum's theory of contextual integrity. SIGCAS Comput. Soc. **41**(1), 38–47 (2011)
11. Hall, D.C.: Making risk assessments more comparable and repeatable. Syst. Eng. **14**(2), 173–179 (2011)
12. Henriksen-Bulmer, J., Faily, S.: Applying contextual integrity to open data publishing. In: Proceedings of the 31st British HCI Group Annual Conference on People and Computers: Digital Make Believe. British Computer Society (2017)
13. ICO: Preparing for the general data protection regulation (GDPR): 12 steps to take now. Technical report, V2.0 20170525, Information Commissioner's Office, May 2017
14. ICO: Data protection impact assessments (DPIAs) (2018)
15. ICO: General data protection regulation (GDPR) FAQs for charities (2018). https://ico.org.uk/for-organisations/charity/charities-faqs/
16. ISO/IEC 29100: BS ISO/IEC29100: Information technology – security techniques – privacy framework. Technical report, British Standard and the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) (2011)
17. Krupa, Y., Vercouter, L.: Handling privacy as contextual integrity in decentralized virtual communities: the privacias framework. Web Intell. Agent Syst. **10**(1), 105–116 (2012)
18. Mulligan, D.K., Koopman, C., Doty, N.: Privacy is an essentially contested concept: a multi-dimensional analytic for mapping privacy. Philos. Trans. Ser. A Math. Phys. Eng. Sci. **374**(2083), 20160118 (2016)
19. National Drug Evidence Centre: National drug treatment monitoring system (NDTMS) (2018)
20. Nissenbaum, H.: Privacy as contextual integrity. Wash. Law Rev. **79**(1), 119–158 (2004)
21. Nissenbaum, H.F.: Privacy in Context: Technology, Policy, and the Integrity of Social Life. Stanford Law Books, Stanford (2010)
22. NIST: Guide to protecting the confidentiality of personally identifiable information (PII). Technical Report, National Institute of Standards and Technology (NIST), U.S. Department of Commerce, pp. 800–122 (2010)

23. NIST: Guide for conducting risk assessments. Technical Report SP 800-30, National Institute of Standards and Technology (NIST), U.S. Department of Commerce, Gaithersburg, September 2012
24. Palen, L., Dourish, P.: Unpacking 'privacy' for a networked world. In: CHI-CONFERENCE, pp. 129–136 (2003)
25. Rooney, T., Lawlor, K., Rohan, E.: Telling tales: storytelling as a methodological approach in research. Electron. J. Bus. Res. Methods **14**(2), 147–156 (2016)
26. Sanchez Abril, P., Levin, A., Del Riego, A.: Blurred boundaries: social media privacy and the twenty-first-century employee. Am. Bus. Law J. **49**(1), 63–124 (2012)
27. Sar, R.K., Al-Saggaf, Y.: Contextual integrity's decision heuristic and the tracking by social network sites. Ethics Inf. Technol. **16**(1), 15–26 (2013)
28. Solove, D.J.: A taxonomy of privacy. Univ. Pennsylvania Law Rev. **154**(3), 477–564 (2006)
29. Warren, S.D., Brandeis, L.D.: The right to privacy. Harvard Law Rev. **IV**(5), 193–220 (1890)
30. Westin, A.F.: Science, privacy, and freedom: issues and proposals for the 1970's. Part I-the current impact of surveillance on privacy. Columbia Law Rev. **66**(6), 1003–1050 (1966)
31. Data protection act 2018, May 2018. http://www.parliament.uk/
32. Yin, R.K.: Case Study Research : Design and Methods. SAGE, Los Angeles (2013)

Me and My Robot - Sharing Information with a New Friend

Tanja Heuer(✉), Ina Schiering, and Reinhard Gerndt

Ostfalia University of Applied Sciences, Wolfenbüttel, Germany
{ta.heuer,i.schiering,r.gerndt}@ostfalia.de

Abstract. This paper investigates user perception regarding social robots and personal information disclosure. During a study two participant groups stated their attitude towards functionality, shared personal information and the interest of transparency and intervenability. The impact of technical background knowledge regarding users attitude and perception was examined. Participants working with robots have a more open-minded attitude to share personal information achieving a wider range of functionality. Both groups care about transparency of collected data and the possibility of intervenability.

Keywords: Social robot · HRI · Human-Robot Interaction · Privacy

1 Introduction

Since Amazon Echo, Google Home and iRobot Roomba are already part of daily life, Asus Zenbo is the next generation robot entering homes. Although the current focus on social robots is on care and most studies are conducted with older people or children with chronic impairments [15], the features of social robots are not only addressing *people in need of care*, but everyone who wants to have support and assistance at home or a hand's free way to read news or listening to music when, e.g. the phone is not within reach. To contribute assistance, it communicates with other smart devices, informs about news and appointments, supports music streaming, helps cooking and supervises the health care status. To assist in all these areas, robots make use of sensors like cameras or microphones. To show itself as assistant and companion, the robot collects a huge amount of personal data to be able to simulate a natural interaction. This collected data is often processed using cloud services to allow a fast response and reaction time on requests.

Using smart home technologies and devices leads to a variety of possible privacy risks. With a microphone there is a risk of eavesdropping. A camera increases the risk of spying. Cloud service connections enable unauthorized access to personal data shared with the devices. With every additional sensory input a list of new hazards appear. It is important to make users aware of potentially risks and of the implication of disclosing personal information. To achieve this

E. Kosta et al. (Eds.): Privacy and Identity 2018, IFIP AICT 547, pp. 189–204, 2019.
https://doi.org/10.1007/978-3-030-16744-8_13

awareness, users need to know which data is required for specific features and for what purposes the data is processed.

To investigate users preferences regarding functionality and shared personal information, a questionnaire is conducted as a first step. Furthermore it should be examined, if users are interested in the operating principles of robots to get an idea how and for what purposes personal information is processed. Even though, a questionnaire is not as representative as conducting user studies because of effects as the privacy paradox - sharing attitude and sharing behavior [10], the aim is to get a predication of younger people's attitude towards social robots and how technical background knowledge affects the way of thinking regarding robot expectations. In a second step, the correlation between features and personal information is investigated. The importance of transparency in the context of this relation needs to be outlined unambiguously.

2 Background Social Robots

We are facing a world of smart homes and connected things. Amazon Echo and Google Home are already part of it and this constitutes several risks. In May 2018, Amazon Echo recorded a private conversation and transferred that to a friend. Thereupon, Amazon argued that there were bad circumstances. Already in 2017, it was determined that a previous version of Google Home was permanently eavesdropping its users. The information was then transferred to a Google server although the Google Home device only should react on *Ok, Google*.

As Amazon Echo and Google Home, also other robots are integrated into the smart home. The vacuum cleaning robot (Roomba980 [1]) can be started via Alexa Echo, saying "Alexa, ask Roomba to start cleaning."[1] or via Google Home saying"Ok Google, tell Roomba to start cleaning"[2]. Figure 1 shows the example scenario using a smart home vacuum cleaning robots. With the help of a camera and a laser range scanner it maps the home and processes information like cleaned and non cleaned areas. Initially, this feature was developed for intelligent cleaning service of the robot and it was processed on the robot and no one had access to that. With an update, iRobot decided to offer a new feature and revealed this map in an application where users are able see the robots status and by default everything is sent to the iRobot Cloud[3]. Furthermore, it was already discussed if this data could be sold to third parties like Amazon or Google because they are cooperating anyway[4].

[1] https://homesupport.irobot.com/app/answers/detail/a_id/1412/~/compatible-commands-for-a-wi-fi-connected-roomba-and-alexa.

[2] https://homesupport.irobot.com/app/answers/detail/a_id/1509/~/compatible-commands-for-a-wi-fi-connected-roomba-and-the-google-assistant.

[3] http://desupport.irobot.com/app/answers/detail/a_id/1406/~/clean-map%E2%84%A2-report-data-and-the-irobot-cloud.

[4] https://www.theguardian.com/technology/2017/jul/25/roomba-maker-could-share-maps-users-homes-google-amazon-apple-irobot-robot-vacuum.

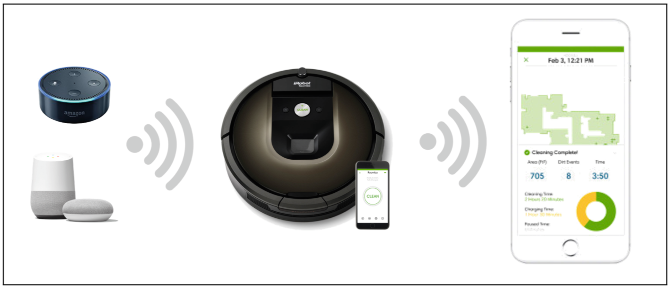


Fig. 1. Smart vacuum cleaning using the Roomba980

A more complex robot for home use is Zenbo, developed by Asus[5]. This robot is marketed with a wide range of functionalities. On the one hand, it is able to remind of medication or meetings, sends notifications to family members in case of emergency and controls other smart devices in the home. On the other hand, it supports online activities like shopping, music and video streaming and searches for recipes. To allow also special functions as e.g. healthcare checks, a broad variety of personal data needs to be collected and the robot needs to be connected to the internet all the time. The collected information includes voice, video and communication records[6] but it is not mentioned explicitly, how privacy and security of this information is ensured.

3 Related Work

As already mentioned, the research of social robots is often linked to the healthcare sector. In many different ways, older people are increasingly integrated into the development process of robots for assistive tasks. As one method, surveys and interviews are conducted to ask about user preferences. Various studies investigated the attitude of people towards robots and in which situations robots or human assistance is preferred [16,19,20,23].

Apart from design and features, some studies dealt with the topic of privacy and possible concerns of users towards social robots. Syrdal et al. [21] focused on data storage and which data users would agree with to be stored. Caine et al. [6] investigated changing behavior of older people when they were recorded. Concerning privacy, more studies focused on the identification of general privacy risks [11,17] or considered privacy concerns in a more general smart home context [7,24]. Aroyo et al. [4] investigated the disclosure of personal information when participants started to trust a robot. Because of smart home systems communicating with the robots (see Fig. 1), it is also worth to take a look at topics investigated in this area. For the healthcare sector, there are smart home solutions [22] and smartphone applications [8] where it is investigated which personal

[5] https://benchmarkreviews.com/41895/asus-zenbo-robot-announced/.
[6] https://www.asus.com/Terms_of_Use_Notice_Privacy_Policy/Privacy_Policy.

data participants are willing to disclose. More general privacy issues using smart home systems are analysed in the context of data tracking and activity monitoring [5,12,14,18].

4 Research Approach

The aim of this survey was to evaluate younger people's technical background knowledge and how this might affect the perception and use of social robots. In doing so, the focus is on privacy concerns because they have a negative implication on the usage of robots [2]. Therefore this influencing determinant needs to be investigated further. Next to preferred features, participants should state their willingness to disclose several personal information.

One the one hand, we want to investigate whether potential users are aware of the connection between use of features and provision of personal information. On the other hand, we want to know if in general users are interested in transparency of the operating principles of a robot, how and which personal data is processed and for what purpose it is used. Because we assume that the interest in transparency and the awareness is related to technical know-how, we investigated two user groups.

5 Methodology

5.1 Questionnaire

Before handing in the survey, participants had to give their informed consent. They were informed about the research topic of the survey and how the collected data will be handled. All collected information is stored anonymously and encrypted and will not be given to third parties.

The questionnaire was conducted on paper. It is arranged in three major sections. The first part determines favored features for a robot. Participants had the ability to choose between different features and second, they could describe additional, not mentioned features. Furthermore, they should state their attitude towards usage of robots by more than one person and if the robot should be able to distinguish between people. The second section asked for preferences according to collected data. Participants should decide what personal information the robot is allowed to collect.

In a third step, the interest of intervention towards data collection and processing should be stated (scale from $-2 =$ strongly disagree to $2 =$ strongly agree). Partly, the questions are adapted from the UTAUT model [3] and a study investigating preferences of older adults [23]. In the last section, users were asked for demographical information of participants as age, sex, education, technical affinity and technical devices which are used regularly.[7]

[7] The whole questionnaire can be downloaded under the following link: https://powerfolder.sonia.de/getlink/fiZbw6TM9E8Vb2aCM6DyEor/Questionnaire_IFIP2018.pdf.

5.2 Participants

In total, quantitative data from 73 participants was collected. Participants for *group 1* were chosen randomly during the RoboCup competition 2018 in Montreal, Canada. Participants for *group 2* were chosen randomly during a social event in Germany also in 2018.

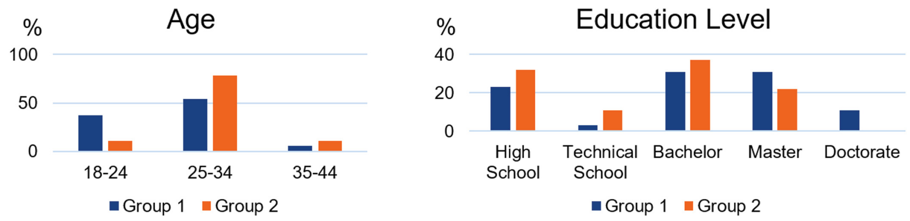


Fig. 2. Age and graduation distribution of both groups

Group 1 consisted of 35 participants - 7 female, 25 male and 3 non-disclosed participants. The age and level of degree of both groups is opposed in 2.37.1% were in the range of 18–24 years, 54.3% were in the range of 25–34 years. The most completed educational qualification in *participant group 1* are Bachelor's degree (31,4%) and Master's degree (31,4%). 22.9% had a high school degree. 68.6% had a student status, 25.7% are working full time. Smartphones (100%) and laptops (97%) are daily used technical devices. Game consoles are regularly used by 28.6% of the respondents, a vacuum cleaning robot by 22,8% of the participants, Alexa Echo and Google Home by only 14.3%.

Group 2 consisted of 38 participants - 16 female, 18 male and 4 non-disclosed participants. 79% were in the range of 25–34 years. The two highest qualification degrees are Bachelor degree (36.8%) and high school degree (31.6%). 63.2% are employed full time, the rest of the respondents were students. Smartphones (100%) and laptops (94.7%) are daily used technical devices. Games consoles and vacuum cleaning robots are regularly used by 15.8% of the participants, Alexa Echo and Google Home by only 10.5%.

Figure 2 illustrates the age distribution and the educational level of both groups.

Additionally, participants were asked to rate their self-evaluation of technical affinity (scale from $-2 =$ strongly disagree to $2 =$ strongly agree):

1. *easy usage:* It would be easy for me using a robot at home.
2. *easy learning:* Learning how to use a home robot would be easy for me.
3. *willingness of usage:* I am willing to use technical devices.
4. *usage is fun:* I have fun using technical devices.
5. *problem solving:* I am able to solve technical problems on my own.

Figure 3 shows the average of level of agreement for technical affinity. For all statements, *group 1* states a higher level of agreement than *group 2*. Whereas the average level of agreement is between 1.3 (easy usage) and 1.8 (willingness of usage) for *group 1*, *group 2* only reaches an average level of agreement between 0.3 (problem solving) and 1.2 (easy learning).

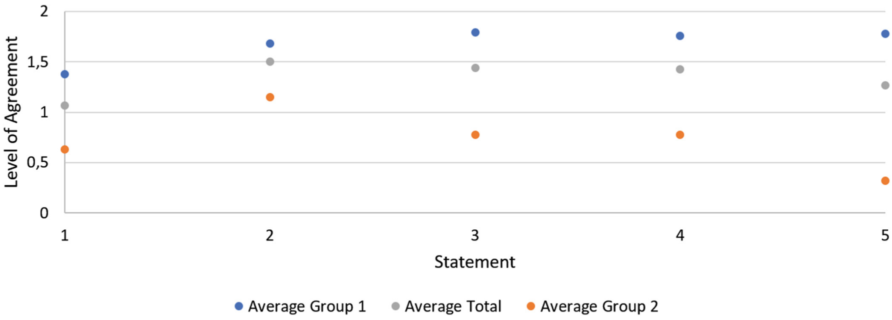


Fig. 3. Technical affinity of participant groups in average

6 Results

6.1 Functionality of a Robot

The first part of the questionnaire addressed applications of the robot. Respondents could select as many features as they like and in a second step add their own ideas of features a robot needs to have. The chosen features are already part of various social robots. As a third part of this category, participants should state, if a robot should be used by more than one person and if it should able to distinguish the users. Prescribed possibilities of features are the following:

- The robot reminds me when to take my medicine.
- The robot keeps an eye on me, possibly calls for rescue.
- The robot takes over cleaning activities.
- The robot keeps me company.
- The robot moves autonomously.
- The robot provides cognitive exercises.
- The robot interfaces with other technologies in my home.
- I can cuddle and hug my robot.
- I video conference with my friends and family via the robot.
- Friends and family can monitor in case of problems.

As a result, the radar chart (see Fig. 4) shows the percentual answers of every feature. It can be seen that both groups have the same peaks. The most chosen features are cleaning assistance, autonomously moving, communication with other technologies in home and video conferencing with other people. But only for cleaning assistance, both groups show the same high interest (>90%).

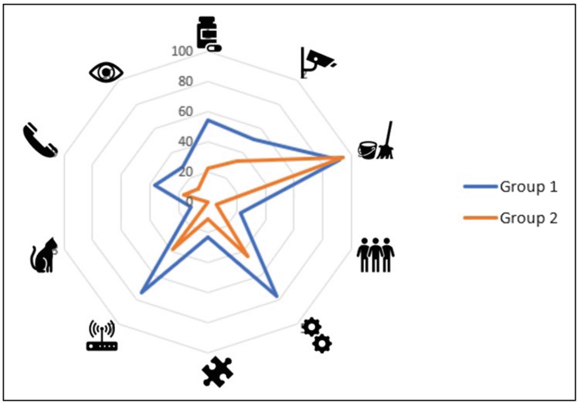


Fig. 4. Possible selectable features for the robot

For all other features, approximately one third less of *group 2* wanted to have these features. Reminder for medication and surveillance in case of emergency are also two interesting features for half of *group 1*.

In a second step, participants were able to state their own ideas. 77% of the participants of *group 1* wrote down their own ideas or specified tasks, in *Group 2* only 50% did. *Group 1* listed specific tasks for cleaning activities like dish washing, floor cleaning, clearing away things, laundry and take out garbage. Additionally, the robot shall be able to cook or help cooking (20%), it shall have the same features as Alexa Echo or Google Home (11%), can remind you of things (11%) and it shall be part of home security - surveillance in case of emergency when no one is at home(11%). In *group 2* the most popular answer was vacuum cleaning. Other answers of *group 2* were a reminder function (11%) and security issues (11%).

It is accepted by 85% of *group 1* and 100% of *group 2* that the robot can be used by everyone at home. The distinction of people at home is important for 92% of *group 1* but only for 60% of *group 2*.

6.2 Disclosed Information

Participants selected personal data they are comfortable with sharing it. The right radar chart (see Fig. 5) shows the answers. *Group 1* is more willing to share personal information with the robot. More than 60% allow the robot to know personal information like name, age and sex but also grocery lists and calenders they would agree to share. Additionally, speech and face recognition will be also allowed by more than 60%. In contrast, most participants of *group 1* would not share their hobbies, login information for social media accounts or websites and their address or GPS location. *Group 2* is more reluctant. The robot should know as little as possible. The most accepted informations to share are calender entries, grocery list and birthdays. The only information, participants of *group 2* would share more often than *group 1* are login informations.

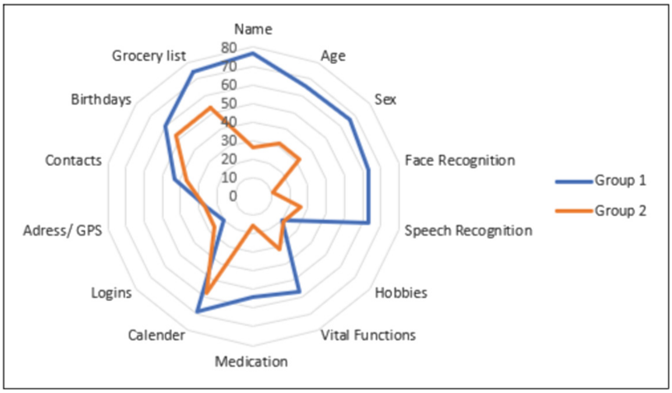


Fig. 5. Personal information, respondents are comfortable with to share with the robot

6.3 Ability of Intervention

In the third section, participants were asked about their interest towards background information of collected personal data. Therefore it will be differentiated between transparency and intervenability. Whereas the transparency questions ask for getting an inside view of collected data, intervenability asks for the ability to modify feature settings and processed information. Transparency is necessary to understand the use and purpose of different features of technical devices. Intervenability allows to exercise the right of informational self-determination by choosing between several data sharing options.

6. I am interested in when the robot is recording.
7. I am interested in what the robot is recording.
8. I am interested in when the robot is storing.
9. I am interested in how the robot uses certain recorded information.

10. I want to decide when the robot is recording.
11. I want to decide what the robot is recording.
12. I want to decide what the robot is storing.
13. I want to be able to turn on/off certain robot functionalities.

6.4 Transparency

Participants would like to have transparency concerning the data robots collect. More than 60% of *group 1* as well as *group 2* agree with most of the statements (see Fig. 6. Whereas the interest in time-dependent personal information for both groups is almost the same, the non-expert *group 2* shows even higher interest in type and intended purpose of personal information. Almost 80% of *group 2* are interested in the data which is collected and they want to know for what reasons and features it is processed and needed. Figure 8 contrasts the average of both groups. It can be seen, that the average of *group 2* is higher for all of the statements 6 to 9 and the questions *how* and *what* information is used, have the highest level of agreement for transparency.

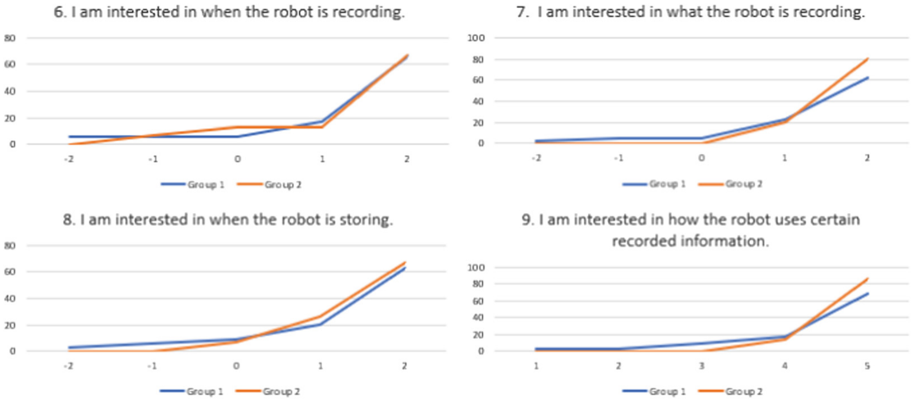


Fig. 6. Transparency for collected and processed personal information

6.5 Intervenability

The opinions towards intervenability of both groups are slightly different. Again, *group 1* showed a high interest in all possible methods of intervention. *Group 2* is less motivated in this category. They showed a great interest in transparency of personal information, but they do not want to have the ability to entirely control their data. It should be possible to decide when and what personal information is collected, but only 20% care about the storage of this data and for them it is not necessary to turn on/off certain features if they are privacy relevant for example. Again here, only 20% showed an interest for this opportunity. As an interesting factor, Fig. 8 shows that the average of question 12 and 13 is higher for *group 2* even though Fig. 7 might lead to a different result. This is affected by negative results of *group 1*. Some participants of the robotic group strongly disagree with these statements, whereas no one of the other group did.

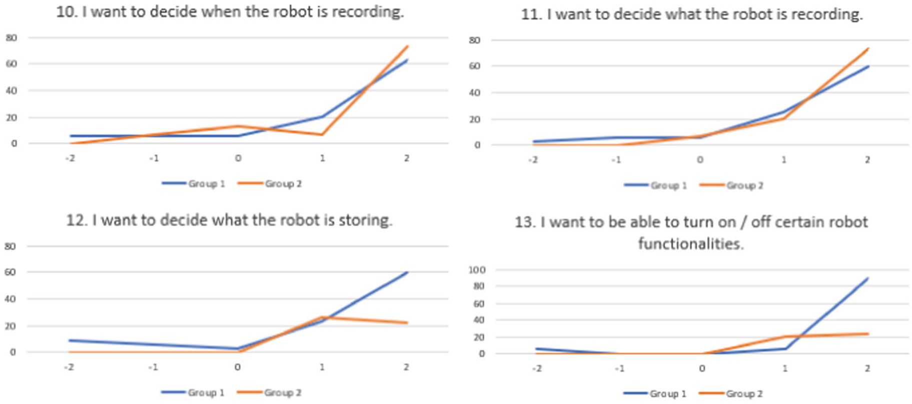


Fig. 7. Intervenability for collected and processed personal information

6.6 Features and Disclosed Information

In this section the five most requested features and their realization and implementation will be discussed. Therefore three gradations of every feature and the personal information, which needs to be disclosed to make use of it, will be proposed. Figure 9 shows the percentage of users who are willing to disclose their personal information which is necessary for the single stages of the features.

Vacuum Cleaning/Autonomous Moving. This category will propose two of the features. Because vacuum cleaning requires autonomous movement, these two features are merged. As already known from Roomba and other vacuum cleaning robots, there are different possibilities of realizing this feature:

1. *Easy cleaning:* The robot drives around in the room or apartment and when it thinks it has finished cleaning, it stops cleaning the floor.
2. *Smart cleaning:* The robot creates a map of the room or apartment and drives through the room in an intelligent way, controlled by an algorithm.
3. *Supervised cleaning:* The robot creates a map, cleans in an intelligent way and additionally, the owner gets information about cleaning status, where the robot already drove and where it did not get.

The general feature of cleaning is requested from 91% of *group 1* and 94% of *group 2*. For the first level of the service no information is needed, but the user cannot check, where the robot has been and where it did not get. For the second level, any kind of sensor is needed, e.g. a laser range scanner or a camera to record home data to create a map. It can also be controlled via Alexa Echo or Google Home. Smart home communication as extra service is allowed by 51% of *group 1* and 21% of *group 2*. If additionally, the robot should be able to transmit personal information to the smartphone, internet and something as a login for the application is needed. Only 12% of *group 1* want the robot to clean the floors, interface with other technologies and would give share their login information. In *group 2*, out of 91%, 5% would use the third level.

Connected Devices. In a smart home, devices are able to communicate among one another. Possible levels in this category are:

1. *Easy communication:* A hub is able to create a local network through which devices can communicate and process their data on a local storage device.
2. *Smart communication:* A cloud based approach is able to handle a lot of more and more complex information in a faster way than a local processing solution.
3. *Location-based communication:* A location based system is able to track activities of the owner and detects if the owner is near the home, the light or music turns on or the heating or air conditioning is starting.

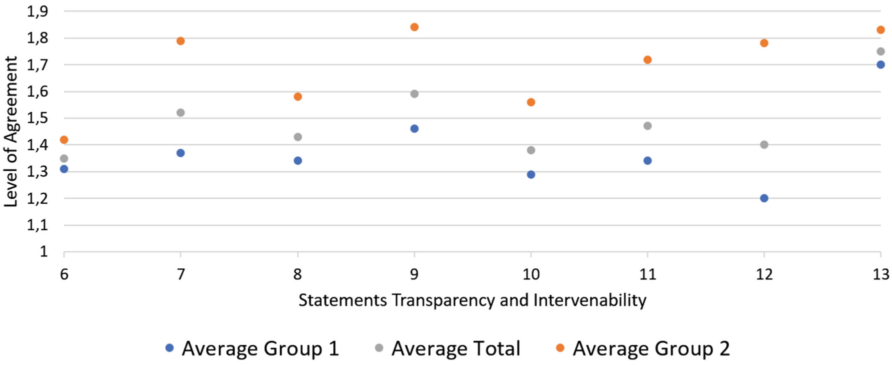


Fig. 8. Avergade level of agreement for transparency and intervenability

In a first step, devices are able to communicate inside the home, information is processed on local storage and no personal information is transfered outside the home (74% of *group 1*, 39% of *group 2*). In a second step, being connected to a cloud means being connected to the internet. For using Alexa Echo and Google Home, speech recognition is necessary, which will be sent to a cloud service and analyzed there. This provides a proficient response for all requests in a fast way. Additionally, both services need full access to accounts. 11% of *group 1* and none of *group 2* would be willing to use those devices with this configuration. Thirdly, smart home devices are allowed to track a person, e.g. to adjust the home conditions according to the time of arrival. Excluding full account access, this is allowed by 14% of *group 1*, 0% of *group 2*.

Surveillance in Case of Emergency. Surveillance is a problematic topic. It is categorized in the following steps:

1. *Easy surveillance:* Data can be stored on internal memory and users are able to watch the videostream connecting to the ip address in the same network or it can be viewed later.
2. *Smart surveillance:* Cameras are connected to the internet, process the collected data and give notification in case of detected motions.
3. *Autonomous surveillance:* The system calls for rescue in case of emergency.

The easiest way of surveillance is to be able to take a look at the video stream and see what happens. No special information is needed for that solution. The second variant gives the possibility to analyze the video stream and in case of changes or movements it would notify e.g. the user. 22% of *group 1* and 5% of *group 2* would still use this, if devices are connected for communication. In the third stage, an algorithm decides what is shown on the video stream, e.g. emergency or burglary and calls for help. Initially, informed persons can be the user, but also family members, the police or the ambulance might get informed, as already introduced in a similar way with eCall for car crashes. In this case

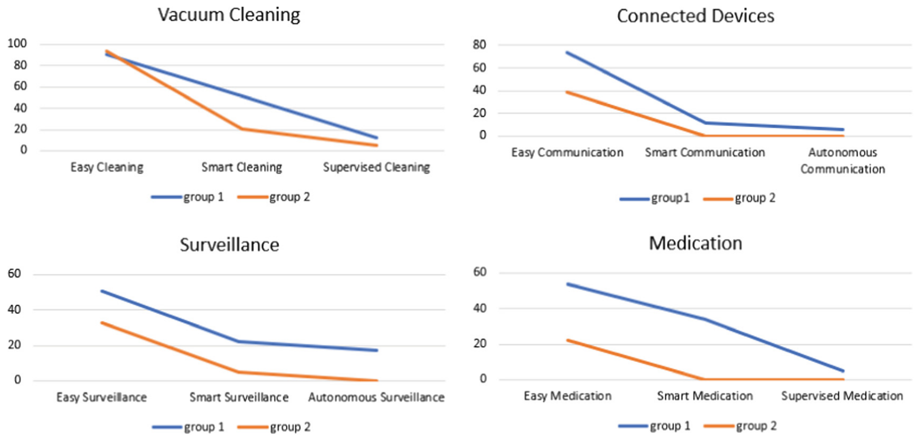


Fig. 9. Gradiations of the features and the percentage of people might using it.

at least emergency numbers needs to be collected. 17% of *group 1* and no one of *group 2* would allow this. (Face recognition is not taken into account in this scenario.)

Medication. Next to surveillance, medication is one of the most critical and sensible features. Distinguished can be in this way:

- *Easy medication:* In regular intervals you are reminded to take your medication. This requires, that the user of this application knows which medicine needs to be taken and sets a timer.
- *Smart medication:* If medication needs to be given dependent on the vital status, e.g. blood pressure, personal information needs to be gathered to ensure a right medication.
- *Supervised medication:* One step further is the involvement of an eHealth application where everything according to your health status is stored.

The first feature requires a timer function only. For every medicine a timer with a special name is set, according to intake intervals. 54,3% of *group 1* would use it, only 22,2% of *group 2*. From those who have choosen this feature, all of *group 1* would also allow the information about medication intake, none of *group 2* would. If only the user is allowed to modify these timers, this feature is uncritical. In a second step, health care status needs to be tracked. Medication often depends on vital functions as blood pressure or blood sugar level. By giving information about medication intake and vital functions, in *group 1* still 34% can use this feature and none of *group 2* would use it, even though 30% of *group 2* would share their vital functions but they did not choose the feature. In a final step, the application is linked to an eHealth application where everything is collected. Therefore it might be necessary to share login information. Only 5,7%

of *group 1* would use this feature. Additionally, in case of anomalies, family members or doctors receive a call, when telephone numbers are allowed. Only 5% of *group 1* and noone of *group 2* would allow this.

7 Analysis

Group 1 is more open-minded to use robots. They selected a wider range of features and are more willing to disclose personal information than *group 2*. Participants of *group 1* have potentially the competence to consider that a larger amount of data collection leads to a larger feature set for the robot. Additionally, in case of privacy risks or problems they are able, to intervene on their own.

Group 2 is more cautious. The chosen features are already available, taking a look at Amazon Echo, Google Home, Roomba or other smart home technologies. Although they are using smartphones every day, they cannot assess the risks and therefore only select already known, available features. Interestingly, compared to smartphones, the conservative disclosure of information of *group 2* is surprising. Personal information is shared with applications on the smartphones to gain the full functional possibilities without thinking about privacy risks [9]. Especially in healthcare, people a willing to share their data if that supports a healthy lifestyle [8,13]. But taking a look at the results of the survey, *group 2* is very reserved. This might lead to the fact, that robots are not in common use for the majority and if they work with robots it is mostly for industrial purposes. Though it needs to be mentioned again, that attitude differs from behavior.

One interesting fact is the distinction between persons. Although 60% of *group 2* wants to have this feature, only 10,5% would allow the robot to use speech recognition and no one would allow face recognition. In *group 1*, 40% of participants would allow both of it.

8 Discussion and Future Work

The questionnaire gave first quantitative hints on user preferences and attitude towards robots. As it can be seen in the results, it almost doesn't matter if people have a technical background. Most of the participants are interested in having a transparent view on the data processed by the robot and they would like to have to possibility to intervene. Even though this survey is not as representative as a user study or a workshop, it shows that users need to be involved during the development process.

Therefore, as a next step users need to be asked and observed directly. They need to be integrated in the development process to incorporate their perspective. Thereupon, with the help of participatory design strategies, solution ideas need to be developed. As presented, desired features can be implemented on different levels. As one example, distinction between people normally is realized using face or speech recognition. If a user does not want to allow the usage of a camera, there might be less-invasive possibilities to implement the feature, e.g.

using RFID tags, specific code words or color identifications. Especially interdisciplinary teams would be able to give new perspectives in robot design and development. Figure 10 shows the overall development process of for privacy protecting robots using a privacy by design approach.

This allows the user, to get a holistic view of certain features and its capabilities. Users have to be able to be aware of features and its accompanying privacy risks simultaneously. They need to understand and consider actively what happens and how sensors, features and personal information are related to each other. As an example, a red light is blinking when a camera is recording but we only see the content of the video watching it. Installing smartphone applications requires access authorization for camera, microphone and gps sensor - but the purpose of the sensor use is not clearly visible. The questionnaire shows that non-experts does not want to share personal information. On the one hand, this needs to be respected in creating privacy-friendly solutions. On the other hand, sometimes personal information is required but users should be able to decide on its own allowing the access or not. And if the user allows the access, the purpose needs to be clear and the corresponding personal data must be kept confidential.

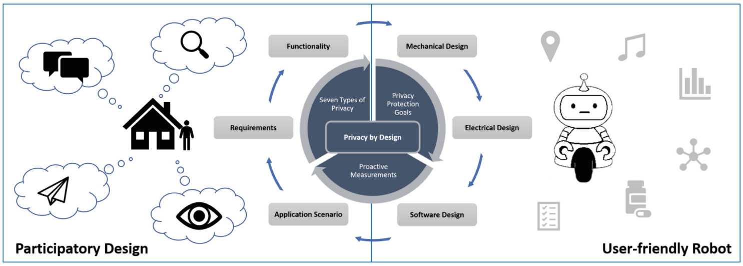


Fig. 10. Development process

This survey makes clear, that users should be made aware of common and possible privacy risks because of the imbalance of privacy attitude and sharing behavior [10]. Although both groups are stating their interest about collected data, they do not want to share personal information like location and logins, and *group 2* is critical towards speech and face recognition, there is a gap between attitude and behaviour. As an example, depending on the purpose, users are willing to share their personal data to be able to use different kinds of applications [8,9,13]. With the prototype they are able to take a look at certain features and it should be made clear, what sensors are used to provide this feature and which personal information needs to be collected. With the addition or removal of personal information or processed data, users can get an idea of how the results or operating principles of features change and if they still matche the requirements or if users need to allow more collection of personal data. The more complex a feature gets, the more difficult it is to get a holistic view. An important aspect for introduction of robots is to make people aware of which personal information is needed for specific tasks and how they are protected.

Soon, robots will enter our private homes and going to be part of a smart home. With autonomous movements and decision-making of a robot, it is becoming a very complex and inherent part of our home and our life. Therefore, we first of all need to make people aware of privacy risks and how they can protect themselves. Secondly, there needs to be a possibility to use robots in a privacy-friendly way such that users can decide about functionalities and using a feature on different levels depending on the privacy perception of the user.

Acknowledgment. This work was supported by the Ministry for Science and Culture of Lower Saxony as part of the program “Gendered Configurations of Humans and Machines (KoMMa.G)”.

References

1. irobot store - roomba980. http://store.irobot.com/default/roomba-vacuuming-robot-vacuum-irobot-roomba-980/R980020.html
2. Alaiad, A., Zhou, L.: The determinants of home healthcare robots adoption: an empirical investigation. Int. J. Med. Inf. **83**(11), 825–840 (2014)
3. Alaiad, A., Zhou, L., Koru, G.: An empirical study of home healthcare robots adoption using the UTUAT model (2013)
4. Aroyo, A.M., Rea, F., Sandini, G., Sciutti, A.: Trust and social engineering in human robot interaction: will a robot make you disclose sensitive information, conform to its recommendations or gamble? IEEE Robot. Autom. Lett. **3766**(c), 1–8 (2018)
5. Bugeja, J., Jacobsson, A., Davidsson, P.: On privacy and security challenges in smart connected homes (2016)
6. Caine, K., Šabanovic, S., Carter, M.: The effect of monitoring by cameras and robots on the privacy enhancing behaviors of older adults. In: Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction, pp. 343–350. ACM (2012)
7. Caine, K.E., Fisk, A.D., Rogers, W.A.: Benefits and privacy concerns of a home equipped with a visual sensing system: a perspective from older adults. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 50, pp. 180–184. Sage Publications Sage, Los Angeles (2006)
8. Chen, J., Bauman, A., Allman-farinelli, M.: A study to determine the most popular lifestyle smartphone applications and willingness of the public to share their personal data for health research. Telemed. e-Health **22**(8), 655–665 (2016)
9. Chin, E., Felt, A.P., Sekar, V., Wagner, D.: Measuring user confidence in smartphone security and privacy, p. 1 (2012)
10. Coopamootoo, K.P., Groß, T.: Why privacy is all but forgotten. Proc. Priv. Enhancing Technol. **2017**(4), 97–118 (2017)
11. Denning, T., Matuszek, C., Koscher, K., Smith, J.R., Kohno, T.: A spotlight on security and privacy risks with future household robots: attacks and lessons. In: Proceedings of the 11th International Conference on Ubiquitous Computing, pp. 105–114. ACM (2009)
12. Fallmann, S., Psychoula, I., Chen, L., Chen, F., Doyle, J., Triboan, D.: Reality and perception: activity monitoring and data collection within a real-world smart home (2017)

13. Felt, A.P., Egelman, S., Wagner, D.: I've got 99 problems, but vibration ain't one: a survey of smartphone users' concerns. In: Proceedings of the Second ACM Workshop on Security and Privacy in Smartphones and Mobile Devices, pp. 33–44. ACM (2012)
14. Geneiatakis, D., Kounelis, I., Neisse, R., Nai-fovino, I., Steri, G., Baldini, G.: Security and privacy issues for an IoT based smart home, pp. 1292–1297 (2017)
15. Heuer, T., Schiering, I., Gerndt, R.: Privacy and socially assistive robots - a meta study. In: Hansen, M., Kosta, E., Nai-Fovino, I., Fischer-Hübner, S. (eds.) Privacy and Identity 2017. IAICT, vol. 526, pp. 265–281. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-92925-5_18
16. Lee, H.R., Tan, H., Šabanović, S.: That robot is not for me: addressing stereotypes of aging in assistive robot design. In: 2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pp. 312–317. IEEE (2016)
17. Lera, F.J.R., Llamas, C.F., Guerrero, Á.M., Olivera, V.M.: Cybersecurity of robotics and autonomous systems: privacy and safety. In: Robotics-Legal, Ethical and Socioeconomic Impacts. InTech (2017)
18. Lin, H., Bergmann, N.W.: IoT privacy and security challenges for smart home environments (2016)
19. Pino, M., Boulay, M., Jouen, F., Rigaud, A.S.: Are we ready for robots that care for us? Attitudes and opinions of older adults toward socially assistive robots. Frontiers Aging Neurosci. **7**, 141 (2015)
20. Smarr, C.A., Prakash, A., Beer, J.M., Mitzner, T.L., Kemp, C.C., Rogers, W.A.: Older adults preferences for and acceptance of robot assistance for everyday living tasks. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 56, pp. 153–157. SAGE Publications Sage, Los Angeles (2012)
21. Syrdal, D.S., Walters, M.L., Otero, N., Koay, K.L., Dautenhahn, K.: He knows when you are sleeping-privacy and the personal robot companion. In: Proceedings of Workshop Human Implications of Human-Robot Interaction, Association for the Advancement of Artificial Intelligence (AAAI 2007), pp. 28–33 (2007)
22. Theoharidou, M., Tsalis, N., Gritzalis, D.: Smart Home Solutions for Healthcare: Privacy in Ubiquitous Computing Infrastructures (2011)
23. Wu, Y.H., Cristancho-Lacroix, V., Fassert, C., Faucounau, V., de Rotrou, J., Rigaud, A.S.: The attitudes and perceptions of older adults with mild cognitive impairment toward an assistive robot. J. Appl. Gerontol. **35**(1), 3–17 (2016)
24. Ziefle, M., Rocker, C., Holzinger, A.: Medical technology in smart homes: exploring the user's perspective on privacy, intimacy and trust. In: 2011 35th Annual IEEE Computer Software and Applications Conference Workshops (COMPSACW), pp. 410–415. IEEE (2011)

chownIoT: Enhancing IoT Privacy by Automated Handling of Ownership Change

Md Sakib Nizam Khan[1](✉), Samuel Marchal[2], Sonja Buchegger[1], and N. Asokan[2]

[1] KTH Royal Institute of Technology, Stockholm, Sweden
{msnkhan,buc}@kth.se
[2] Aalto University, Espoo, Finland
samuel.marchal@aalto.fi, asokan@acm.org

Abstract. Considering the increasing deployment of smart home IoT devices, their ownership is likely to change during their life-cycle. IoT devices, especially those used in smart home environments, contain privacy-sensitive user data, and any ownership change of such devices can result in privacy leaks. The problem arises when users are either not aware of the need to reset/reformat the device to remove any personal data, or not trained in doing it correctly as it can be unclear what data is kept where. In addition, if the ownership change is due to theft or loss, then there is no opportunity to reset. Although there has been a lot of research on security and privacy of IoT and smart home devices, to the best of our knowledge, there is no prior work specifically on automatically securing ownership changes. We present a system called chownIoT for securely handling ownership change of IoT devices. chownIoT combines authentication (of both users and their smartphone), profile management, data protection by encryption, and automatic inference of ownership change. For the latter, we use a simple technique that leverages the context of a device. Finally, as a proof of concept, we develop a prototype that implements chownIoT inferring ownership change from changes in the WiFi SSID. The performance evaluation of the prototype shows that chownIoT has minimal overhead and is compatible with the dominant IoT boards on the market.

Keywords: Ownership · Privacy · Smart home · IoT

1 Introduction

Internet of Things (IoT) devices produce and store sensitive information related to their sensing capabilities and contextual awareness. Similarly, they contain information related to configuration settings, credentials for network and user authentication, etc., all of which are privacy sensitive. Security has been one of the major concerns of the IoT paradigm due to a combination of factors, such

Published by Springer Nature Switzerland AG 2019
E. Kosta et al. (Eds.): Privacy and Identity 2018, IFIP AICT 547, pp. 205–221, 2019.
https://doi.org/10.1007/978-3-030-16744-8_14

as a potentially large number of networked devices, unprecedented use cases, resource constraints, and often sensors or other collections of data about user behavior, adding new privacy concerns.

Ownership here refers to the ability to control, manage, and access a particular device. The large growth in deployment of smart home IoT devices has introduced the possibility of device ownership change (due to selling, loss, theft, moving house, lending - handing over to use elsewhere or to a guest at the same location). This change can compromise data or access rights for both the previous and the new owner. For instance, a user forgets to log out from his smart TV box before selling it. In such a scenario, the available data or credentials can easily be misused by the buyer/new owner, ranging from browsing the history of what the previous owner watched to charging downloads to the associated credit card. Therefore, handling ownership change in a secure manner becomes necessary.

Contributions. The major contributions of the paper are the following:

– We present the problem of automatic handling of ownership change of IoT devices, with adversary model and requirements (Sect. 3).
– chownIoT, the first system capable of protecting owner privacy without user interaction during ownership change of IoT devices (Sect. 4). We implemented a prototype on Raspberry Pi and evaluated its performance (Sect. 5). chownIoT has the following features:
  - Automatic detection of ownership change using the context of IoT devices. For the prototype, this is based on the WiFi SSID.
  - A profile management system for authenticating owners.
  - Encryption of data and isolation of owner profiles for owner privacy.
  - A communication protocol between the IoT device and the smartphone device (i.e. used for controlling the IoT device) enabling its implementation.
– We discuss extensions to chownIoT in two directions, more sophisticated context 6 and vendor independence 7.

2 Related Work

Recently, several research works have been done to secure and ensure smooth operation of IoT devices. Our proposed solution is closely related to smart-home device privacy, authentication and access control of IoT devices (especially for ownership change), and context-aware security. We thus divide the related works into these three major categories.

Smart Home Device Privacy. Recently, studies are focusing on mitigating the privacy issues of smart-home devices. Apthorpe et al. [2], examined different smart home IoT devices and found that even with encrypted traffic, the network-traffic rates of the devices can reveal potentially sensitive user interactions. In another work [1], the same authors proposed mechanisms for preventing network observers from inferring consumers' private in-home behaviors. While the general concern of privacy matches ours, the main difference to our proposed work is that these works only focus on privacy issues of smart-home devices related to passive network observation. The privacy issues related to *ownership* of smart-home IoT devices, the main focus of our work, are not addressed.

Authentication and Access Control of IoT Devices, Ownership Change. Due to their often limited computational capabilities, IoT devices require light-weight yet secure authentication and access control mechanisms. Several research works talked about authentication requirements during ownership change of IoT devices. Tam et al. [21] and Bohn [3] proposed ownership transfer mechanisms for smart devices in their individual works for securely transferring ownership. Similarly, Pradeep et al. [17] also proposed a concept of ownership-authentication transfer for securely handling the ownership transfer of a device to a new owner. The main difference to our proposed system is that in ownership-authentication transfer the seller has to initiate the transfer process, there is no notion of detecting ownership change *automatically*. In addition, the protocol requires a central key server for key management which may not be feasible in the case of smart homes. The protocol does not also mention any data protection mechanism.

Context Aware Security. In recent years, several security solutions [7,24] have started using context to provide better security. Several works have integrated context for providing better automatic access-control techniques [8,9,15,18,25]. Besides access control, Miettinen et al. [14] proposed a new approach for secure zero-interaction pairing intended for IoT and wearable devices, which uses context to identify the pairing devices. Apart from proposing new context-aware security solutions, some studies also focused on finding vulnerabilities in already existing solutions [20]. All of these works leverage context either to take access control, pairing or key agreement decisions to improve security. In our work, we leverage context of a smart-home device for a new purpose: *detecting ownership change.*

3 Models and Requirements

3.1 System and Adversary Model

In our system model, IoT devices are connected to the owner's account with a cloud service through a network connection mediated by an access point. The owner also has a control device (typically her smartphone) to interact with the IoT device directly. Ownership change in our context refers to the IoT device only.

Ownership change of a device involves two parties, the previous and the new owner, that need to be protected from each other, as they are both potential adversaries and targets. We model the adversary as malicious, i.e., assume that they can mount active attacks on security and privacy, with standard assumptions on computational power. We conservatively assume that access to one asset (cloud account, IoT device), unless specifically prevented, implies access to the other. The adversary is interested in access to the other party's data, including credentials and metadata. We now list our specific attacker types and capabilities.

Previous Owner Adversary (POA): access to (and credentials for) the cloud account and the control device associated with the IoT device.
New Owner Adversary (NOA): access to the IoT device.
Advanced New Owner Adversary (ANOA): NOA plus special equipment and dedication to read out from IoT device storage while the device is turned off as well as ability to spoof the AP the device was associated with. To spoof an AP with the correct SSID and MAC, the attacker needs to 1. know these, 2. know the protocol and other authentication parameters used, and 3. participate in the protocol while accepting any credential.

3.2 Requirements

Smart home devices have some special characteristics, such as limited resources, different sensor modalities and application dependencies, which differentiate them from traditional devices. Based on these characteristics, the requirements of an intended solution are:

1. Resource Constraints: Ability to work on resource constrained devices in terms of computation (and thus energy), network, storage, and memory.
2. Security Goal: equivalent to timely reset to factory defaults of the IoT device upon ownership change in terms of confidentiality (privacy), integrity, and availability. Specifically, protect the data on the cloud, the control device, and the IoT device from the respective other owner.
3. Deployability: Adaptable to the largest possible class of devices.
4. Usability: The added functionality may not be outweighed by any burden put on the user. That means minimal user involvement and waiting time as well as minimal consequences for any wrong decisions made automatically or exceptions such as loss of the control device.

4 chownIoT

IoT covers a wide range of heterogeneous devices and diverse scenarios. We therefore first present the algorithmic view of our solution that can be adapted to different environments and protocols. We then explain one concrete instantiation and reasoning for design choices.

4.1 Algorithmic Solution Overview

To protect privacy-sensitive data against adversaries (previous or new owner), chownIoT 1. automatically detects change of ownership, 2. manages owners by maintaining individual profiles for each owner, and 3. verifies the ownership change and protects data based on owner authentication.

The smart home device maintains a profile for each owner, which enables the isolation of one user profile from another. Each *profile* contains owner authentication credentials, cloud credentials, known contexts and user data that is specific to the owner. All data of a particular owner is managed under a profile. The main goal of chownIoT is to protect this data on the device (and any associated data stored in the cloud).

Figure 1 illustrates the flow diagram of chownIoT. The process starts with inferring ownership change based on the change in context of the device. Most smart home devices are either static or semi-static in terms of their mobility. The static devices never move once deployed except if sold (e.g. smart AC, smart fridge) whereas the semi-static devices move rarely after deployment (e.g. baby monitoring camera, smart TV) within a fixed boundary. The deployment context for such devices usually never changes while they are under the same owner. However, when their ownership changes, the deployment context also changes. Therefore, by identifying the change in the deployment context of a device we can infer ownership change. Thus, if chownIoT detects a context change, it tries to authenticate the owner based on her control device (smartphone) or her own credentials. If the authentication is successful, chownIoT infers that only the context of the device has changed (e.g., the owner has moved the device from the home to the summer cottage) but not the ownership. Thus it creates a new known context for the same owner in the same profile.

In contrast, if the authentication is not successful, chownIoT concludes that the ownership has changed, hence it protects the profile data, using encryption. In chownIoT, only one profile remains active at a time and all others are protected. Once the profile data is protected, the owner can either retrieve an existing profile or create a new profile. If the owner chooses to retrieve an existing profile, chownIoT authenticates the owner for the selected profile. After successful authentication, it releases the profile data and stores the context as a new known context for the selected profile. However, if the authentication is not successful or the user chooses to create a new profile, chownIoT creates a new profile for the owner. With the new profile, the owner gets full control of the device except access to the data of other profiles which remain protected.

This algorithm is executed once the device has been deployed. The life-cycle of an IoT device, however, begins with configuring the device into the deployment network. Currently, most available smart home IoT devices require a smartphone and a vendor provided smartphone application for the initial configuration as well as for later management/control [16]. chownIoT requires some additional steps besides traditional configuration.

chownIoT builds the security mechanism by trusting the control device used during initial configuration of a smart home device. *Control device* here refers to the smartphone used to configure, control and manage the smart home device. During the initial configuration it establishes a security association (rendering the control device a *trusted device*) as well as an owner authentication mechanism

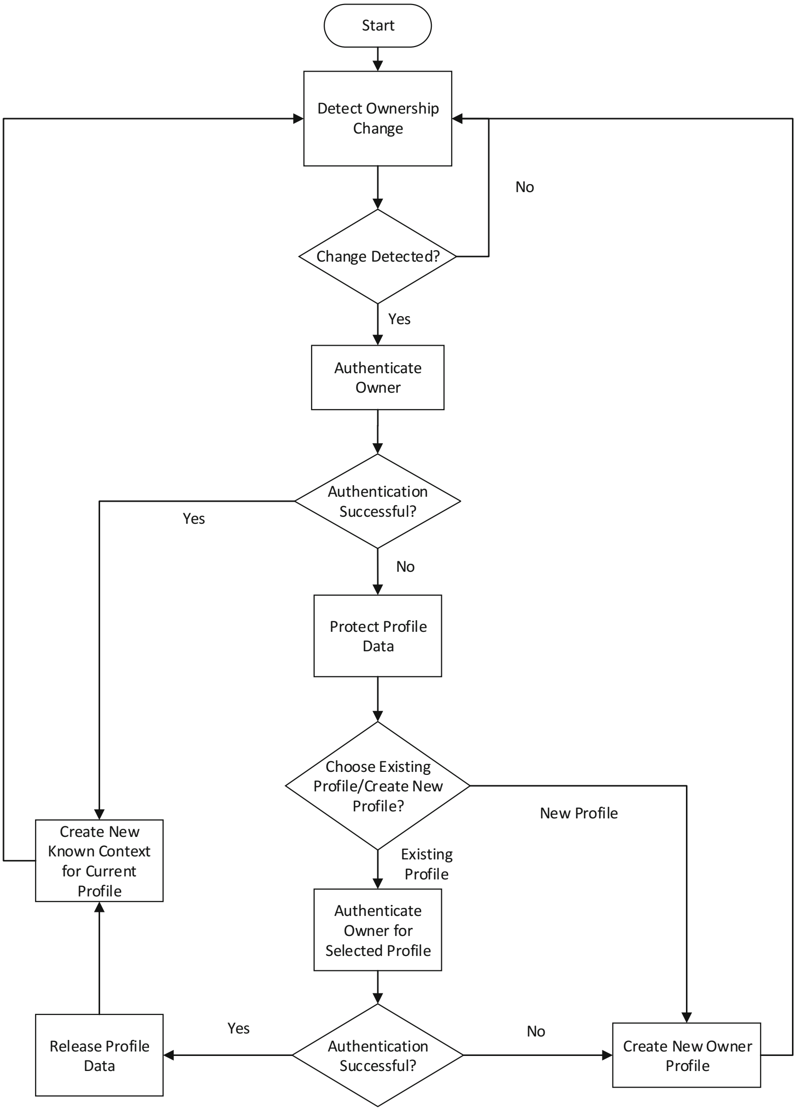

Start
Detect Ownership Change
Change Detected?
No
Yes
Authenticate Owner
Authentication Successful?
Yes
No
Protect Profile Data
Choose Existing Profile/Create New Profile?
New Profile
Existing Profile
Create New Known Context for Current Profile
Authenticate Owner for Selected Profile
Release Profile Data
Yes
Authentication Successful?
No
Create New Owner Profile

Fig. 1. Flow diagram of chownIoT

for future verification of the owner in case of ownership change. The interactions between the smart home device and the trusted device during initial configuration are depicted in Fig. 2.

Apart from the security association, the smart home device also establishes an owner authentication mechanism. With the security association, the authentication mechanism is bound to a particular control device. But realistically, the owner should be able to authenticate with any device.

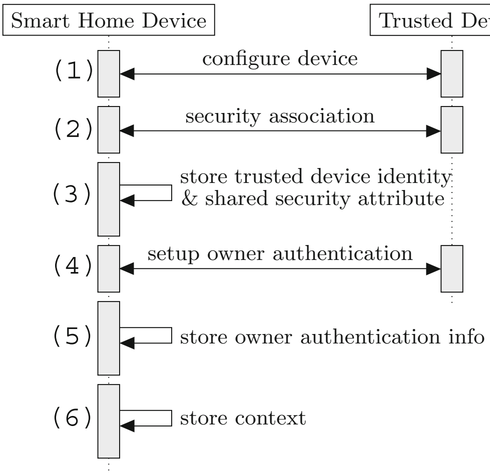

Smart Home Device
Trusted Device
(1)
configure device
(2)
security association
(3)
store trusted device identity
& shared security attribute
(4)
setup owner authentication
(5)
store owner authentication info
(6)
store context

Fig. 2. Sequence diagram of chownIoT during initial configuration

4.2 Prototype Design and Implementation Choices

To realize chownIoT, we made some design and implementation choices. In this section we discuss and reason about our choices.

Initial Configuration. First, the control device configures the smart home device (1) in Fig. 2. The configuration steps includes device specific configuration and some additional steps for chownIoT, namely, turning on the discoverable mode of Bluetooth and creating a server socket which listens for packets. To facilitate the communication between the smart home device and the control device/trusted device we define a simple protocol based on User Datagram Protocol (UDP) (for details see [10]). The *configure device* feature is implemented using Bluetooth pairing. The control device discovers the smart home device and sends a pairing request. The smart home device responds to the pairing request and once paired, the control device performs the necessary configuration.

For the next step, establishing a security association (2) that serves to check whether the context change likely implies an ownership change, there are different techniques available. For instance, the involved parties can have each other's public key and their own private key. Key agreement is another alternative where the involved parties establish a shared secret key between them. The public/private-key mechanism requires larger key sizes than symmetric key to achieve similar security [12]. In addition, they also require more computational processing power [11]. As most smart home IoT devices are resource constrained, we establish a shared secret between the smart home device and the trusted device using Diffie-Hellman key exchange protocol [19]. After establishing a shared secret, chownIoT stores the identity (3) of the trusted device for future verification of ownership change. The identity includes the Bluetooth device name and MAC address of the trusted device and the established shared secret as depicted in Table 1. The trusted device identity is stored in persistent storage.

Table 1. Elements stored during initial configuration

Trusted device identity	Bluetooth device name	Bluetooth MAC address	Shared secret
Known context	AP SSID	AP MAC address	AP Access credential
Owner profile	Known context	Trusted device identity	Profile name

In addition to the security association with the trusted device, chownIoT implements an authentication mechanism (4) for the user independent of the control device she is using. There are several candidates for authentication mechanisms, such as password-based authentication, public key based authentication protocol, Authentication and Key Agreement protocol (AKA), and Extensible Authentication Protocol (EAP). Password-based authentication mechanisms are widely used as they are convenient to use and implement [26]. Moreover, using password-based authentication, an owner can be authenticated on any control device very easily. This is due to the fact that by using password-based authentication, we do not need to store any key on the trusted device. Thus, we chose to implement a password-based authentication mechanism. The owner is prompted to choose a profile name and a password during the initial configuration. The smart home device receives the hash of the owner password and a profile name from the trusted device which it stores (5) in the persistent storage. In addition, it stores the triplet <SSID, MAC address of the Access Point (AP), access credential of the AP> as known context (6) in the persistent storage. A *known context* here refers to a context that has been already observed and approved by a particular owner for a particular device. Finally, chownIoT generates a profile which is identified by the profile name provided by the user, and the known contexts and trusted device identity are stored under the profile. The stored elements for both context and owner profile are as depicted in Table 1.

Handling Ownership Change. After the initial configuration, the smart home device starts detecting possible ownership change based on change in context. We chose the Wi-Fi SSID as a simple indicator of potential ownership change. IoT devices mostly achieve Internet connectivity through a Wi-Fi connection with an AP [16]. Devices know the SSID of the wireless network that they are connected to. In a typical smart home scenario, the SSID is the same for the whole house or apartment. Thus, for static and semi-static devices, the connected SSID is unlikely to change while a particular device has the same owner. However, a new owner needs to connect the IoT device to a different network, which also changes the SSID of the particular device.

In chownIoT, the context of a smart home IoT device is linked to the SSID it is currently connected to. The smart home device continuously monitors the SSID of the AP that it is currently connected to and compares it with the stored known context list. If the current SSID is not in the list, it infers a possible ownership change. Once it detects a possible ownership change, at first it triggers a Bluetooth discovery looking for the trusted device based on the identity it stored during initial configuration. If the trusted device is discovered, the smart home device tries to authenticate it using the shared secret established during the initial configuration. The authentication is performed using a challenge-response authentication mechanism. The details of the challenge-response mechanism can be found in [10]. If the trusted device is not available, then password-based authentication is triggered. Although SSID has clear limitations, as it can easily be forged, it has the advantage of being present at every smart home, is independent of which or how many IoT devices there are, and is easy to detect. However, depending on the setup, other context information can be used and easily be integrated into chownIoT.

Profile Management and Data Protection. During ownership change when the authentication process fails, chownIoT protects the profile data by means of encryption. For data encryption, we use AES-CCM [13] authenticated encryption technique. It is a technique that provides both authentication and encryption at the same time. We use a key derived from the owner-provided password as the encryption key for AES-CCM, with a key length of 128 bits as recommended by NIST [6]. We do not store the encryption key on the device as physical memory access can leak the encryption key. In addition, we also require a profile retrieval mechanism using any control device. Thus, we cannot store the key on the trusted devices only. To fulfill these requirements, we chose to derive the encryption key from the owner password. The main benefit of password-based key derivation is that it can be instantly derived from the owner-provided password without requiring to be stored and it also facilitates profile retrieval using any control device. Password-Based Key Derivation Function 2 (PBKDF2) with 256 bits random salt and 4096 iterations of SHA256 hash algorithm is used to derive the key from the password. The key derivation function is given in Equation (1).

$$Key = H_{\mathrm{SHA256}}(Password_{\mathrm{SHA256}}, Salt_{\mathrm{256bit}}, 4096_{\mathrm{iterations}}) \tag{1}$$

As the smart home device receives the hash of the owner password from the trusted device, it derives the key according to Eq. (1). The salt for the key derivation is randomly generated. Once the key derivation is completed, the smart home device only stores the salt and the key in persistent storage, and deletes the password hash. During an ownership change, the smart home device encrypts the profile data except for the salt and the profile name. Once the encryption process is completed, it also deletes the derived key. There are two cases when the profile needs to be encrypted: either the inferred ownership change was a false positive, or it was only temporary. When the owner wants to retrieve the encrypted profile, she is prompted to provide the owner password. Upon receiving the password hash, the smart home device again derives the key using the provided password hash and stored salt value. Once the key is derived, it performs authenticated decryption using AES-CCM and the derived key. If the process is successful, the owner is authenticated and the profile data gets decrypted. chownIoT limits the (configurable) number of failed user authentication attempts, after which the whole profile gets deleted.

5 Evaluation and Discussion

For evaluating chownIoT, we implemented the smart home device features on a Raspberry Pi 3 using C++ and control device features on the Android platform. We evaluate chownIoT based on the requirements identified in Sect. 3.2.

5.1 Resource Constraints

For **CPU usage**, we measure the performance of the major resource-intensive operations in chownIoT, namely encryption/decryption[1], key derivation, and hashing.

Table 2. Encryption CPU usage

Data size	CPU usage (seconds)
10 KB	0.010
10 MB	5.89
100 MB	60.67

The data to be encrypted includes chownIoT protocol data and user data produced by the specific device. Table 2 lists the CPU usage measured in seconds for different data sizes. 10 KB and 10 MB data requires only 0.010 s and 5.89 s, respectively, for encryption which is not a large computational overhead. By experimenting with a range of IoT devices (e.g. weather stations, smart switches, and different kinds of sensors), we found that most of them produce data between 10 KB and 100 MB. Thus, encrypting data on such devices is quite feasible. Apart from this, we can also see that 100 MB data requires 60.67 s. Devices such as surveillance cameras produce this amount of data and such devices do not need to be always active. Thus, spending 60 CPU seconds for encrypting data seems feasible for such devices. For devices that produce larger amounts of data, the corresponding encryption time can be very long. While by default

[1] We only measure encryption, as it yields more conservative results than decryption [4].

chownIoT encrypts the entirety of the data, depending on the device type not all data are necessarily privacy sensitive. For such cases, it may make sense to allow for device-specific adaptations of the system to identify and protect only more privacy-sensitive data. This opens the possibility of different choices in the trade-off between privacy and performance/usability.

The time required for key derivation and hashing are 0.060 and 0.010 s, respectively.

In the current implementation of chownIoT, the smart-home device continuously loops to detect change of SSID for detecting ownership change. This is also resource intensive, as the monitoring process needs to run all time. This can be improved by implementing call backs for when there is any disconnection or state change of the Wi-Fi connection.

Table 3 lists the CPU specifications of three of the most popular IoT boards. Arduino Tian and Intel Edison have less powerful CPUs than Raspberry Pi 3. The resource-hungry operations may need twice the CPU time of Raspberry Pi 3 to execute on these boards. Thus, operations such as encryption or decryption of large amounts data can degrade the usability of the system on these boards.

Table 3. Specifications of IoT boards

Board	CPU	RAM
Raspberry Pi 3	ARM Cortex-A53, 1.2 GHz	1 GB
Arduino Tian	Atheros AR9342 560 MHz	64 MB
Intel Edison	Dual-Core Intel Atom 500 MHz	1 GB

We measured the **RAM usage** of each resource-intensive operation, see Table 4. It is constant (approximately 1000 KB or 1 MB) regardless of data size and operation performed and only uses a small fraction of the RAM available on the boards in Table 3.

Table 4. Memory usage for different operations

Operation	RAM usage (KB)
Encryption 10 KB	1031
Encryption 10 MB	1051
Encryption 100 MB	1094
Key derivation	1047
Hashing	1045

The **network overhead** is negligible as all operations, except the initial configuration and authentication, are performed only on the smart-home device itself and do not involve any network communication. In terms of **storage**, chownIoT does not add much overhead either, as encryption using CCM adds minimal message expansion [23].

chownIoT fulfills the requirements from Sect. 3.2 in terms of computation, network, memory, and storage even for quite resource-constrained devices.

5.2 Security and Privacy

To fulfill the requirements of protecting the cloud account privacy and device data privacy as identified in Sect. 3.2, chownIoT isolates owner profiles from one another. This isolation is done by encryption with carefully chosen parameters (see Sect. 4.2) as soon as the absence of the trusted device is detected in case of a suspected ownership change.

The potential entry points for an adversary (NOA, POA, ANOA, as defined in 3.1) to break this isolation are 1. the context used to infer change of ownership (NOA), 2. authentication (both user and device) (NOA, POA), and 3. physical access to the active, unencrypted profile (ANOA).

From an implementation point of view, the known context can be spoofed by replicating the AP SSID and MAC address. They are stored along with access credentials for the AP in the known context information, which is encrypted.

Although chownIoT deletes both the owner password and the derived encryption key on the device once the data gets encrypted, the security of user authentication is limited to that of password authentication in general, meaning vulnerability to guessing and brute-force attacks. Online guesses for the device are limited and thus protect against a NOA that does not know the password.

Offline attacks are, however, possible for the ANOA, meaning the attacker has physical access to the device and the right equipment to read out the ciphertext and salt.

Regarding device authentication, the challenge-response based authentication is in theory vulnerable to relay attacks [5]. In chownIoT, however, the proximity requirements enforced by Bluetooth communication during device authentication (for inference of ownership change or profile reactivation) makes relay attacks unrealistic and ineffective[2].

In case of a device getting stolen and/or powered off before detecting ownership change, the data of the active profile remains unencrypted and can be read by ANOA. Even if the data were encrypted at all times, the ANOA can do offline password cracking.

In summary, chownIoT withstands the NOA and POA, but in the time window between the device changing hands and being powered on, it is vulnerable to the ANOA, a determined attacker that can spoof the access point, perform an offline brute-force attack on the password by reading from the storage of a powered-off device if the current profile has not yet been encrypted.

[2] Nevertheless, they can easily be mitigated by implementing user consent/notification during the authentication process, at the cost of reduced usability.

5.3 Deployability

chownIoT does not depend on any particular operating system or hardware. The implementation of chownIoT only depends on Wi-Fi communications for ownership-change detection and does not involve any other sensors. The ownership change detection technique can also be adapted for other communication technologies, such as Bluetooth and ZigBee, for instance by analyzing/monitoring the available nearby devices of a smart home device. Thus, it is possible to implement chownIoT on any device with a communication interface. There can, however, be extremely resource-constrained devices that cannot run chownIoT due to the requirements discussed in Sect. 5.1. Some devices do not store user data deemed sensitive (or any user data at all) and thus do not need chownIoT. We hypothesize that the overlap between these two types of devices is large (e.g. smart light bulbs).

Vendor Dependency. The current solution of chownIoT requires vendor cooperation for deploying it on existing and also upcoming IoT devices on the market. According to the present implementation, to deploy chownIoT, a vendor needs to include the smart-home device part of the solution in the firmware of the intended device. In addition, the vendor also needs to include the control device part of the solution in the vendor provided control application. For existing devices, deploying chownIoT would require a device firmware as well as an application update. However, vendor dependency is a major limitation of the current system due to the lack of universally adopted standards, and a workaround to overcome this dependency is needed. Our proposal for reducing/eliminating vendor dependency and the ensuing trade-offs are discussed in Sect. 7.

5.4 Usability

In chownIoT, besides the regular configuration, we additionally setup an owner-authentication mechanism and a shared secret with the trusted device. While the generation of the shared secret for the security association with the control device is automatic, the owner authentication mechanism requires user participation. In our prototype implementation that means the user sets up a profile name and password. The ownership-change detection is performed automatically, requiring no user interaction. Loss of the trusted control device requires user authentication and a new security association, for the user that only means entering the password again and enabling Bluetooth.

One potential limitation of chownIoT in terms of usability is the possibility of false positives, i.e., detection of ownership change when none occurred. This happens if the change of context and both the device and user authentication fail and yet the inference of ownership change is invalid. While this should be rare, since the change of context with continued ownership most likely involve the owner and/or her control device, it entails user involvement to fix. Once chownIoT assumes an ownership changes, the profile gets encrypted. To retrieve the profile, the owner needs to authenticate herself, in our implementation that means selecting the profile she wants to access and supplying her password. She then has to wait until chownIoT has decrypted her profile data; the time required depends on the size of the profile data and the processing power of the device.

The reverse problem, false negatives, can happen when the ownership and context changes, but the previous owner and her trusted control device are nearby and thus authentication succeeds (or due to a successful relay attack). While this could be remedied by user notifications and consent for each automatic authentication based on the security association, we opted to not include that and err on the side of usability.

6 Advanced Ownership Change Detection

chownIoT uses SSID as a simple indicator of change in the context of an IoT device. Extending chownIoT to overcome the limitations of SSID, Artur Valiev, in his master's thesis [10] proposed a more robust and forge-proof system called *FoundIoT* that uses richer context. The main idea behind FoundIoT is that if a device stays in the same context, it will observe other devices in its vicinity over time. However, if the device goes to a new context, the devices in the vicinity will also change. Thus in FoundIoT, the context change is inferred by monitoring nearby devices over wireless communication channels. Once the data is captured, FoundIoT performs statistical analysis on the data to detect a change in context, in multiple stages. First, change is detected based on Wireless Stations (STA) which are the other IoT devices in the device vicinity, then APs in vicinity of the device, and finally, Bluetooth-enabled devices. It uses the Jaccard Index and Kullback-Leibler (KL) divergence for finding similarity metrics changes over consecutive scans. If the similarity is low, FoundIoT concludes that there is an ownership change. The author shows that by using such techniques, it is possible to detect ownership change with high accuracy and low false alarms in their system model.

7 iChownIoT

Currently, chownIoT requires vendor cooperation for deployment. One way to eliminate this dependency is to move ownership-change detection to an independent device in the smart-home environment. This device needs to infer ownership change of *other* devices and thus needs context information *about* these other devices instead of *for* the devices themselves, as is the case in chownIoT. This new system, independent chownIoT or iChownIoT, adapts FoundIoT to infer ownership change from the perspective of other devices instead of its own as originally designed.

During the FoundIoT-inspired monitoring process, if iChownIoT notices that a device is missing from the expected devices in the environment, it can notify the user on their smartphone. If the user agrees that indeed there was a change of ownership, then they can take the necessary action to secure their personal data. In such a scenario, even false detection of ownership change will be helpful for the user as a notification of a missing device due to some other technical failures.

In the space between complete vendor independence (iChownIoT) and vendors implementing chownIoT, vendors of IoT devices can choose to use these notifications from the independent monitoring device as a service. For instance, the vendors can open up some web APIs using their cloud services to receive information about changes in the context of their particular device from iChownIoT and take necessary measures to secure the user's personal data. However, in terms of robustness, the basic chownIoT on IoT device is still a better solution as it is able to apply a protection mechanism immediately on the device in case of ownership change rather than depending on a third party such as user or vendor for that.

8 Conclusions

In this work, we present an automatic context-based technique to improve the privacy of smart-home IoT devices during ownership change. While there exists related work for several different aspects of our solution, we are aiming to bridge the research gap for the specific problem. In our evaluation, we found that we can protect owners from each other, unless they have special equipment to read from the IoT device while it is switched off, at low cost of overhead and resource requirements suitable for most IoT setups. Even if the detected context change is a false positive, the current owner need not be inconvenienced. There are, however, some limitations of chownIoT in its pure form: our solution hinges on the adoption by IoT device manufacturers or service providers, which may be an unrealistic assumption, and a more sophisticated context may improve accuracy. We present a vendor-independent version, iChownIoT. Though limited by the lack of vendor cooperation, such a system can at least alert the user that an ownership change was detected and action is needed. iChownIoT adapts FoundIoT [22], which builds on the first version of chownIoT [10] to make use of richer and forge-proof context.

Acknowledgements. This work is supported by the Academy of Finland under the WiFiUS program (grant 309994), the Wallenberg AI, Autonomous Systems and Software Program (WASP), and the Swedish Foundation for Strategic Research (grant SSF FFL09-0086).

References

1. Apthorpe, N., Reisman, D., Feamster, N.: Closing the blinds: four strategies for protecting smart home privacy from network observers. arXiv preprint arXiv:1705.06809 (2017)
2. Apthorpe, N., Reisman, D., Feamster, N.: A smart home is no castle: Privacy vulnerabilities of encrypted IoT traffic. arXiv preprint arXiv:1705.06805 (2017)

3. Bohn, J.: Instant personalization and temporary ownership of handheld devices. In: 2004 Sixth IEEE Workshop on Mobile Computing Systems and Applications, WMCSA 2004, pp. 134–143. IEEE (2004)
4. Ertaul, L., Mudan, A., Sarfaraz, N.: Performance comparison of AES-CCM and AES-GCM authenticated encryption modes. In: Proceedings of the International Conference on Security and Management (SAM), p. 331. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp) (2016)
5. Francis, L., Hancke, G., Mayes, K., Markantonakis, K.: Practical NFC peer-to-peer relay attack using mobile phones. In: Ors Yalcin, S.B. (ed.) RFIDSec 2010. LNCS, vol. 6370, pp. 35–49. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-16822-2_4
6. Giry, D.: Keylength - NIST report on cryptographic key length and cryptoperiod (2016). https://www.keylength.com/en/4/ (2017). Accessed 26 May 2017
7. Hu, J., Weaver, A.C.: A dynamic, context-aware security infrastructure for distributed healthcare applications. In: Proceedings of the First Workshop on Pervasive Privacy Security, Privacy, and Trust, pp. 1–8. Citeseer (2004)
8. Jih, W.R., Cheng, S.y., Hsu, J.Y., Tsai, T.M., et al.: Context-aware access control in pervasive healthcare. In: Computer Science and Information Engineering, National Taiwan University, Taiwan (2005)
9. Kapsalis, V., Hadellis, L., Karelis, D., Koubias, S.: A dynamic context-aware access control architecture for e-services. Comput. Secur. **25**(7), 507–521 (2006)
10. Khan, M.: Enhancing privacy in IoT devices through automated handling of ownership change. Master's thesis, School of Science, Aalto University, Finland 28 August 2017. http://urn.fi/URN:NBN:fi:aalto-201709046805
11. Kumar, Y., Munjal, R., Sharma, H.: Comparison of symmetric and asymmetric cryptography with existing vulnerabilities and countermeasures. Int. J. Comput. Sci. Manag. Stud. **11**(03), 60–63 (2011)
12. Lenstra, A.K., Verheul, E.R.: Selecting cryptographic key sizes. J. Cryptol. **14**(4), 255–293 (2001)
13. McGrew, D., Bailey, D.: AES-CCM cipher suites for Transport Layer Security (TLS). Technical report (2012)
14. Miettinen, M., Asokan, N., Nguyen, T.D., Sadeghi, A.R., Sobhani, M.: Context-based zero-interaction pairing and key evolution for advanced personal devices. In: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, pp. 880–891. ACM (2014)
15. Miettinen, M., Heuser, S., Kronz, W., Sadeghi, A.R., Asokan, N.: ConXsense: automated context classification for context-aware access control. In: Proceedings of the 9th ACM Symposium on Information, Computer and Communications Security, pp. 293–304. ACM (2014)
16. Miettinen, M., Marchal, S., Hafeez, I., Asokan, N., Sadeghi, A.R., Tarkoma, S.: IoT Sentinel: Automated Device-Type Identification for Security Enforcement in IoT. arXiv preprint arXiv:1611.04880 (2016)
17. Pradeep, B., Singh, S.: Ownership authentication transfer protocol for ubiquitous computing devices. arXiv preprint arXiv:1208.1712 (2012)
18. Ren, B., Liu, C., Cheng, B., Hong, S., Zhao, S., Chen, J.: EasyGuard: enhanced context-aware adaptive access control system for android platform: poster. In: Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking, pp. 458–459. ACM (2016)
19. Rescorla, E.: RFC 2631 - Diffie-Hellman key agreement method. https://tools.ietf.org/html/rfc2631 (1999). Accessed 04 May 2017

20. Shrestha, B., Saxena, N., Truong, H.T.T., Asokan, N.: Contextual proximity detection in the face of context-manipulating adversaries. arXiv preprint arXiv:1511.00905 (2015)
21. Tam, P., Newmarch, J.: Protocol for ownership of physical objects in ubiquitous computing environments. In: IADIS International Conference E-Society 2004, pp. 614–621 (2004)
22. Valiev, A.: Automatic ownership change detection for IoT devices. G2 pro gradu, diplomityö 20 August 2018. http://urn.fi/URN:NBN:fi:aalto-201809034781
23. Whiting, D., Housley, R., Ferguson, N.: Counter with CBC-MAC (CCM). Technical report (2003)
24. Wullems, C., Looi, M., Clark, A.: Towards context-aware security: an authorization architecture for intranet environments. In: Proceedings of the Second IEEE Annual Conference on Pervasive Computing and Communications Workshops 2004, pp. 132–137. IEEE (2004)
25. Zhang, G., Parashar, M.: Context-aware dynamic access control for pervasive applications. In: Proceedings of the Communication Networks and Distributed Systems Modeling and Simulation Conference, pp. 21–30 (2004)
26. Zhang, L., McDowell, W.C.: Am i really at risk? Determinants of online users' intentions to use strong passwords. J. Internet Commer. **8**(3–4), 180–197 (2009)

Is Privacy Controllable?

Yefim Shulman(✉) and Joachim Meyer

Tel Aviv University, 6997801 Tel Aviv, Israel
efimshulman@mail.tau.ac.il, jmeyer@tau.ac.il

Abstract. One of the major views of privacy associates privacy with the control over information. This gives rise to the question how controllable privacy actually is. In this paper, we adapt certain formal methods of control theory and investigate the implications of a control theoretic analysis of privacy. We look at how control and feedback mechanisms have been studied in the privacy literature. Relying on the control theoretic framework, we develop a simplistic conceptual control model of privacy, formulate privacy controllability issues and suggest directions for possible research.

Keywords: Privacy · Feedback · Information disclosure · Human control · Closed-loop control · Feedback control

1 Introduction

Casually used in colloquial conversations, the term “privacy” in all its complexity and prominence appears in philosophical, legal, political, scientific and technological discussions. Although there exist numerous definitions, Solove [46] states in his *A Taxonomy of Privacy* that “Privacy is a concept in disarray”, which “suffers from an embarrassment of meanings” (p. 477). Incidentally, the situation has hardly improved ever since.

While scholars struggle with privacy definitions, the general public (the users) struggle with their online privacy settings, as has been abundantly demonstrated in the security and privacy literature. The seeming futility of information control and the lack of functional transparency lead people to feel helpless.[1] In spite of the diversity of approaches, many discussions of privacy tie it to some form of control (control of access to information, use of information, distribution of information, etc.). We take these terms literally.

If mainstream privacy research embraces the understanding of privacy as control, is there a way to analyse the controllability of privacy? Can we borrow from formal methods of control theory to broaden our understanding of privacy issues, at least when it is appropriate to define privacy as control over information?

[1] See [50] for a case of American consumers being resigned to giving up their data in exchange for commercial offers, rather than engaging in cost-benefit analyses.

Published by Springer Nature Switzerland AG 2019
E. Kosta et al. (Eds.): Privacy and Identity 2018, IFIP AICT 547, pp. 222–238, 2019.
https://doi.org/10.1007/978-3-030-16744-8_15

This paper is organised in the following way. Section 2 of this paper provides the scope of the problem. Here we discuss the extent of privacy determined by control and the reach of control theory.

In Sect. 3 of this paper we apply a control theoretic framework for a conceptual analysis of privacy as control over information.

Section 4 presents a discussion of the applicability, limitations and relevance of our analysis to contemporary research in the privacy literature.

This paper is a first attempt to analyze privacy within the framework of control theory. We map control theory onto privacy, manifested as control over personal information from a user's perspective. Our analysis looks at privacy on a micro level, dealing with the topic in the meaning with which it is used in social and computer science discussions (as opposed to legal, political and philosophical interpretations). Starting from Sect. 3, we use the term "privacy" interchangeably and as a shorthand for "personal information" and its disclosure. Our analysis can serve as a conceptual framework for discussions of privacy and its implications in different contexts.

2 Privacy as Control over Information, Control as a Theory

Privacy is a permeating concept, which has no generally accepted definition throughout all disciplines. Privacy definitions[2] are often formulated through descriptions of the features and properties of privacy and even by writing off constructs which are not privacy [45]. Incidentally, a bibliometric analysis of computer and information ethics literature has revealed privacy as one of three major concepts in that field [19]. It must be noted, however, that the authors make an unsubstantiated claim about differences between the American and the European approaches to privacy, based on their clustering results, where "data protection" fell into the ethics, rather than the privacy cluster.[3]

In ontological attempts to determine what privacy is, scholars often arrive at the same conclusion: general privacy is contextual. It may be internalised through different conceptualisations by different individuals [45].

Additional peculiarities of the concept of privacy come to light, when one is reminded that privacy and the underlying notions may be relative. For example, they may not have simultaneously direct and corresponding translations into other languages. Smith et al. ([45], p. 996) write: "Privacy corresponds to the desire of a person to control the disclosure of personal information [...]" while "[...] confidentiality corresponds to the controlled release of personal information to an

[2] In this paper, we are not concerned with formal definitions of privacy used in cryptography and privacy-enhancing technologies (i.e., differential privacy [15], l-diversity and (n,t)-closeness [29], etc.).

[3] The observed effect could be an artefact of their literature sample (which was not focused on- and, thus, might not be representative of privacy research), and (or) sampling method (picking selected journals in computer and information ethics without attending to the geographical and authorship scope of those journals).

information custodian under an agreement that limits the extent and conditions under which that information may be used or released further". Nevertheless, the term "privacy policy" is conventionally used in English. Simultaneously, in any software application or website in Russian the same document is referred to as "policy of confidentiality" (literal translation), when it contains specifications of data processing, data protection measures and personal information collection.[4]

We can, however, resort to some general conceptions and phenomena observed in the privacy literature. Thus, major approaches to define privacy in philosophical, legal and scientific writings include: privacy as control, privacy as a right, and privacy as an economic good. By and large, the meaning of privacy is attributed arguably to control over information, restriction of access, human dignity, social relationship and intimacy ([12,34,45]).

Privacy as control is a prominent and distinctive approach in philosophical and legal thought, and most definitions include features and properties, which are associated with the term "control". In fact, major theoreticians of privacy, including Warren and Brandeis [52], Fried [16], and Parent [36] refer to privacy as some form of control over information.

Alan Westin defined privacy as "the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others" ([53], p. 7).

Joseph Kupfer argues that "by providing control over information about and access to ourselves, privacy enables us to define ourselves socially in terms of intimate relationship" ([28], p. 86).

Adam Moore derives the following definition: "A right to privacy is a right to control access to and uses of- places, bodies, and personal information" ([34], p. 421).

Perhaps the view of privacy as control over information is so widespread, because it resonates more easily (enables operationality) with research in information systems, behavioral and cognitive psychology, and marketing management.

The term "control" has a specific meaning in engineering, where it is used within the framework of control theory (see [4,14] and [30] for the theoretical framework and applications). Control theory is the basis for engineering models of control of systems and processes, including those that involve a human in the control loop (see [21] and [42] for applications to human performance).

Control theory has been successfully applied to the modelling of manual control over a physical system in human factors research. Concepts of control theory have been borrowed by, and have been productively adjusted to the field of social psychology [32]. Optimal control models in economics belong to a family of optimal control strategies of control theory. The use of computational models of behavior, including control theory, is advocated by psychology scholars in

[4] In reality, "privacy" is directly translated as "privateness", while the latter corresponds to a "degree of inviolability of private life", whereas, in fact, "privacy" corresponds to several control-, protection- or jurisprudence-related terms in the Russian language (confidentiality being one).

management and organisation science [51]. Control theory has found its way into life sciences [8], as an alternative form of Powers' perceptual control theory (PCT)[5], when the goal of a dynamic system resides within the system itself (see [7] for a survey of biological, neurobiological and psychological implementations of negative feedback loops).

It does not seem a stretch to assume that people get some form of feedback on their behavior, including privacy-related actions. The information people receive about outcomes of their actions may alter their behavior, which aims to reach some comfort zone, i.e., a certain level of physical, mental or emotional well-being. Of course, people may not always be able to associate the feedback with the action (and cause with effect for that matter), a growing concern for privacy researchers and data protection professionals. With recent developments, it also becomes a concern for the general public.

Control theory is instrumental and productive when it is applied to phenomena where feedback plays some role. In this conceptual paper we ask what could be the implications from analyzing privacy as control in the framework of control theory?

Section 3 presents our attempt to tackle this question.

3 Control Theoretic Analysis of Privacy

Control theory distinguishes between *open-loop* and *closed-loop (feedback) control.* In an open-loop control system, some input is fed into the system, and a process runs its course, without further interventions in the process. In closed-loop (or feedback) control the output of the process is measured, and some information about the output is provided as feedback, serving to minimize the difference between a desired state and an existing state.

In the context of privacy, our system consists of:

- a person (the controller) who performs some actions (e.g., permits an app to access information about location or contacts, or posts some information on a social network);
- some process that runs, depending partly on the person's actions;
- the controlled output, which is the disclosure of information about the person or its use;
- and the evaluation of the level of disclosure of personal information[6].

Any part of the process may be affected by external factors (the environment) that may introduce noise, or disturbances.

A control theoretic analysis of the user actions assumes that the output (i.e., the information disclosure) has some value that can be compared to a

[5] Originates in [39].

[6] From this point on, for the sake of convenience, we may use the term "privacy" as a shorthand for "personal information" and its disclosure.

desired value (e.g., expressed through some personally comfortable level of disclosure). We can assume that information disclosure has some benefits (financial, emotional, social, etc.) and some possible costs. The overall outcome is the sum (or other combination) of the benefits and the costs. The exact functions by which the benefits and the costs change as more information is revealed, depend, of course, on the person, the information, the party receiving access to the information, and the specific context, in which the information is revealed. We depict a demonstration of the behavior of the controlled and output variables in Fig. 1. For the sake of the demonstration, we assume that benefits increase monotonously with diminishing marginal returns, and that costs increase exponentially as the amount of revealed information increases.

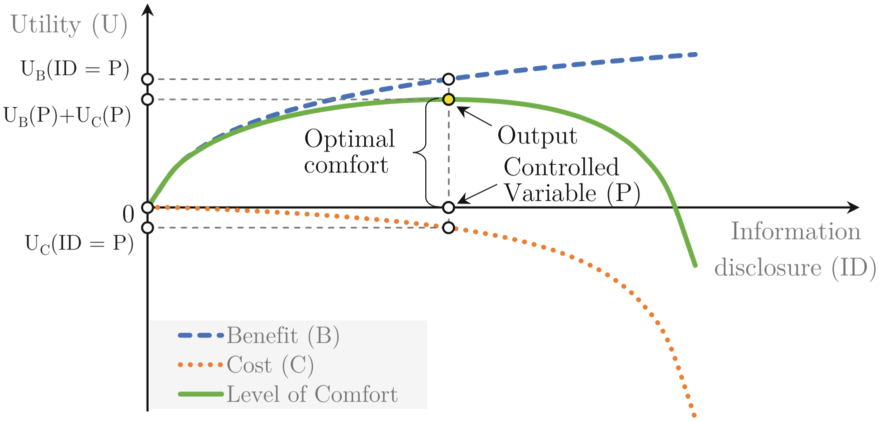


Fig. 1. Privacy comfort as utility based on costs and benefits curves

The change in the *Information Disclosure* (ID) leads to desired and undesired consequences for the controller, i.e., benefits and costs, respectively. Estimating the difference between costs and benefits for each level of ($ID = P$) while exercising control over P, along the abscissa axis, we seek to maximize the *Level of Comfort* from disclosure. The character of control, as well as the optimality criterion will differ, based on the shapes of the benefits and costs functions, as defined by individual and momentary factors.

We may assume a more complex scenario, where the control variable P is multidimensional: i.e., it may contain multiple types and corresponding amounts of disclosed information ($p_{[type,amount]} \in P$). The controller wants to maximize both pleasurable effects and privacy. The optimal comfort level can be reached when there is no possibility to improve either of the two outcomes, while maintaining the same value for the other, mapping an optimal *Output* as a Pareto frontier, as we show in Fig. 2.

The Pareto frontier represents the *Output* space, while the area under the curve contains suboptimal solutions that can be improved. The area above the

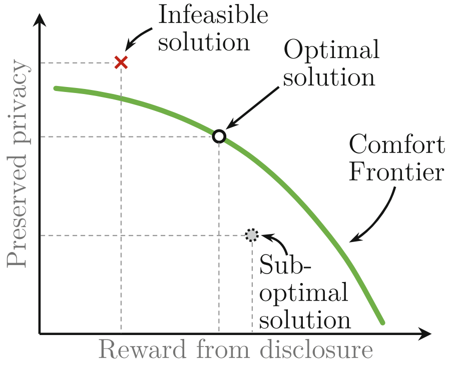


Fig. 2. Privacy comfort as Pareto-optimal frontier

frontier contains infeasible solutions, due to existing constraints on the amount of privacy preserved and the benefits gained for each level of personal information disclosure.

For the person it is desirable to be on the Pareto frontier. The person has to consider two important questions: (1) Am I on the frontier? If not, what can I do to get there? (2) Where on the frontier do I prefer to be?

Figure 3 contains a depiction of the proposed privacy control model in the form of a block diagram – a widely used way to depict dynamic systems in control theory[7].

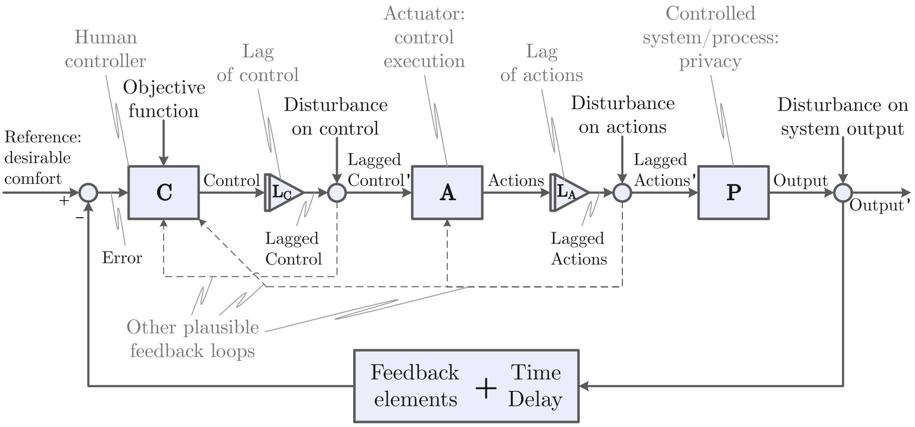


Fig. 3. A block diagram representation of the privacy control model

[7] Both in dynamic systems (e.g., [4,14] and [30]) and human factors (e.g., [21] and [42]) block diagrams are used for concise depictions of systems.

Each box in the block diagram is a separate subsystem. A privacy level P is the amount of disclosed information, which is a controlled variable in our conceptual model. The $Output$[8] of our control model is the utility variable, representing the level of comfort, given the level of personal disclosure.

A human controller C is a person, performing actions and seeking to achieve some comfortable level of personal information disclosure. The human controller C adjusts the privacy level P using an actuator A, which is some form of a tool, system or service through which privacy-related actions are taken (e.g., settings adjustment, information sharing, etc.).

Arrows represent signals flowing between the elements of the system. Arrows going into a certain element are input signals for this element, and arrows going out of an element are output signals of this element. An input signal to the whole system is $Reference$, which is some comfortable level of personal information disclosure that a person desires to achieve. An output signal of the whole system is $Output$ or $Output'$ described above. The circular blocks are "comparators" that sum up inflowing input signals, producing output(s).

Each triangular block represents a lag of output (that is an effect when the output of the action or process is not proportional to the input). Intermediate outputs may be non-linear, due to disturbances from the environment and the properties of the medium. The order of the lag is undefined, and the symbol is used for representation.

A time lag (delay) may also be present throughout the system. The most important time delay appears with the feedback loop (shown explicitly in Fig. 3). The feedback loop includes both the information on the reached comfort value ($Output$ or $Output'$) and the time delay until the human controller receives the feedback. The time delay is a varying quantity for each individual at each point in time, which limits its predictability.

We approach the conceptual privacy control model through different topics from the systems control literature, and we reveal multiple controllability issues from the standpoint of the individual, summarized in Table 1.

These issues preclude us from asserting that humans and their personal information make up controllable systems on their own. Yet, this is not a reason for despair. It only shows that a one-to-one straightforward mapping of a control theoretic framework onto personal information disclosure cannot immediately produce beneficial results. We discuss possible implications and contributions of control theory to privacy in the following Sect. 4, alongside the discussion on the relevance and current standings of the notions of "control" and "feedback" in the privacy literature.

In Sect. 4 we also proceed to discuss the contribution, applicability and limitations of our model. We further investigate the existing empirical privacy research to better understand how our conceptual analysis fares with the observed reality.

[8] Or $Output'$, if any disturbance is introduced into the system after a certain $Output$ is achieved.

Table 1. Privacy controllability issues

Issue	Description	Element
Feedback time delay	The consequences of actions arrive at an uncertain time, and they may be not attributed to the actions	Feedback loop
Physical feedback lag	The consequences of actions may arrive non-linearly and are prone to alterations within feedback elements	Feedback loop
Multiple feedback loops	Each element of the control system (Fig. 3) may have its own feedback loop(s)	Model
Complexity	Elements of the system may constitute control subsystems with all the corresponding issues	Model
Order of control	Intermediate signals of controls and actions may have different non-linear profiles and may require learning from the human controller	Forward control path
Multiple physical lags	Multiple linear and non-linear relations exist between the elements of the system	Outputs, signals
Momentariness and individual differences	The privacy control model may have to be non-stationary, as privacy behavior and preferences may vary over time for different individuals	Concept, assumptions
Linearity and time-invariance	Humans and privacy are perhaps non-linear time-variant systems: the output is not proportional to the input; and at different points in time, the system output may differ for the same system input	Concept, assumptions

4 Discussion

In Sect. 4.1 we present an overview of, and discussion on how privacy research has handled the notions of "control" and "feedback", and what benefits and contributions the control theoretic analysis can potentially bring. Section 4.2 describes potential research directions, driven by the control theoretic approach and clarifies the scope and limitations of this paper.

4.1 Control, Feedback and Privacy Research

The effects and implications of providing users with control (and the feeling of control) over their personal information and its use have been abundantly studied. Control over personal information constitutes a whole dimension of privacy concerns for users ([24,31]). Interestingly, perceived control over information does not seem to impact the level of related privacy concerns, whenever this control is perceived to be low [26]. That result is in line with findings on feeling resigned regarding one's privacy ([50], discussed in Sect. 2), and that a perceived higher level of control may increase the willingness to disclose personal information [6]. On a more narrow approach, it has been shown that an incremental increase in controllability over information collection may make users more tolerant towards tailored online advertisement [9]. Users may give away control over personal information as a result of the framing of an online service offer [3]. Additionally, there are multiple studies in the privacy decision-making literature that operationalize "privacy control" and "perceived privacy control" as either dependent or independent variables in their corresponding models (e.g., [13,18,27]; see [43] for a review of more papers on the topic).

Technological implementations of control over privacy have been mostly concerned with cryptography and systems architecture (e.g., [11,17,22,23,44] and others), or functionality enabling and interface design (e.g., [10,25,33,40,47], and many more).[9]

However, the privacy control literature so far has used the term "control" mostly as a mean of adjusting disclosure preferences (e.g., adjust settings, contact support, etc.), a level of disclosure adjustments that may be introduced (e.g., change settings and give or revoke consent in full or partially, on a level of a server, an application, a location, an enterprise, etc.), or as a plausible adjustment that can be made realistically (e.g., start disclosing, stop disclosing, delete information, etc.). Using a control theoretic analysis, and introducing a closed-loop control that can inform users about achieved disclosure outcomes or privacy states, we can start to use the term "control" more productively. With properly associated feedback we may know about achieved privacy states and disclosure outcomes. It may enable us to talk about "controllability" of privacy states and disclosure outcomes. Analysing controllability of privacy may help us answer questions about whether desired states or outcomes are reachable, and whether they have been reached.

Empirical privacy research in computer science and human-computer interaction has already given some attention to feedback processes and their impact on privacy behavior and perceptions (e.g., [41] and [49] dealing with feedback design, [37] and [38] looking at effects of feedback recency).

Trying to answer the question of how important a feedback mechanism can be for managing personal privacy, Tsai et al. [48] demonstrate that the presence of feedback in a location-sharing scenario makes people feel more comfortable with disclosure of personal information and alleviates the level of privacy concerns. Thus, both the aforementioned increase in perceived control and the presence of feedback raise the people's information sharing propensity. These findings bear risks, alongside obvious benefits, and they should be treated with caution.

[9] We invite our readers to explore independently the world of patents on privacy controls.

In a paper concerned with the state of transparency enhancing technologies Murmann and Fischer-Hübner [35] provide a categorisation and assessment of existing transparency enhancing technologies, which greatly rely on feedback mechanisms. The authors note that without a feedback mechanism, users may be unable to make rational decisions about the use of transparency enhancing technologies and exercise control over them.

Hoyle et al. [20] study relationships between content publishers and content users. Their findings lead to the conclusion that feedback mechanism may be a useful instrument for balancing personal information disclosure and exposure of the publishers' content.

Bargh et al. [5] explore relationships between data controllers and data processors. The authors define "feedback" as any backwards-directed data flow from data processors to data controllers that facilitates forward-directed data flow. Their conceptual paper is concerned with the public policy discussion on procedural feedback between different agents dealing with personal data.

Discussing nudges in privacy and security decision-making performed with the use of information, Acquisti et al. [2] distinguish between education and feedback, where education is responsible for affecting future decision-making, while feedback is capable of altering behavior at the current moment or over time. In terms of control theory, education corresponds to open-loop system dynamics, and feedback naturally relates to close-looped systems. It must be noted, however, that the authors use some colloquial understanding of the term "feedback", resulting in a debatable claim that "feedback can also inform about expected and actual outcomes before or immediately after making a decision" ([2], p. 44:13). A process that informs about "expected outcomes" before an action perhaps constitutes a separate notion (e.g., predictive modelling, feed-forward control, predictive inference, hypothesising, etc. – depending on the context), which is different from what is understood by feedback in control theory.

As we show, the term "feedback" in the privacy literature is often only loosely defined. If we want to proceed with analyses in the control theoretic framework, then we should align our understanding of the term "feedback" with the control theoretic definitions. One can use the following definition as an anchor point: *Feedback* is "the modification, adjustment, or control of a process or system (as a social situation or a biological mechanism) by a result or effect of the process, esp. by a difference between a desired and an actual result; information about the result of a process, experiment, etc.; a response" [1].

Application of the control theoretic framework may not only imply the stricter definition of feedback. It is not any information related to privacy choices that is provided to the decision-maker. Control theoretic feedback returns information on achieved levels of outcomes, which decision-makers can compare to their own goal levels. This feedback hardly appears in a simple obvious way in reality.

As was mentioned before, the feedback mechanism can perhaps be implemented in technology. This technology, if it is built with control theoretic considerations, will differ from existing privacy-enhancing technologies and basic recommender systems. Existing systems provide recommendations derived from:

- profiles of users and associations between profiles and specific users;
- the accumulated statistics (historical data) on privacy outcomes, which allow predictions of desired or unwanted outcomes for specific types of users;
- the best practices and advice from scholars and professionals;
- the "raw" information about who, how and when can, may and will access the users' personal data, if they proceed with a given option;
- the same "raw" information about who, how and when (and possibly for what purpose) someone actually accessed specific users' personal data;
- and other external data.

The desired state of privacy, however, changes over time (a person may become more informed, more or less concerned, more or less alert, etc., as life changes). In order to resort to some comfort levels, a person needs to figure out what privacy-related action to perform. A person would need sufficient understanding of causal and temporal relationships between actions and privacy-entailing consequences, as well as of the character and form of these relationships. It is questionable that people are capable and willing to do that. Conversely, adjusting one's privacy to some comfort zone can be facilitated with the addition of a control-theoretic feedback loop, providing the following advantages:

1. Privacy outcomes of actions may be traced back to those actions in terms of cause and time, through the nature of feedback, accounting for time delays.
2. Desired privacy outcomes may be compared with actual privacy outcomes.
3. Effects of actions on privacy may be associated with these actions, even when the form of relationships between the actions and their effects is not proportional (more complex than one-to-one mapping, e.g., "if-then" rules). This is by accounting for physical lag and multiple elements and loops.

We face several issues, when we attempt to model the privacy feedback loop with control theory:

- Is there a way for the user or decision-maker to know and define their desirable outcomes and states of privacy?
- Is there a way for the user or decision-maker to associate feedback about privacy outcomes and implications with actions that have led to these outcomes and implications?
- Should users and decision-makers be nudged towards some optimal privacy configuration? What would be the optimality criteria in that case?
- Should users and decision-makers be nudged towards some specific privacy actions? What would be the justification for and against certain actions?
- General controllability issues highlighted in Table 1.

Technological implementation of the feedback loop may help people make better adjustments of their privacy behavior. The feedback may be partially approximated with quasi-linearity, modelled with anticipation (e.g., a quickened display), Kalman filter and finite state control with time lag and other concepts from control theory.

Alternatively, we may model privacy as an open-loop system. The improvement of personal disclosure behavior may be achieved through enriching people's prior knowledge. One way to implement that is to provide relevant privacy education, and training.

Thus, the control theory framework can be used to come up with an analysis of privacy and to inform the development of privacy solutions.

4.2 Future Work, Scope and Limitations

The control theoretic approach provides various constructs and ideas to be tested in privacy-related research, including privacy attitudes and behaviors, especially decision-making.

This conceptual analysis reveals several potential research directions:

- Study of feedback elements and their effects on privacy attitudes and behavior: time lag (delay) for feedback to flow between outcomes and actions, physical lag between an action and its effect on privacy, etc.
- Study of relations between different elements in a system, involving a user, the user's privacy state, the user's desirable privacy, information disclosure outcomes, evaluations of outcomes, external factors, and feedback loops between these elements.
- Study of users' decision-making, when feedback loops are involved.
- Modelling certain elements of the privacy decision-making process in more detail as separate control subsystems. Some of these subsystems may be suitable for more formal control theoretic representation (e.g., application or server permission management).
- Modelling individual differences in the control theoretic framework.
- Feedback-control loop implementation in practice.
- And, without a doubt, others.

We emphasize that, even though this paper is devoted to a conceptual control theoretic analysis of privacy, it can be naturally developed towards modelling and analyzing privacy control in information systems. Technological implementations of the privacy feedback loop, based on control theoretic principles, may facilitate individuals' control over personal information and its disclosure and may raise awareness of their current privacy states.

An applied control theoretic analysis of privacy may be appropriate and particularly valuable when it comes to the implementation of privacy-by-design principles. On the one hand, it may help in the evaluation of a system's compliance with the privacy-by-design principles via assessing controllability (and stability) of users' personal information disclosure. It may also consult the development of information systems with privacy-by-design in mind. On the other hand, an explicit feedback loop mechanism is easier to develop for a system, which is adhering to the privacy-by-design principles.

We also note that this paper is not an exhaustive analysis of privacy in terms of control theory. We did not extend our paper with an alternative analysis, based on Powers' perceptual control theory, for the sake of keeping the scope and the rationale of the paper within reason and to avoid theoretical debates around PCT's assumptions and applicability in psychology. We also did not venture into the analysis of open-loop control of privacy. However, analyses of privacy in the literature so far have already treated privacy in a somewhat similar way to an open-loop system. It must also be noted that control theory is instrumental, when we deal with closed-loop control.

The presented conceptual analysis of privacy as an object of control does not map the whole body of control theoretic constructs onto privacy research. We have omitted multiple domain- and application-specific concepts and tools. Our attempt has been to evaluate, transfer and adjust those control theoretic constructs that seem to bear benefits and can be fit to privacy-related research. For the sake of simplicity we also omitted more formal or specialized aspects and items (e.g., underlying partial differential equations, Kalman filter, feed-forward models, etc.), which still may be useful in further analyses of the subject and in relation to specific problems.

5 Conclusion

In this paper we apply the theoretical framework of control theory to privacy, according to one of the major understandings of privacy as a person's control over information. We conceptualize privacy control with a human controller at its core, and raise questions about the controllability of such a system.

The conceptual model of privacy that we developed and presented in this paper allows us to reveal multiple controllability issues of privacy, and we propose several directions for future research with control theory in mind.

We further discuss the relation and relevance of our proposed model and of the control theoretic analysis to privacy. We study the existing body of empirical privacy research and find multiple connections in how the privacy literature used and highlighted the notions of control and feedback. We also present our analysis of applicability of the proposed approach, as well as the challenges and opportunities of modelling personal information disclosure as a dynamic system with open and closed-loop control.

One particular question we raise concerns the plausibility of a feedback control loop of privacy. If and when the implementation of a feedback control loop is infeasible, privacy may be analyzed as an open-loop control system. Our analysis shows that privacy may be a phenomenon that is inherently difficult to control. Some aids can perhaps be used to make it more controllable, such as indications about possible privacy implications of actions, or recommendations on privacy optimisation through the development of privacy-related feedback control loops.

Funding. This research is partially funded by the EU Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 675730 "Privacy and Us".

References

1. feedback — feed-back, n.: OED Online. Oxford University Press, July 2018. http://www.oed.com/view/Entry/68965
2. Acquisti, A., et al.: Nudges for privacy and security: understanding and assisting users' choices online. ACM Comput. Surv. **50**(3), 44:1–44:41 (2017). https://doi.org/10.1145/3054926
3. Angulo, J., Wästlund, E., Högberg, J.: What would it take for you to tell your secrets to a cloud? In: Bernsmed, K., Fischer-Hübner, S. (eds.) NordSec 2014. LNCS, vol. 8788, pp. 129–145. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11599-3_8
4. Aström, K.J., Murray, R.M.: Feedback Systems: An Introduction for Scientists and Engineers. Princeton University Press, Princeton (2010)
5. Bargh, M.S., Meijer, R., Choenni, S., Conradie, P.: Privacy protection in data sharing: towards feedback based solutions. In: Proceedings of the 8th International Conference on Theory and Practice of Electronic Governance ICEGOV 2014, pp. 28–36. ACM, New York (2014). https://doi.org/10.1145/2691195.2691279
6. Brandimarte, L., Acquisti, A., Loewenstein, G.: Misplaced confidences: privacy and the control paradox. Soc. Psychol. Pers. Sci. **4**(3), 340–347 (2013). https://doi.org/10.1177/1948550612455931
7. Carey, T., Mansell, W., Tai, S.: A biopsychosocial model based on negative feedback and control. Front. Hum. Neurosci. **8**, 94 (2014). https://doi.org/10.3389/fnhum.2014.00094
8. Carver, C.S.: Control processes, priority management, and affective dynamics. Emot. Rev. **7**(4), 301–307 (2015). https://doi.org/10.1177/1754073915590616
9. Chanchary, F., Chiasson, S.: User perceptions of sharing, advertising, and tracking. In: Eleventh Symposium On Usable Privacy and Security (SOUPS) 2015, pp. 53–67. USENIX Association, Ottawa (2015). https://www.usenix.org/conference/soups2015/proceedings/presentation/chanchary
10. Colnago, J., Guardia, H.: How to inform privacy agents on preferred level of user control? In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: UbiComp Adjunct 2016, pp. 1542–1547. ACM, New York (2016). https://doi.org/10.1145/2968219.2968546
11. Coppersmith, D., Mintzer, F.C., Tresser, C.P., Wu, C.W., Yeung, M.M.: Fragile imperceptible digital watermark with privacy control, vol. 3657, pp. 79–85 (1999). https://doi.org/10.1117/12.344705
12. DeCew, J.: Privacy. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy, Spring 2018 edn. Metaphysics Research Lab, Stanford University (2018)
13. Dinev, T., Xu, H., Smith, J.H., Hart, P.: Information privacy and correlates: an empirical attempt to bridge and distinguish privacy-related concepts. Eur. J. Inf. Syst. **22**(3), 295–316 (2013). https://doi.org/10.1057/ejis.2012.23
14. Doyle, J.C., Francis, B.A., Tannenbaum, A.R.: Feedback Control Theory. Courier Corporation (2013)
15. Dwork, C.: Differential privacy: a survey of results. In: Agrawal, M., Du, D., Duan, Z., Li, A. (eds.) TAMC 2008. LNCS, vol. 4978, pp. 1–19. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-79228-4_1
16. Fried, C.: An anatomy of values: problems of personal choice (1970)
17. Gong, Y., Wei, L., Guo, Y., Zhang, C., Fang, Y.: Optimal task recommendation for mobile crowdsourcing with privacy control. IEEE Internet Things J. **3**(5), 745–756 (2016). https://doi.org/10.1109/JIOT.2015.2512282

18. Griffin, C., Rajtmajer, S., Squicciarini, A.: Invited paper: a model of paradoxical privacy behavior in online users. In: 2016 IEEE 2nd International Conference on Collaboration and Internet Computing (CIC), pp. 206–211, November 2016. https://doi.org/10.1109/CIC.2016.037
19. Heersmink, R., van den Hoven, J., van Eck, N.J., van den Berg, J.: Bibliometric mapping of computer and information ethics. Ethics Inf. Technol. **13**(3), 241 (2011). https://doi.org/10.1007/s10676-011-9273-7
20. Hoyle, R., Das, S., Kapadia, A., Lee, A.J., Vaniea, K.: Viewing the viewers: publishers' desires and viewers' privacy concerns in social networks. In: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW 2017, pp. 555–566. ACM, New York (2017). https://doi.org/10.1145/2998181.2998288
21. Jagacinski, R.J., Flach, J.M.: Control Theory for Humans: Quantitative Approaches To Modeling Performance. CRC Press, Boca Raton (2003)
22. Jiang, X., Landay, J.A.: Modeling privacy control in context-aware systems. IEEE Pervasive Comput. **1**(3), 59–63 (2002). https://doi.org/10.1109/MPRV.2002.1037723
23. Jones, S., O'Neill, E.: Feasibility of structural network clustering for group-based privacy control in social networks. In: Proceedings of the Sixth Symposium on Usable Privacy and Security SOUPS 2010, pp. 9:1–9:13. ACM, New York (2010). https://doi.org/10.1145/1837110.1837122
24. Kitkowska, A., Wästlund, E., Meyer, J., Martucci, L.A.: Is it harmful? Re-examining privacy concerns. In: Hansen, M., Kosta, E., Nai-Fovino, I., Fischer-Hübner, S. (eds.) Privacy and Identity 2017. IAICT, vol. 526, pp. 59–75. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-92925-5_5
25. Knijnenburg, B.P., Kobsa, A.: Increasing sharing tendency without reducing satisfaction: finding the best privacy-settings user interface for social networks. In: ICIS (2014)
26. Kowalewski, S., Ziefle, M., Ziegeldorf, H., Wehrle, K.: Like us on facebook! analyzing user preferences regarding privacy settings in Germany. Proc. Manuf. **3**, 815–822 (2015). https://doi.org/10.1016/j.promfg.2015.07.336. http://www.sciencedirect.com/science/article/pii/S2351978915003376. 6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE
27. Krasnova, H., Spiekermann, S., Koroleva, K., Hildebrand, T.: Online social networks: why we disclose. J. Inf. Technol. **25**(2), 109–125 (2010). https://doi.org/10.1057/jit.2010.6
28. Kupfer, J.: Privacy, autonomy, and self-concept. Am. Philos. Q. **24**(1), 81–89 (1987). http://www.jstor.org/stable/20014176
29. Li, N., Li, T., Venkatasubramanian, S.: Closeness: a new privacy measure for data publishing. IEEE Trans. Knowl. Data Eng. **22**(7), 943–956 (2010). https://doi.org/10.1109/TKDE.2009.139
30. Luenberger, D.G.: Introduction to Dynamic Systems: Theory, Models, and Applications, vol. 1. Wiley, New York (1979)
31. Malhotra, N.K., Kim, S.S., Agarwal, J.: Internet users' information privacy concerns (IUIPC): the construct, the scale, and a causal model. Info. Syst. Res. **15**(4), 336–355 (2004). https://doi.org/10.1287/isre.1040.0032
32. Mansell, W., Marken, R.S.: The origins and future of control theory in psychology. Rev. Gen. Psychol. **19**(4), 425–430 (2015)

33. Mehta, V., Bandara, A.K., Price, B.A., Nuseibeh, B.: Privacy itch and scratch: on body privacy warnings and controls. In: Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems CHI EA 2016, pp. 2417–2424. ACM, New York (2016). https://doi.org/10.1145/2851581.2892475
34. Moore, A.: Defining privacy. J. Soc. Philos. **39**(3), 411–428 (2008)
35. Murmann, P., Fischer-Hbner, S.: Tools for achieving usable ex post transparency. a survey. IEEE Access **5**, 22965–22991 (2017). https://doi.org/10.1109/ACCESS.2017.2765539
36. Parent, W.A.: Privacy, morality, and the law. Philos. Public Aff. **12**(4), 269–288 (1983). http://www.jstor.org/stable/2265374
37. Patil, S., Hoyle, R., Schlegel, R., Kapadia, A., Lee, A.J.: Interrupt now or inform later?: Comparing immediate and delayed privacy feedback. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems CHI 2015, pp. 1415–1418. ACM (2015). https://doi.org/10.1145/2702123.2702165
38. Patil, S., Schlegel, R., Kapadia, A., Lee, A.J.: Reflection or action?: How feedback and control affect location sharing decisions. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems CHI 2014, pp. 101–110. ACM, New York (2014). https://doi.org/10.1145/2556288.2557121
39. Powers, W.T., Clark, R.K., McFarland, R.L.: A general feedback theory of human behavior: Part i. Percept. Mot. Skills **11**(1), 71–88 (1960). https://doi.org/10.2466/pms.1960.11.1.71
40. Schaub, F., Balebako, R., Durity, A.L., Cranor, L.F.: A design space for effective privacy notices. In: Eleventh Symposium On Usable Privacy and Security SOUPS 2015, pp. 1–17. USENIX Association, Ottawa (2015). https://www.usenix.org/conference/soups2015/proceedings/presentation/schaub
41. Schlegel, R., Kapadia, A., Lee, A.J.: Eyeing your exposure: quantifying and controlling information sharing for improved privacy. In: Proceedings of the Seventh Symposium on Usable Privacy and Security, SOUPS 2011, pp. 14:1–14:14. ACM, New York (2011). https://doi.org/10.1145/2078827.2078846
42. Sheridan, T.B., Ferrell, W.R.: Man-Machine Systems; Information, Control, and Decision Models of Human Performance. The MIT Press (1974)
43. Shulman, Y.: Towards a broadening of privacy decision-making models: the use of cognitive architectures. In: Hansen, M., Kosta, E., Nai-Fovino, I., Fischer-Hübner, S. (eds.) Privacy and Identity 2017. IAICT, vol. 526, pp. 187–204. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-92925-5_12
44. Sivaraman, V., Gharakheili, H.H., Vishwanath, A., Boreli, R., Mehani, O.: Network-level security and privacy control for smart-home IoT devices. In: 2015 IEEE 11th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), pp. 163–167, October 2015. https://doi.org/10.1109/WiMOB.2015.7347956
45. Smith, H.J., Dinev, T., Xu, H.: Information privacy research: an interdisciplinary review. MIS Q. **35**(4), 989–1016 (2011). http://dl.acm.org/citation.cfm?id=2208940.2208950
46. Solove, D.J.: A taxonomy of privacy. Univ. PA Law Rev. **154**(3), 477–560 (2006)
47. Toch, E., et al.: Locaccino: a privacy-centric location sharing application. In: Proceedings of the 12th ACM International Conference Adjunct Papers on Ubiquitous Computing - Adjunct, UbiComp 2010 Adjunct, pp. 381–382. ACM, New York, (2010). https://doi.org/10.1145/1864431.1864446

48. Tsai, J.Y., Kelley, P., Drielsma, P., Cranor, L.F., Hong, J., Sadeh, N.: Who's viewed you?: The impact of feedback in a mobile location-sharing application. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI 2009, pp. 2003–2012. ACM, New York (2009). https://doi.org/10.1145/1518701.1519005
49. Tsai, L., et al.: Turtle guard: helping android users apply contextual privacy preferences. In: Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017), pp. 145–162. USENIX Association, Santa Clara (2017). https://www.usenix.org/conference/soups2017/technical-sessions/presentation/tsai
50. Turow, J., Hennessy, M.: The tradeoff fallacy: how marketers are misrepresenting American consumers and opening them up to exploitation. SSRN Electron. J., 24 p. (2015). https://doi.org/10.2139/ssrn.2820060
51. Vancouver, J.B., Weinhardt, J.M.: Modeling the mind and the milieu: computational modeling for micro-level organizational researchers. Organ. Res. Methods **15**(4), 602–623 (2012). https://doi.org/10.1177/1094428112449655
52. Warren, S.D., Brandeis, L.D.: The right to privacy, pp. 193–220. Harvard law review (1890)
53. Westin, A.: Privacy and Freedom. Atheneum, New York (1967)

Assessing Theories for Research on Personal Data Transparency

Anette Siebenkäs(✉) and Dirk Stelzer

Fachgebiet Informations- und Wissensmanagement,
Technische Universität Ilmenau, Postfach 10 05 65, 98693 Ilmenau, Germany
{anette.siebenkaes,dirk.stelzer}@tu-ilmenau.de

Abstract. A growing number of business models are based on the collection, processing and dissemination of personal data. For a free decision about the disclosure of personal data, the individual concerned needs transparency as insight into which personal data is collected, processed, passed on to third parties, for what purposes and for what time (Personal Data Transparency, or PDT for short). The intention of this paper is to assess theories for research on PDT. We performed a literature review and explored theories used in research on PDT. We assessed the selected theories that may be appropriate for exploring PDT. Such research may build on several theories that open up different perspectives and enable various fields of study.

Keywords: Personal Data Transparency · Literature review · Theory · Information privacy · Ex-ante transparency · Real-time transparency · Ex-post transparency

1 Introduction

An increasing number of business models are based on the collection, processing and dissemination of personal data [16, 39, 48, 65].

Personal data is defined as "any information relating to an identified or identifiable natural person" [46]. Personal data may be used for purposes that could harm the data subject. This threatens the right to informational self-determination [37]. For a free decision about disclosing personal data, the individual concerned needs transparency.

Transparency requires insight into which personal data is collected, processed, passed on to third parties, for what purposes and for what time. We call transparency of personal data processing Personal Data Transparency (or PDT for short). PDT is a privacy principle and a prerequisite for informational self-determination [22, 46].

Although PDT is demanded by legislators, consumer protection associations, privacy commissioners and data protection officers to ensure consumers' privacy [46] and consumers explicitly ask for transparency [31], findings from several research projects suggest that enhanced transparency may overstrain consumers [26, 40, 59, 62] and decrease users' privacy concerns and risk beliefs [2, 8, 14, 45, 47]. Therefore, enhanced PDT – originally meant as a means of increasing consumer protection – may indeed lead to less privacy.

Published by Springer Nature Switzerland AG 2019
E. Kosta et al. (Eds.): Privacy and Identity 2018, IFIP AICT 547, pp. 239–254, 2019.
https://doi.org/10.1007/978-3-030-16744-8_16

Theories provide a lens for issues and challenges worthy of scientific research. They also help to pose interesting research questions and guide the selection of research methods. Theories are practical because they help to accumulate knowledge and to integrate findings of different scholars and research projects in a systematic manner [21]. Our research may support scholars in identifying, assessing, selecting or adapting theories when exploring PDT.

Research into PDT is a subset of information privacy research. When starting our research, we were working on the assumption that many scholars who explore PDT apply theories that are also used in other areas of information privacy research.

The intention of this paper is to assess theories for research on PDT. In particular, we address the following research questions:

RQ1: Are there theories that can substantially support research in the field of PDT?
RQ2: What are strengths and weaknesses of theories used to investigate PDT?

We followed a two-step approach. First, we performed a literature review to analyse which theories scholars use when investigating PDT. We based our review on Rowe's [49] recommendations for conducting literature reviews. To distinguish conceptual foundations from theories, we drew on Sutton, Staw and Gregor [21, 55]. Then, we defined criteria that a theory appropriate for exploring PDT should cover. We used these criteria for assessing the selected theories.

2 Theories Used in PDT-Research

2.1 Literature Review

For identifying theories appropriate for exploring PDT, we focused on papers published between 2000 and 2017 and queried the following databases: ACM Digital Library, AIS Electronic Library, EBSCO, Elsevier ScienceDirect, IEEE Xplore Digital Library, INFORMS PubsOnline, SpringerLink and Web of Science.

We searched titles, abstracts and keywords with the following search term: *(transparent OR transparency) AND (privacy OR personal data OR personal information)*. Our focus was on journal articles, conference proceedings and book chapters written in English. By reading article titles, abstracts and introductions, we identified and selected papers for further review. We conducted backward and forward searches, following Webster and Watson [64]. We identified 157 papers relevant to PDT. Within these articles, we searched for "theor*" in the full texts of the papers which led to 42 papers for in-depth review. We read relevant passages, in particular theoretical and conceptual foundations. Subsequently, we identified and analysed the original sources of the theories quoted in the papers. Theories quoted in only one of the 42 papers or theories that refer to contexts not directly relevant for the purpose of our research (such as the Theory of Cryptography) were excluded. We included 21 papers in the final selection. Several authors base their research not only on one theory, but combine different theories into a new research construct. Papers that we considered relevant in this context, were assigned to the theory that was predominantly used in the respective papers.

Table 1 gives an overview of theories identified in our study, the original sources and the papers that apply – or at least quote – these theories (Table 1).

Table 1. Theories used in research on PDT

Theories	Sources explaining the theories	Sources applying the theories
Agency Theory Signaling Theory	Eisenhardt [15], Spence [53]	Greenaway et al. [19], Monteleone [36], Pollach [44]
Theory of Reasoned Action (TRA), Theory of Planned Behavior (TPB)	Ajzen and Fishbein [4], Ajzen and Fishbein [3]	Awad and Krishnan [6], Cabinakova et al. [9], Kowatsch and Maass [31]
Technology Acceptance Model (TAM)	Davis [11]	Cabinakova et al. [9], Kowatsch and Maass [31], Zhang and Xu [66]
Theory of Bounded Rationality	Simon [51, 52]	Acquisti et al. [1], Adjerid et al. [2], Brandimarte et al. [8], Monteleone [36], Zhang and Xu [66]
Prospect Theory	Tversky and Kahneman [60, 61], Kahneman and Tversky [29]	Acquisti et al. [1], Adjerid et al. [2], Monteleone [36], Walker [62]
Information Boundary Theory (IBT) Communication Privacy Management Theory (CMPT)	Altmann [5] Petronio [42, 43]	Dinev et al. [14], Hauff et al. [24], Karwatzki et al. [30], Rader [47], Stutzman et al. [54]
Restricted Access/Limited Control Theory of Privacy (RALC)	Tavani and Moor [58], Tavani [56, 57]	Brandimarte et al. [8], Pardo and Siemens [41]
Theory of Contextual Integrity	Nissenbaum [38–40], Barth et al. [7]	Hildén [27], Ifenthaler and Schumacher [28], Tene and Polonetsky [59]
Procedural Fairness Theory/Procedural Justice (adapted to privacy)	Greenberg [20], Lind and Tyler [34], Culnan and Armstrong [10]	Cabinakova et al. [9], Dinev et al. [12, 14], Greenaway et al. [19], Hauff et al. [24], Karwatzki et al. [30], Pollach [44]
Social Contract Theory (adapted to privacy) Privacy Calculus Extended Privacy Calculus Dual Calculus	Milne and Gordon [35], Laufer and Wolfe [32], Culnan and Armstrong [10], Dinev and Hart [13], Li [33]	Awad and Krishnan [6], Cabinakova et al. [9], Dinev et al. [12, 14], Greenaway et al. [19], Kowatsch and Maass [31]
Utility-maximization Theory (adapted to privacy)	Rust [50]	Awad and Krishnan [6], Kowatsch and Maass [31]

2.2 Assessing Theories

In our literature review, we identified authors referring to established theories and concepts from other disciplines such as psychology, sociology and economics. Other authors, mostly engaged in design research, refer to concepts from information systems and computer science. Several authors draw on privacy theories. In the selected papers mentioned in Table 1, authors exploring PDT either use general theories or privacy theories or general theories that have been contextualized and adapted to the privacy sphere. We call a theory a general theory when it is highly abstract and separate from specific application areas. Privacy theories are theories that were developed solely for exploring privacy. None of the authors identified in our literature review has developed a specific theory for PDT or has drawn on a native PDT theory.

We use the following questions to assess the theories:

1. Does the theory address information privacy?
2. Have scholars adapted the theory for privacy research?
3. Does the theory cover aspects that may be relevant for the study of PDT?
4. Which aspects of PDT are or can be considered when using the theory?

Table 2 provides answers to questions 1 to 3.

Table 2. Assessment of theories (Questions 1 to 3)

Theory	1. Information privacy theory?	2. Adaption to privacy research?	3. Aspects of PDT considered?
Agency Theory Signaling Theory	No	Yes [19]	Yes
Theory of Reasoned Action (TRA), Theory of Planned Behavior (TPB)	No	Partly used for studying privacy decision-making	Yes
Technology Acceptance Model	No	Yes, adapted in [9, 31, 66]	Yes [9, 31]
Theory of Bounded Rationality	No	Partly used for studying privacy decision-making	No
Prospect Theory	No	Partly used for studying privacy decision-making	No
Information Boundary Theory (IBT), Communication Privacy Management Theory (CMPT)	Yes	-	Yes [24, 30, 47, 54]
Restricted Access/Limited Control Theory of Privacy (RALC)	Yes	-	No
Theory of Contextual Integrity	Yes	-	Yes
Procedural Fairness Theory/Procedural Justice (adapted to privacy)	Yes	-	Yes [19]
Social Contract Theory (adapted to privacy), Privacy Calculus, Extended Privacy Calculus, Dual Calculus	Yes	-	Yes [6, 14]
Utility-maximization Theory (adapted to privacy)	Yes	-	Yes [6, 14]

For answering question 4, we draw on the following characterization of transparency:

> *"Transparency aims at an adequate level of clarity of the processes in privacy-relevant data processing so that the collection, processing and use of the information can be understood and reconstructed at any time. Further, it is important that all parties involved can comprehend the legal, technical, and organizational conditions setting the scope for this processing. This information has to be available before, during and after the processing takes place. Thus, transparency has to cover not only the actual processing, but also the planned processing (ex-ante transparency) and the time after the processing has taken place to know what exactly happened (ex-post transparency)."* [23]

Based on this characterization, a theory for describing, analysing, explaining or predicting PDT should address at least one of the following questions:

(a) Does the theory address supply of ***information about collection, processing, use or dissemination of personal data***?
(b) Which ***parties involved*** in processing personal data does the theory address?
(c) Is the focus of the theory on the ***process of providing information*** or on ***individual traits of data subjects*** (e.g. intention, decision-making or behaviour)?
(d) Does the theory deal with ***the point in time*** at which information is made available?

We regard the data subject as the producer and owner of the personal data on the one hand and the data controller as the representative for all parties involved in the collection, processing, use and distribution on the other hand. In this context, the data subject is "an identified or identifiable natural person" [46], whose personal data is provided to a data controller as a "natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data" [46]. Other parties involved can be the data processor that processes the personal data on behalf of the controller, recipients of the personal data and third parties. Transparency-enhancing information about data protection measures taken by the controller is also relevant for supervising authorities and consumer protection associations. We have shown this in Fig. 1. In the following text however, we focus on data subjects and data controllers and abstract from supervisory authorities.

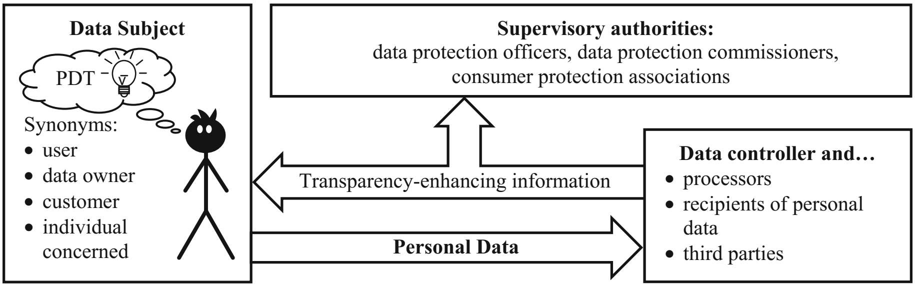


Fig. 1. Parties involved

In the following sections, we discuss which aspects of PDT are considered by the selected theories. A brief characterization of the theories is included. Core concepts are marked in italics. The characters in brackets refer to the questions (a) to (d) mentioned above.

Agency Theory, Information Asymmetry and Signaling Theory

Agency Theory describes *principal-agent relationships* in transactions with *information asymmetries* [15]. **Signaling Theory** addresses options to reduce information asymmetry by *screening* (the principal monitors the agent) or *signaling* (the agent provides information to the principal) [53].

Agency Theory, Information Asymmetry and Signaling Theory are economic theories. These theories can be applied in the data protection/privacy context. Greenaway et al. [19] developed a "Company information privacy orientation framework" based on several theories, including Agency Theory. A lack of PDT for a user as the principal can be considered as an information asymmetry. In this case, screening (e.g. the user as the data subject monitors the company as the data controller with a transparency-enhancing tool) or signaling (the data controller provides information for understanding collection, processing and use of personal data) are opportunities to reduce information asymmetries (a). The parties involved are the data subject as the principal and the data controller as the agent (b). The focus of Agency Theory is on information asymmetry and exchange and not on individual traits of principal or agent. The relationship between principal and agent and the exchange of transparency-enhancing information can be investigated (c). The time of information availability is irrelevant in this context (d).

Theory of Reasoned Action (TRA), Theory of Planned Behavior (TPB), Reasoned Action Approach (RAA)

The **TRA** by Ajzen and Fishbein [4] and the **TPB** by Ajzen and Fishbein [3] are two classical behavioural theories from psychology. They aim to explore the effect of *attitudes* and *subjective norms* on *behaviour intention* and *behaviour*. TPB also considers *perceived behaviour control*. In 2010, Fishbein and Ajzen released a joint theory named the **Reasoned Action Approach** which aims at "predicting and changing social behaviour" [17]. This approach extends TRA and TPB. *Attitude, perceived norm and perceived behaviour control* are influenced by the individual beliefs (*behavioural beliefs, normative beliefs and control beliefs).* These beliefs are based on *background factors*: *individual factors* (e.g. personality and past behaviour), *social factors* (e.g. education, age, gender, culture) and *information factors* (knowledge, media, intervention). Intention and behaviour are moderated by *actual control* (skills, abilities, environment). These theories are an important basis for the development of further, adapted theories, e.g. the Technology Acceptance Model.

TRA, TPB and RAA are general theories focussing on human behaviour and not on PDT. Several papers included in our review use one of these theories or elements thereof [6, 9, 31]. In RAA, actual control moderates individual intention and behaviour. With a theory based on RAA, disclosing personal data can be studied by specifying actual control with transparency-enhancing measures. These measures provide information about collection, processing, use or dissemination of personal data (a). The focus is on the data subject (b) and her/his data disclosure behaviour (c). For actual control, the time when transparency-enhancing measures are available is of particular interest as only information provided before or during disclosure of personal data enables a well-informed decision about disclosing, i.e. actual control (d).

Technology Acceptance Model (TAM)

The **Technology Acceptance Model (TAM)** was developed on the basis of TRA by Davis [11] as an instrument for evaluating the acceptance of information technologies. Users' *perceived usefulness* and *perceived ease of use* determine *behavioural intention to use* and *actual system use*.

TAM was not specifically designed for exploring privacy, but it was adapted for evaluating acceptance of transparency-enhancing tools, e.g. in [9, 31, 66]. Cabinakova et al. [9] base their empirical analysis on TAM, TPB and the Privacy Calculus. They studied how *information about personal data processing* presented by the Google dashboard influences trust in the dashboard and in the dashboard provider, Google (a). Data subject (service user) and data controller (dashboard provider) are addressed by the theory (b). The focus lies on individual behaviour intention (c). The time of information availability is not taken into account (d). Kowatsch and Maass [31] draw on TAM, Utility-maximization Theory and Extended Privacy Calculus for exploring usage intentions and individuals' willingness to provide personal information to Internet of Things services. Survey participants were asked about their expectations on being informed about personal data usage (a). The focus lies on the opinion of potential users of the Internet of Things services (b). The authors studied participants' expectations but they did not explore how transparency affects planned usage of Internet of Things services (c). Participants were asked whether they prefer to be informed every time personal data is used or only the first time (d).

Theory of Bounded Rationality

Unlike Agency Theory, TRA, TPB and RAA, which consider human decisions to be rational, the **Theory of Bounded Rationality** assumes *limited rationality*. *Cognitive limitations, time constraints* and *environmental factors* influence human decisions. Instead of striving to achieve an optimum, individuals try to reach satisfactory levels. The individual uses *heuristics* in decision-making to deal with a complex situation [51].

In the privacy context, the Theory of Bounded Rationality has been used to describe the issue of people not understanding the consequences of personal data disclosure [8, 66] and of information overload as potential inhibitor of disclosing personal data [36]. The theory does not address supply of information about collection, processing, use or dissemination of personal data (a). The focus lies on individual traits of data subjects, i.e. decision-making behaviour (b, c). As the time when information is made available may affect decision-making, this theory is appropriate for describing how and when PDT should be provided in the context of bounded rationality (d).

Prospect Theory

Another behavioural aspect in decision-making is considered in **Prospect Theory**. Prospect Theory explores *decision-making under risk* when an individual selects from *probabilistic alternatives*. The *losses and gains* of this process seem to be more important than the final outcome, leading to a *risk-avoiding behaviour* [29, 60, 61].

Prospect Theory in privacy research can be used to explore the decision-making process of individuals who consider disclosing personal data. PDT, however is not in the focus of the theory (a). In studies of privacy behaviour based on Prospect Theory, "biases in judgements" [60] and the heuristics that data subjects use for decision-making are of interest (b, c). The theory does not explicitly deal with the time when information is made available for transparency reasons (d).

Information Boundary Theory (IBT), Communication Privacy Management Theory (CMPT)

Theoretical contributions to privacy date back to Warren and Brandeis [63] stressing the "right to be left alone". In the **Information Boundary Theory (IBT,** also called Privacy Regulation Theory), Altman [5] discusses *privacy as a dynamic process of boundary regulation* and a "selective control of access to the self or to one's group". Altman states five properties of IBT: *Temporal dynamic process of interpersonal boundary, desired and actual levels of privacy, non-monotonic function of privacy, bi-directional nature of privacy, two levels of privacy (individual and group privacy)* [5].

Petronio [43] integrated these concepts into the **CPMT** and shifted Altman's theory into virtual space. *Private boundaries* separate private and public information. Sharing private information leads to a *collective boundary*, including the individual and the group with which the information was shared. For the individual, it is important to know the communication context for deciding about personal data disclosure. She or he creates a set of rules for the disclosure decision, for example 'I always share my party pictures with my friends, but not with my employer'. The rules are based on five criteria: two *core criteria (cultural and gender)* and *three catalyst criteria (context, motivation, and risk/benefit ratio)* [43].

Applying IBT, Hauff et al. [24] investigate the disposition to value privacy in the context of personalized services. According to IBT, situational factors moderate a person's privacy concerns and risk assessment. "Situation factors represent the degree of personalization and transparency offered to a customer." [24]. The theory does not specifically address supply of transparency-enhancing information. However, the degree of PDT may influence individual information boundaries (a). Hauff et al. [24] explore disclosure behaviour of data subjects (b) as a function of enhanced or reduced PDT in service personalisation (c). The time of providing information to service users is not explicitly considered (d). Other examples of applying components of IBT and CMPT are described by Karwatzki et al. [30], Rader [47] and Stutzmann et al. [54].

Recently proposed privacy theories are more involved with the idea of data protection. The RALC Theory by Tavani and Moor [58] and Nissenbaum's Theory of Contextual Integrity [38–40] are two of them [25].

Restricted Access/Limited Control Theory of Privacy (RALC)

The Restricted Access/Limited Control Theory of Privacy (RALC) by Tavani and Moor [58] seeks to join *limitation* and *control* as two concepts of former privacy theories and to lay a foundation for further privacy theories [25, 58]. Tavani [56, 57] distinguishes between restricted access theories, control theories and restricted access/limited control theories (RALC) of privacy [18]. In the first set of theories, privacy is ensured by *restricting access to personal data*. Control theories place greater emphasis on the individual. They perceive *privacy as control and self-determination of* *data subjects over information about themselves*. The RALC theory combines both approaches. *Restricted access* refers to a sphere that protects individuals from privacy intrusions. *Limited control* refers to management of privacy that enables consumers to grant different levels of access and different usage rights of personal data to different data controllers in different contexts [18, 57].

The theory does not explicitly address the supply of transparency enhancing information (a). The focus is on privacy management of the individual data subject (b). Providing information on personal data processing is a way to support privacy management (c). The time of providing information is not addressed in the theory (d).

Theory of Contextual Integrity

Nissenbaum's **Theory of Contextual Integrity** frames privacy in terms of personal information flows. She calls for not simply restricting the flow of personal information but ensuring that it flows appropriately. She introduces the framework of contextual integrity for determining appropriateness. The framework includes factors determining when people will perceive information technologies and systems as threats to privacy. It helps to predict how people will react to such systems [38–41]. Barth et al. formalized essential elements of contextual integrity in a framework to support technical implementation of data protection requirements [7].

In her paper "A Contextual Approach to Privacy Online" [40], Nissenbaum maintains that the so-called notice-and-consent (or transparency-and-choice) approach has failed. She defines transparency as "conveying information handling practices in ways that are relevant and meaningful to the choices individuals must make" [40]. She claims that in most contexts data subjects are either provided with too little or too much or to detailed information and thus cannot easily make informed decisions (a, b). The focus of the theory is on providing appropriate information on personal information flows. Since the question of what is appropriate also depends on personal characteristics of the data subjects, these are also taken into account. In a broader sense, the appropriateness of transparency-enhancing information for supervising authorities can be considered, too (c). The time at which information is made available is not explicitly addressed in the Theory of Contextual Integrity. However, time is an essential element of appropriate information on personal information flows (d).

Procedural Fairness Theory/Procedural Justice

The **Procedural Fairness Theory**, also known as procedural justice, deals with the perception of individuals *whether a procedure is fair and complies with specified rules*. [20, 34]. The Procedural Fairness Theory was adapted to privacy by Culnan and Armstrong. When a company's privacy practices are considered questionable and customers suspect misuse of their personal data, they feel being treated unfairly and are unwilling to disclose additional personal data [10].

The adaption of Procedural Fairness Theory to privacy is used by Greenaway et al. [19] in their "Company information privacy orientation (CIPO) framework". They also build on Agency Theory and the Privacy Calculus. The authors use two dimensions to distinguish four company information privacy orientations: (1) control "as a way to differentiate how and the extent to which organisations offer their customers the ability to make choices about how their information is collected, used and reused" (p. 584) and (2) procedural justice which "emphasises the extent to which organisations offer transparency to their customers" (p. 584). The second dimension addresses supplying information on personal data processing to customers (a). The parties involved are the costumer as data subject and the company as data controller (b). The CIPO-Framework is a strategy model that focuses on providing information to customers (c). The authors did not examine the point in time when transparency-enhancing information is presented (d).

Social Contract Theory, Privacy Calculus, Extended Privacy Calculus, Dual Calculus, Utility-maximization Theory

The **Social Contract Theory** was adapted to the privacy context by Milne and Gordon [35]. It assumes that disclosing personal data to an organisation can be regarded as a *social exchange* besides the economic exchange. The resulting social contract is considered fair by the data subject if she/he retains *control* over her/his data. The *cost/benefit analysis* a consumer as data subject makes in entering the social contract leads to a *decision about disclosing personal data*. This *calculus of behavior* by Laufer and Wolfe [32] later was called **Privacy Calculus** [10]. Adapted to e-commerce transactions, Dinev and Hart developed the **Extended Privacy Calculus** [13]. Li proposed an integrated framework named **Dual Calculus** based on the Privacy Calculus and taking a *Risk Calculus* into account [33].

Utility-maximization Theory is based on economic exchange theories. Applied to privacy, it assesses how the overall benefit or satisfaction of a person in terms of data protection can be maximised. The *decision to disclose personal data* is a function of the difference between *expected benefits* (e.g. personalised services) and *expected costs* (e.g. privacy losses). Individuals strive for achieving an appropriate optimum [50]. This utility function is usually referred to as Privacy Calculus [33].

Social Contract Theory, Privacy Calculus, Extended Privacy Calculus, Dual Calculus and Utility-maximization Theory are closely connected to each other. Several authors combine two or more of these theories in information privacy research. For this reason, we assess these theories together using the following examples.

Awad and Krishnan [6] use the Utility-maximization Theory and the Privacy Calculus to explore the "relationship between information transparency and consumer willingness to partake in personalization" [6] (a). The authors concentrate on utility functions of data subjects (b). Awad and Krishnan [6] consider providing information and individual traits of data subjects. They focus on the effect of *privacy concerns*, *former privacy invasion experiences* and other factors on the *importance of information transparency* and willingness to be profiled online (c). The time of information availability is not taken into account (d).

Dinev et al. [14] build their research on the Privacy Calculus and Procedural Fairness Theory. They study the effect of "importance of information transparency" (defined as in [6]) and "regulatory expectations" (data protection provisions) on perceived risk. The theoretical framework takes the importance of PDT into account (a), concentrating on the individual's behaviour (b, c) but not on the time of information availability (d).

3 Conclusion

In our study, we have found that 42 out of 157 papers, i.e. only about a quarter, mention a theory. Yet, our literature review has revealed several theories that scholars have used to explore PDT. Some authors base their research not only on one theory, but combine different theories into a new research construct. None of the authors identified in our literature review has developed a specific theory for PDT or has drawn on a native PDT theory. Although PDT has evolved to a considerable research topic within information privacy research, no native PDT theory seems to have emerged yet. Nevertheless, our assessment shows that there are several theories that can substantially support research into PDT. We have identified papers referring to established theories from other disciplines such as psychology, sociology, economics, information systems and computer science, e.g. Agency Theory, Information Asymmetry and Signaling Theory, the Theory of Reasoned Action (TRA), the Theory of Planned Behavior (TPB), the Reasoned Action Approach, the Technology Acceptance Model, the Theory of Bounded Rationality, Prospect Theory, the Procedural Fairness or Procedural Justice Theory, Social Contract Theory and Utility-maximization Theory. Several authors draw on privacy theories as theoretical foundation, e.g. Information Boundary Theory (IBT), Communication Privacy Management Theory (CMPT), the Restricted Access/Limited Control Theory of Privacy (RALC), the Privacy Calculus, the Extended Privacy Calculus and the Dual Calculus or the Theory of Contextual Integrity (RQ1).

From a characterization of PDT we have deduced the following requirements that a theory for exploring PDT should address:

- supply of information about collection, processing, use or dissemination of personal data,
- data subjects and data controllers,
- the process of providing information or individual traits of data subjects, and
- the point in time at which information is made available.

Supply of information about collection, processing, use or dissemination of personal data is addressed by all theories with the exception of the Theory of Bounded Rationality, Prospect Theory and the Restricted Access/Limited Control Theory of Privacy (RALC).

Most theories focus on data subjects only. In the context of PDT, these theories help to explore which forms of PDT result in which disclosure willingness or actual disclosure of personal data. It is noticeable that the vast majority of the theories do not even consider other parties involved. However, an appropriate theory of PDT would take into account not only the data subject but data controllers, data processors, third parties and supervising authorities, since they must at least be involved in providing PDT. Agency Theory, Information Asymmetry, Signaling Theory, the Theory of Contextual Integrity and the “Company information privacy orientation framework” introduced by Greenaway et al. [19] could provide clues for further research in this area.

The process of providing information about collection, processing, use or dissemination of personal data from the data controller to the data subject is addressed by Agency Theory, Information Asymmetry and Signaling Theory, Information Boundary Theory (IBT), Communication Privacy Management Theory (CMPT), the Restricted Access/Limited Control Theory of Privacy (RALC), the Theory of Contextual Integrity, the Procedural Fairness Theory/Procedural Justice and the research approach by Awad and Krishnan [6]. All the other theories focus on individual traits of data subjects.

The point in time at which information is made available by the data controller to data subjects is addressed by only very few theories, namely, the Reasoned Action Approach (RAA), the Theory of Bounded Rationality, and the Theory of Contextual Integrity.

It is also striking that previous research has mainly taken ex-post and ex-ante transparency into account. In this context Adjerid et al. point out "the need to expand the concept of transparency to … making … privacy risks salient and readily available to consumers when they most require them, at the point of disclosure" [2]. Adjerid et al. refer here to an aspect of transparency that we call real-time transparency, i.e. PDT at the time of the decision to disclose personal data. This facet of PDT is probably particularly interesting. However, it has been neglected in previous research and, unfortunately, we have not identified a single theory that could support research in this area (RQ 2).

Only few theories address potential drawbacks of PDT, i.e. the presumption that enhanced PDT may lead to less information privacy. Nissenbaum has explicitly addressed this issue [40] and presented the Theory of Contextual Integrity that may help to further explore this challenge for information privacy research and practice.

Our assessment provides an overview of theories that are used in the context of PDT. However, we do not claim that our study is comprehensive. We have only included papers in our research that explicitly explore PDT and label research foundations with the string "theor*". Our study is based on the assumption that a theory is present when the author of the paper in question uses the term "theory". However, the concept of theory is ambiguous and ambivalent. Therefore, we may have included constructs that are not considered theories in some research disciplines. Furthermore, we have excluded some theories from our study which, in our opinion, do not fit into our research context. Some of these theories, e.g. the Theory of Cryptography, may be relevant for privacy research but not for research into PDT.

Scholars from a wide range of scientific disciplines, e.g. computer science, information systems, privacy, law and media science, have contributed to exploring PDT. Consequently, PDT can most likely not be explored on the basis of a single theory alone. However, research on PDT may build on several theories that open up different perspectives and enable various fields of study.

References

1. Acquisti, A., Adjerid, I., Brandimarte, L.: Gone in 15 seconds: the limits of privacy transparency and control. IEEE Secur. Priv. **11**, 72–74 (2013). https://doi.org/10.1109/MSP.2013.86
2. Adjerid, I., Acquisti, A., Brandimarte, L., Loewenstein, G.: Sleights of privacy. Framing, disclosures, and the limits of transparency. In: Cranor, L.F., Bauer, L., Beznosov, K. (eds.) SOUPS Proceedings, pp. 1–17 (2013). https://doi.org/10.1145/2501604.2501613
3. Ajzen, I.: The theory of planned behavior. Organ. Behav. Hum. Dec. **50**, 179–211 (1991). https://doi.org/10.1016/0749-5978(91)90020-t
4. Ajzen, I., Fishbein, M.: Understanding Attitudes and Predicting Social Behavior. Prentice-Hall, Englewood Cliffs (1980)
5. Altman, I.: The Environment and Social Behavior. Privacy, Personal Space, Territory, Crowding. Brooks-Cole Publishing Co., Monterey (1975)
6. Awad, N.F., Krishnan, M.S.: The personalization privacy paradox: an empirical evaluation of information transparency and the willingness to be profiled online for personalization. MIS Q. **30**, 13–28 (2006). https://doi.org/10.2307/25148715
7. Barth, A., Datta, A., Mitchell, J.C., Nissenbaum, H.: Privacy and contextual integrity: framework and applications. In: IEEE Symposium on Security and Privacy, pp. 184–198 (2006). https://doi.org/10.1109/sp.2006.32
8. Brandimarte, L., Acquisti, A., Loewenstein, G.: Misplaced confidences. Privacy and the control paradox. Soc. Psychol. Pers. Sci. **4**, 340–347 (2013). https://doi.org/10.1177/1948550612455931
9. Cabinakova, J., Zimmermann, C., Müller, G.: An Empirical Analysis of Privacy Dashboard Acceptance: The Google Case. In: ECIS Proceedings (2016)
10. Culnan, M.J., Armstrong, P.K.: Information privacy concerns, procedural fairness, and impersonal trust: an empirical investigation. Organ. Sci. **10**, 104–115 (1999). https://doi.org/10.1287/orsc.10.1.104
11. Davis, F.D.: Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q. **13**, 319 (1989). https://doi.org/10.2307/249008
12. Dinev, T., Bellotto, M., Hart, P., Russo, V., Serra, I., Colautti, C.: Privacy calculus model in e-commerce – a study of Italy and the United States. Eur. J. Inf. Syst. **15**, 389–402 (2006). https://doi.org/10.1057/palgrave.ejis.3000590
13. Dinev, T., Hart, P.: An extended privacy calculus model for E-commerce transactions. Inform. Syst. Res. **17**, 61–80 (2006)
14. Dinev, T., Xu, H., Smith, J.H., Hart, P.: Information privacy and correlates: an empirical attempt to bridge and distinguish privacy-related concepts. Eur. J. Inf. Syst. **22**, 295–316 (2013). https://doi.org/10.1057/ejis.2012.23
15. Eisenhardt, K.M.: Agency theory. An assessment and review. Acad. Manag. Rev. **14**, 57 (1989). https://doi.org/10.2307/258191
16. Fischer-Hübner, S., Hoofnagle, C., Krontiris, I., Rannenberg, K., Waidner, M., Bowden, C.: Online privacy: towards informational self-determination on the internet. In: Hildebrandt, M., O'Hara, K., Waidner, M. (eds.) Digital Enlightenment Yearbook 2013, pp. 123–138. IOS Press, Amsterdam (2013)
17. Fishbein, M., Ajzen, I.: Predicting and Changing Behavior. The Reasoned Action Approach. Psychology Press, New York (2010)
18. Fuchs, C.: Towards an alternative concept of privacy. J. Inf. Commun. Ethics Soc. **9**, 220–237 (2011). https://doi.org/10.1108/14779961111191039

19. Greenaway, K.E., Chan, Y.E., Crossler, R.E.: Company information privacy orientation. A conceptual framework. Inf. Syst. J. **25**, 579–606 (2015). https://doi.org/10.1111/isj.12080
20. Greenberg, J.: A taxonomy of organizational justice theories. Acad. Manag. Rev. **12**, 9–22 (1987). https://doi.org/10.5465/AMR.1987.4306437
21. Gregor, S.: The nature of theory in information systems. MIS Q. **30**, 611–642 (2006). https://doi.org/10.2307/25148742
22. Hansen, M.: Marrying transparency tools with user-controlled identity management. In: Fischer-Hübner, S., Duquenoy, P., Zuccato, A., Martucci, L. (eds.) Privacy and Identity 2007. ITIFIP, vol. 262, pp. 199–220. Springer, Boston, MA (2008). https://doi.org/10.1007/978-0-387-79026-8_14
23. Hansen, M.: Top 10 mistakes in system design from a privacy perspective and privacy protection goals. In: Camenisch, J., Crispo, B., Fischer-Hübner, S., Leenes, R., Russello, G. (eds.) Privacy and Identity 2011. IAICT, vol. 375, pp. 14–31. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-31668-5_2
24. Hauff, S., Dytynko, O., Veit, D.: The influence of privacy dispositions on perceptions of information transparency and personalization preferences. In: Proceedings of the 50th Hawaii Intern. Conference on System Sciences, pp. 5006–5015. AIS Electronic Library (AISeL) (2017). https://doi.org/10.24251/hicss.2017.607
25. Heath, J.: Contemporary privacy theory contributions to learning analytics. JLA **1**, 140–149 (2014). https://doi.org/10.18608/jla.2014.11.8
26. Hildebrandt, M.: The dawn of a critical transparency right for the profiling era. In: Bus, J., Crompton, M., Hildebrandt, M., et al. (eds.) Digital Enlightenment Yearbook 2012. IOS Press, Amsterdam, Washington, D.C. (2012)
27. Hildén, J.: The normative shift: three paradoxes of information privacy. In: Kramp, L., et al. (ed.) Politics, Civil Society and Participation. Media and Communications in a Transforming Environment, pp. 63–73. edition lumière, Bremen (2016)
28. Ifenthaler, D., Schumacher, C.: Student perceptions of privacy principles for learning analytics. ETR&D-Educ. Tech. Res. **64**, 923–938 (2016). https://doi.org/10.1007/s11423-016-9477-y
29. Kahneman, D., Tversky, A.: Prospect theory. An analysis of decision under risk. Econometrica **47**, 263 (1979). https://doi.org/10.2307/1914185
30. Karwatzki, S., Dytynko, O., Trenz, M., Veit, D.: Beyond the personalization–privacy paradox. Privacy valuation, transparency features, and service personalization. J. Manag. Inform. Syst. **34**, 369–400 (2017). https://doi.org/10.1080/07421222.2017.1334467
31. Kowatsch, T., Maass, W.: Critical privacy factors of internet of things services: an empirical investigation with domain experts. In: Rahman, H., Mesquita, A., Ramos, I., Pernici, B. (eds.) MCIS 2012. LNBIP, vol. 129, pp. 200–211. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33244-9_14
32. Laufer, R.S., Wolfe, M.: Privacy as a concept and a social issue. A multidimensional developmental theory. J. Soc. Issues **33**, 22–42 (1977). https://doi.org/10.1111/j.1540-4560.1977.tb01880.x
33. Li, Y.: Theories in online information privacy research. A critical review and an integrated framework. Decis. Support Syst. **54**, 471–481 (2012). https://doi.org/10.1016/j.dss.2012.06.010
34. Lind, E.A., Tyler, T.R.: The Social Psychology of Procedural Justice. Plenum Pr, New York (1988)
35. Milne, G.R., Gordon, M.E.: Direct mail privacy-efficiency trade-offs within an implied social contract framework. J. Public Policy Mark. **12**, 206–215 (1993)

36. Monteleone, S.: Addressing the ‘failure’ of informed consent in online data protection: learning the lessons from behaviour-aware regulation. Syracuse J. Int. Law Commer. **43**, 69–119 (2015)
37. Murmann, P., Fischer-Hübner, S.: Tools for achieving usable ex post transparency: a survey. IEEE Access **5**, 22965–22991 (2017). https://doi.org/10.1109/ACCESS.2017.2765539
38. Nissenbaum, H.: Privacy as contextual integrity. Wash Law Rev. **79**, 119–158 (2004)
39. Nissenbaum, H.: Privacy in Context. Technology, Policy, and the Integrity of Social Life. Stanford Law Books an Imprint of Stanford University Press, Stanford (2010)
40. Nissenbaum, H.: A contextual approach to privacy online. Daedalus **140**, 32–48 (2011). https://doi.org/10.1162/DAED_a_00113
41. Pardo, A., Siemens, G.: Ethical and privacy principles for learning analytics. Br. J. Educ. Technol. **45**, 438–450 (2014). https://doi.org/10.1111/bjet.12152
42. Petronio, S.: Communication boundary management. A theoretical model of managing disclosure of private information between marital couples. Commun. Theory **1**, 311–335 (1991). https://doi.org/10.1111/j.1468-2885.1991.tb00023.x
43. Petronio, S.S.: Boundaries of Privacy. Dialectics of Disclosure. State University of New York Press, Albany (2002)
44. Pollach, I.: Privacy statements as a means of uncertainty reduction in WWW interactions. J. Organ. End User Comput. **18**, 23–48 (2006)
45. Pope, J.A., Lowen, A.M.: Marketing implications of privacy concerns in the US and Canada. Direct Mark.: Int. J. **3**, 301–326 (2009). https://doi.org/10.1108/17505930911000883
46. Publications Office of the European Union: General Data Protection Regulation. 2016/679 (2016)
47. Rader, E.: Awareness of behavioral tracking and information privacy concern in Facebook and Google. In: SOUPS Proceedings, pp. 51–67. USENIX (2014)
48. Roeber, B., Rehse, O., Knorrek, R., Thomsen, B.: Personal data: how context shapes consumers’ data sharing with organizations from various sectors. Electron. Markets **25**, 95–108 (2015). https://doi.org/10.1007/s12525-015-0183-0
49. Rowe, F.: What literature review is not. Diversity, boundaries and recommendations. Eur. J. Inf. Syst. **23**, 241–255 (2014). https://doi.org/10.1057/ejis.2014.7
50. Rust, R.T., Kannan, P.K., Peng, N.: The customer economics of internet privacy. J. Acad. Market. Sci. **30**, 455–464 (2002). https://doi.org/10.1177/009207002236917
51. Simon, H.A.: A behavioral model of rational choice. Q. J. Econ. **69**, 99 (1955). https://doi.org/10.2307/1884852
52. Simon, H.A.: Theories of Decision-Making in Economics and Behavioural Science. Palgrave Macmillan UK (1966)
53. Spence, M.: Job market signaling. Q. J. Econ. **87**, 355 (1973). https://doi.org/10.2307/1882010
54. Stutzman, F., Capra, R., Thompson, J.: Factors mediating disclosure in social network sites. Comput. Hum. Behav. **27**, 590–598 (2011). https://doi.org/10.1016/j.chb.2010.10.017
55. Sutton, R.I., Staw, B.M.: What theory is not. Admin. Sci. Quart. **40**, 371 (1995). https://doi.org/10.2307/2393788
56. Tavani, H.T.: Philosophical theories of privacy. Implications for an adequate online privacy policy. Metaphilosophy **38**, 1–22 (2007). https://doi.org/10.1111/j.1467-9973.2006.00474.x
57. Tavani, H.T.: Informational privacy. Concepts, theories, and controversies. In: Himma, K., Tavani, H.T. (eds.) The Handbook of Information and Computer Ethics, pp. 131–164. Wiley, Hoboken (2008). https://doi.org/10.1002/9780470281819.ch6
58. Tavani, H.T., Moor, J.H.: Privacy protection, control of information, and privacy-enhancing technologies. SIGCAS Comput. Soc. **31**, 6–11 (2001). https://doi.org/10.1145/572277.572278

59. Tene, O., Polonetsky, J.: To track or “do not track”. Advancing transparency and individual control in online behavioral advertising. Minn. JL Sci. Tech. **13**, 281 (2012)
60. Tversky, A., Kahneman, D.: Judgment under uncertainty. Heuristics Biases. Sci. **185**, 1124–1131 (1974). https://doi.org/10.1126/science.185.4157.1124
61. Tversky, A., Kahneman, D.: The framing of decisions and the psychology of choice. Science **211**, 453–458 (1981). https://doi.org/10.1126/science.7455683
62. Walker, K.L.: Surrendering information through the looking glass: transparency, trust, and protection. J. Public Policy Mark. **35**, 144–158 (2016)
63. Warren, S.D., Brandeis, L.D.: The right to privacy. Harvard Law Rev. **4**, 193 (1890). https://doi.org/10.2307/1321160
64. Webster, J., Watson, R.T.: Analyzing the past to prepare for the future. Writing a Literature Review. MIS Q. **26**, xiii–xxiii (2002)
65. World Economic Forum: Rethinking Personal Data: A New Lens for Strengthening Trust. Cologny/Geneva (2014)
66. Zhang, B., Xu, H.: Privacy nudges for mobile applications: effects on the creepiness emotion and privacy attitudes. In: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, pp. 1676–1690. ACM, San Francisco (2016). https://doi.org/10.1145/2818048.2820073

Data Protection by Design for Cross-Border Electronic Identification: Does the eIDAS Interoperability Framework Need to Be Modernised?

Niko Tsakalakis[1](✉), Sophie Stalla-Bourdillon[2], and Kieron O'Hara[1]

[1] Web and Internet Science, ECS, University of Southampton, Southampton, UK
{N.Tsakalakis,kmo}@soton.ac.uk

[2] Institute for Law and the Web, University of Southampton, Southampton, UK
S.Stalla-Bourdillon@soton.ac.uk

Abstract. This paper contributes to the discussion on privacy preservation methods in the context of electronic identification (eID) across borders through interdisciplinary research. In particular, we evaluate how the GDPR principle of 'Data Protection by Design' applies to the processing of personal data undertaken for identification and authentication purposes, suggesting that, in some cases, unlinkable eIDs should be a key requirement in order to facilitate data minimisation and purpose limitation. We argue that in an attempt to welcome diverse types of architectures, the Interoperability Framework could have the effect of reducing the data protection level reached by some national eID schemes, when transacting with services that do not require unique identification. We consequently propose that data minimisation and purpose limitation principles should be facilitated through the implementation of two methods, pseudonymisation and selective disclosure, through an addition to eIDAS' technical specifications.

Keywords: Electronic identification · eIDAS · GDPR · Privacy by Design · Data Protection by Design · Unlinkability · Selective disclosure · Pseudonymisation

1 Introduction

Electronic identification aims at revolutionising the way users interact with online services. In the EU, electronic identification of citizens is at the discretion of the Member States. A handful of Member States have developed national schemes for electronic identification (eID) provision to their citizens, with their architectures varying to a large extend [9][1]. As a result, national systems differ not only in the amount of citizen data they process but also in the level of data protection they offer to these data.

[1] See also country profiles in http://ec.europa.eu/idabc/en/document/6484.html.

E. Kosta et al. (Eds.): Privacy and Identity 2018, IFIP AICT 547, pp. 255–274, 2019.
https://doi.org/10.1007/978-3-030-16744-8_17

Regulation 910/2014 on electronic identification and trust services (hereinafter eIDAS),[2] which came into force on 1 July 2016, enables cross-border interoperability of the diverse national eID schemes. eIDAS aims to create *"a common foundation for secure electronic interaction between citizens, businesses and public authorities"*[3] in order to *"remove existing barriers to the cross-border use of electronic identification means."*[4] Chapter II *"Electronic Identification"* defines the principles required for cross-border eID use across EU Member States by specifying a common denominator in architecture and policies for national schemes to become interoperable. The eID scheme of Germany is the first that has become accessible by all Member States since 29 September 2018.

Meanwhile, the EU's personal data protection framework has been updated by the General Data Protection Regulation (GDPR),[5] which introduced a risk-based approach to data protection and became directly applicable on 25 May 2018. The GDPR aims to facilitate the *"free movement of personal data within the Union"*,[6] in particular in a cross-border context,[7] while ensuring that the data subjects' rights (and in particular their right to the protection of their personal data) are not violated[8].

Article 25 of the GDPR introduces a new requirement of Data Protection by Design. The term is linked to Privacy by Design, a principle stemming from modern privacy engineering. Privacy by Design[9] is advocating for privacy considerations that are embedded in the technology itself, from the design stage throughout the life-cycle of a system [11], rather than imposed only through soft policy measures.[10] The Privacy by Design Resolution, adopted in 2010 by the International Conference of Data Protection and Privacy Commissioners, stresses that Privacy by Design is a *"holistic concept that may be applied to operations throughout an organization, end-to-end"* [1].

Although Privacy by Design is increasingly explored in literature, the effect of the new requirement of the GDPR on design and architectural choices of online services, such as eID provision, remains partially uncertain. This is especially

[2] Regulation (EU) No 910/2014 of 23 July 2014 on electronic identification and trust services for electronic transactions in the internal market and repealing Directive 1999/93/EC [2014] OJ L257/73.

[3] eIDAS Rec. 2.

[4] eIDAS Rec. 12.

[5] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance), OJ L119/1.

[6] GDPR Art. 1(3).

[7] GDPR Rec. 5 *"The economic and social integration resulting from the functioning of the internal market has led to a substantial increase in cross-border flows of personal data"*.

[8] GDPR Art. 1(2).

[9] A term first coined by Ann Cavoukian [10,11] but referring to concepts that started to emerge in privacy literature since the 1970s; see, for example, [12,13,30].

[10] Which are considered less effective, *"an afterthought"* [40].

true since Data Protection by Design befalls data controllers and processors but not system designers, creating therefore inconsistencies as to how the obligation will be translated into system design. Neither eIDAS nor the GDPR offer specific guidance on the means to achieve Data Protection by Design, allowing room for interpretation by the data controllers. However, although eIDAS is technology-neutral,[11] its provisions and accompanying Implementing Acts define a set of requirements for national schemes. Consequently, there is a need to assess the extent and the means by which Data Protection by Design can be effected in the eIDAS Interoperability Framework. This becomes particularly important when considering that, even though eIDAS primarily targets public-sector online services, voluntary use of the eIDAS framework by private-sector services is actively encouraged.[12] In contrast to public-sector services, whose data dissemination practices are often regulated by national legislation, the private sector remains relatively free to decide how to comply with the data protection requirements of the GDPR.

One key question, therefore, is to assess the implications of Data Protection by Design upon the eIDAS Interoperability Framework, and determine whether the Interoperability Framework could be extended to maintain a high level of data protection in cross-border transactions with both public- and private-sector services.

In order to tackle this question, this paper employs an interdisciplinary approach through the use of three different methods: Desk research on Privacy by Design and its application for eID schemes, a synthesised assessment of the general guidance on Data Protection by Design and Data Protection Impact Assessments, and qualitative data collection through a series of interviews with experts in the field of eID. The desk research is used to identify the goals and methods of Data Protection by Design, and in particular how these goals are met in the context of eID. To fully identify its effects in the context of eID, Article 25 of the GDPR should be read in conjunction with Article 35 on Data Protection Impact Assessments, which are meant to provide a process through which the engineering of data protection principles and security measures shall be assessed [37]. Finally, the interviews, which followed a semi-structured format, were used to confirm the findings of the assessment and gave the opportunity to eID experts to express their opinion on Data Protection by Design for eID and the expected impact of eIDAS' Interoperability Framework on participating

[11] eIDAS Rec. 27: *"This Regulation should be technology-neutral. The legal effects it grants should be achievable by any technical means provided that the requirements of this Regulation are met"*.

[12] See eIDAS Rec. 17: *"Member States should encourage the private sector to voluntarily use electronic identification means under a notified scheme for identification purposes when needed for online services or electronic transactions."* See also [44], p. 2: *"the Commission will further promote interoperability actions, including through issuing principles and guidance on eID interoperability at the latest by 2017. The aim will be to encourage online platforms to recognise other eID means – in particular those notified under the eIDAS Regulation (EC) 910/2014 – that offer the same reassurance as their own"*.

schemes. Through thematic analysis, the transcripts established the current practices in eID schemes and the state-of-the-art in regards to Data Protection by Design. We refer to national eID schemes to illustrate how a state-of-the-art system will be impacted by the Interoperability Framework.

The paper is structured as follows: Sect. 2 provides an overview of the Interoperability Framework as defined by eIDAS and its Implementing Acts. We explain the link to the GDPR in Sect. 3 and examine the domain and effect of Data Protection by Design, through seven 'data-protection goals' as proposed by the German Standard Data Protection Model [15]. In Sect. 4 we examine how the Interoperability Framework meets the data protection goals and note that the goal of unlinkability is only partially met. We focus, thus, on unlinkability and analyse what unlinkability entails for eID schemes. We explain how the Interoperability Framework might in certain cases result in constrains on the level of unlinkability that can be supported in cross-border transactions in Sect. 5 and consequently propose a practical way to assure the Interoperability Framework can be extended to support a higher level of unlinkability in Sect. 6. A summary of our findings and concluding remarks can be found in Sect. 7.

2 The eIDAS Interoperability Framework

The cross-border communication of national eID schemes takes place through a set of nodes and related specifications that eIDAS names *'Interoperability Framework'*.[13] Communication between national eID schemes and service providers happens through *'eIDAS nodes'*.[14] eIDAS names the Member State whose notified eID scheme is used as the *'sending Member State'* [17] and the Member State where the service provider resides as the *'receiving Member State'* [17]. Two configurations are supported: The sending Member State can operate an eIDAS node domestically, which will relay authentication requests and assertions between the service providers of the receiving Member State and the national eID scheme (proxy configuration) [17]. Alternatively, the sending Member State provides an instance of their national eID scheme as an eIDAS node which is deployed to each receiving Member State (middleware configuration). The middleware is operated by operators at the receiving Member State [17].

eIDAS defines a set of *'person identification data'*[15] to be transmitted in cross-border identifications. Participating schemes need to satisfy a *'Minimum Dataset'*, which contains four mandatory and four optional attributes.[16] Mandatory attributes are the (a) first and (b) last names of the person, (c) their date of birth and (d) a unique identifier *"as persistent as possible in time."*[17] In addition, the Minimum Dataset may contain (a) the first and last name(s) at

[13] eIDAS Art. 12.

[14] eIDAS Art. 8(3) and [22].

[15] eIDAS Art. 3(3): *"a set of data enabling the identity of a natural or legal person, or a natural person representing a legal person to be established"*.

[16] IR 2015/1501 ANNEX 1.

[17] IR 2015/1501 ANNEX 1(d).

birth, (b) the place of birth, (c) the current address and (d) the gender.[18] The Minimum Dataset is required in every cross-border identification.

eIDAS recognises that eID services have to perform data processing for the needs of electronic identification. Accordingly, Article 5(1) establishes that all processing should be carried out *"in accordance with Directive 95/46/EC"*, which has since been repealed by the GDPR.[19] Consequently the benchmark for data protection compliance under eIDAS is the GDPR. Interestingly, eIDAS seems to have anticipated the GDPR. Article 12(3)(c) of eIDAS mandates that the Interoperability Framework shall *"facilitat*[e] *the implementation of the principle of Privacy by Design;"* and Article 5(2) provides that *"the use of pseudonyms in electronic transactions shall not be prohibited."* In addition, the explanatory recital in the preamble refers to the principle of data minimisation.[20] However, even if eIDAS seems to acknowledge the importance of Data Protection by Design, it is arguable whether the way the Interoperability Framework has been set up can really facilitate the level of data protection guaranteed by some national eID schemes in cases where full identification of a natural person is not necessary.

In order to derive the potential impact of the GDPR on eIDAS it is necessary to analyse the domain and effects of GDPR Article 25. Such an analysis needs to be coupled with an analysis of GDPR Article 35, which offers a process to contextually derive the requirements of Data Protection by Design.

3 Data Protection by Design

Data Protection by Design, under Article 25 of the GDPR, stems from the literature and practice of Privacy by Design approaches in system engineering. Privacy by Design models have extended and refined the protection goals from the field of computer security (confidentiality, integrity and availability, i.e. the 'CIA' model [7,36]), following the model developed by [52] and [28] and which formed the basis of the German Standard Data Protection Model [15]. Four privacy specific goals have been added to the CIA model, to form seven data protection goals: *confidentiality*, *integrity*, *availability*, *transparency*, *intervenability* *unlinkability* and *data minimisation* [6,14,15,28,29,41,52].

Article 25 of the GDPR obliges data controllers to *"implement appropriate technical and organisational measures"* in order to effectively adhere to data protection principles.[21] Data processors are indirectly captured by GDPR Article 25[22] and system producers are *"encouraged* [...] *with due regard to the state*

[18] ibid.

[19] GDPR Art. 94(2): *"References to the repealed Directive shall be construed as references to this Regulation"*.

[20] eIDAS Rec. 11: *"authentication for an online service should concern processing of only those identification data that are adequate, relevant and not excessive to grant access to that service online"*.

[21] GDPR Art. 25(1).

[22] GDPR Art. 28(1): "[data controllers] *shall use only processors providing sufficient guarantees to implement appropriate technical and organisational measures"*.

of the art, to make sure that controllers and processors are able to fulfil their data protection obligations."[23] Of note, eIDAS' requirement to facilitate Privacy by Design could be seen as going further than the GDPR in that it does not expressly target only data controllers. The measures envisioned by Article 25 have to be in place *"both at the time of the determination of the means for processing and at the time of processing itself"*.[24] In other words, technological and policy support for the privacy of data subjects has to be implemented from the design phase and throughout the processing operations. A failure to comply with this requirement might trigger an administrative fine of up to €10.000.000 or in the case of an undertaking, up to 2% of the total worldwide annual turnover of the preceding financial year, whichever is higher.[25] The controller shall justify the selected measures against a list of contextual factors: *"the cost of implementation and the nature, scope, context and purposes of processing as well as the risks* [...] *posed by the processing"*.[26]

The seven data protection goals align with the data protection principles of Article 5 GDPR.[27] An overarching principle, explicitly mentioned in Article 25, is data minimisation. Data minimisation requires the processing (including the collection) of only the data *"limited to what is necessary"*[28] to accomplish a certain purpose. In tandem, under the purpose limitation principle, processing purposes must be specified, explicit and legitimate;[29] in other words purposes should already be defined before data collection. Therefore, not only collection of data must be limited, but collected data must be strictly necessary to a predefined relevant purpose. Confidentiality refers to non-disclosure of certain aspects in an IT system. In a privacy context it can be translated as the need to ensure that information is accessible only by authorised users. Integrity protects the modification, authenticity and correctness of data. It relates, therefore, to safeguards for the accuracy and completeness of the data and their processing methods. Availability concerns the availability, comprehensibility and processability of data. Transparency relates to 'soft' privacy – the relevant policies, reporting and auditing mechanisms in place. Intervenability ensures that parties to the data processing can intervene in the processing when necessary. Finally, unlinkability regards the inability of an attacker to know if any two points of a

[23] GDPR Rec. 78.

[24] GDPR Art. 25(1).

[25] GDPR Art. 83(4)(a).

[26] GDPR Art. 25(1); the qualification will be determined, among others, through a data protection impact assessment.

[27] Confidentiality under GDPR Art. 5(1)(f); integrity under Art. 5(1)(f); availability under Art. 32(b) in relation to Art. 5(1)(f); transparency under Art. 5(1)(a); intervenability under Art. 5(1)(d) and (e) in relation to Arts. 15–22; unlinkability under Art. 5(1)(c) and (e); data minimisation under Art. 5(1)(c).

[28] GDPR Art. 5(1)(c).

[29] GDPR Art. 5(1)(b).

system are related (for example, an eID and its owner).[30] Of note, this definition as explained below is only partial as it focuses upon external actors only. Yet, we argue that in a data protection context, unlinkability should also take into account internal actors.

The data protection goals systematize the obligations put forth by the GDPR, to assist when performing a Data Protection Impact Assessment [15]. Data Protection Impact Assessments are meant as a tool to effect the engineering of data protection principles in a system and, thus, Data Protection by Design. Examining the Interoperability Framework, therefore, in the light of the data protection goals is a useful way to determine the level of Data Protection by Design afforded by eIDAS.

4 The Goal of Unlinkability for eID Schemes

Looking at the Interoperability Framework through the prism of data protection goals, it is clear that those goals have guided the action of the EU legislature. Although most data protection goals have been taken into account by eIDAS and its Implementing Acts,[31] facilitation of the level of unlinkability might be further extended, especially in cases where the service provider is a private-sector entity.

Data minimisation in eIDAS is dealt with through the definition of the Minimum Dataset. The premise is that the Minimum Dataset represents the absolute minimum of attributes necessary to *"uniquely represe*[nt] *a natural or legal person"*.[32] *Confidentiality* is guarding against unauthorised access and disclosure of data. The Implementing Acts define a series of *"implement*[ed] *security controls"*,[33] following a risk-based approach depending on the applicable Level of Assurance, that aim to secure that access and disclosure happens only against authorised actors. The Levels of Assurance ('Low' – 'Substantial' – 'High')[34] also guarantee that technical controls are in place to effect the *integrity* of the claimed identity and its data.[35] *Availability*, which is an explicit goal of eIDAS Article 7(f), is served through legal[36] and technical controls.[37] *Transparency* is addressed by way of published notices and user information about the service providers and the national schemes.[38] Even though eIDAS does not strictly require service providers to display their identity to the users, it allows service

[30] "[Unlinkability] *ensures that a user may make multiple uses of resources or services without others being able to link these uses together* [...] *Unlinkability requires that users and/or subjects are unable to determine whether the same user caused certain specific operations in the system"* [35].

[31] For a detailed analysis of how the Interoperability Framework meets the data protection goals, see [47].

[32] eIDAS Art. 12(4)(d).

[33] Commission Implementing Regulation (EU) 2015/1501 Art. 6(2).

[34] Commission Implementing Regulation (EU) 2015/1502 ANNEX 2.3.1.

[35] ibid, ANNEX 2.4.6.

[36] eIDAS Art. 11(1) and 11(3).

[37] Commission Implementing Regulation (EU) 2015/1502 ANNEX 2.4.4 and 2.4.6.

[38] ibid, ANNEX 2.4.2.

providers to do so if they wish [18]. *Intervenability*, which relates to the user rights about rectification, revocation and erasure of their data, is left to the responsibility of the national eID schemes, since eIDAS is only meant to relay eID data.

Unlinkability appears to be one of the most challenging goals to meet in the context of the Interoperability Framework. Unlinkability aims to serve data minimisation and purpose limitation. In general unlinkability is used to express the impossibility of linking an action performed inside a system (for example, sending a message) to a particular process or agent of the system (in this example, the sender), or the possibility to infer from an outside standpoint that two different sessions in the system (for example two different messages) are performed by the same agent (have, for example, the same originator).[39] However, in a data protection context, unlinkability refers to the risk of linking personal information to its data subject. Therefore, the goal of unlinkability is to eliminate risks of data misuse by minimising risks of profiling [52]. Unlinkability is a key requirement for eID schemes. Indicated in the literature, and confirmed in the expert interviews,[40] a primary goal of *privacy-enhancing* eID schemes is to prevent different pieces of information to be linked together [28,29,49].

The GDPR elevates unlinkability into a performance standard through the data minimisation and purpose limitation principles. Privacy discourse has identified mechanisms for unlinkability, such as data avoidance, separation of contexts through federated distribution, encryption, access control, anonymisation, data destruction etc. [20]. However, the GDPR refrains from providing design standards to realise purpose limitation and data minimisation. Article 25 and its relevant Recital 78 only provide pseudonymisation as an example. In this paper we limit the focus to two specific measures, pseudonymisation and selective disclosure, since both have been identified in electronic identification literature as of particular importance for unlinkability [16,39,41,52]. Pseudonymisation is explicitly mentioned in the GDPR.[41] Based on the definition of pseudonymisation,[42] it must be assumed that a pseudonymised eID dataset can only exist coupled with selective disclosure, i.e. when no other identifying attributes are present in the dataset.[43]

Data minimisation could be seen as having three dimensions: minimisation of content, where the amount of information collected should be the minimum necessary;[44] temporal minimisation, where information should be stored only

[39] See [3] where the authors define the two as "*strong*" and "*weak*" unlinkability.

[40] Excerpts from the interviews are not included in this paper due to space constraints. For a transcript of the experts' opinions, see Sect. 8 and the appendix in [47].

[41] GDPR Art. 4(5).

[42] *"the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person;"* [emphasis given].

[43] For a thorough explanation of this argument, see [48].

[44] GDPR Art. 5(1)(c): *"limited to what is necessary in relation to the purposes"*.

for the minimum amount of time necessary for the specific processing;[45] and minimisation of scope, where data should be used only for the purposes collected.[46] Selective disclosure addresses data minimisation in the strict sense of Article 5(1)(c) – content minimisation. As a means to effect content minimisation, selective disclosure refers to the ability to granularly release information for a specific purpose. Selective-disclosure-capable systems have the ability to accept and transmit only a subset of the available attributes, depending on the processing at hand [38]. An advanced example of selective disclosure can be seen in Fig. 1, where the system only transmits an inferred claim calculated from the user's age instead of transmitting the user's date of birth.

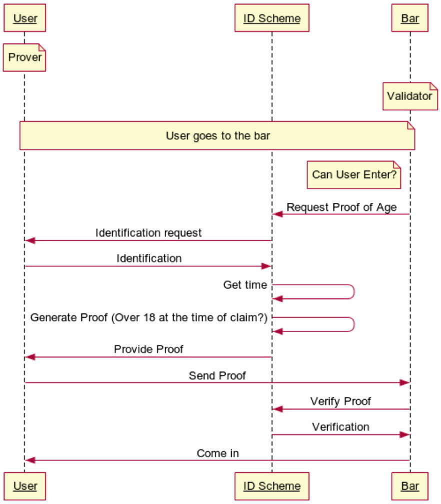


Fig. 1. Simplified verifiable claim using selective disclosure

Pseudonymisation as a means of unlinkability refers to the substitution of direct identifiers with constructed attributes so that the link with the original identifying dataset weakens. There are several degrees of pseudonymisation, with the main impacting factors being the frequency of use of a certain pseudonym and the amount of remaining identifying information in the set. In cases where pseudonyms change across uses ('unidirectional pseudonyms'), linkability between datasets is greatly reduced. On the contrary, where the same pseydonym is deployed regardless of use (an 'omnidirectional' pseudonym) there is a risk of linkability as the pseudonym can act in the form of a *de facto* unique identifier. In eID architectures unidirectional pseudonyms have so far been deployed

[45] GDPR Art. 5(1)(e): *"for no longer than is necessary for the purposes for which the personal data are processed"*.

[46] GDPR Art. 5(1)(b): *"not further processed in a manner that is incompatible with those purposes"*.

in two ways: in 'pairwise persistent' configurations, different pseudonyms are constructed for every pair of pseudonym–user, but remain the same for the specific pair. This way no two service providers receive the same pseudonym, and therefore, service providers cannot easily infer that the pseudonyms refer to the same user. However, since the pseudonym is persistent for that specific user–service pair, it is technically possible for the service to monitor how the pseudonym is used in its system (even if the real identity of the user is not yet known).[47] In contrast, in deployments where the pseudonyms change in between uses even of the same service ('transient pseudonyms') a service provider is not able to distinguish that two uses concern the same user [50]. Although this resolves the issue of linkability it makes it difficult for services to recognise recurring users. For this reason pairwise persistent pseudonyms are preferred in practice[48].

An illustrative example of how unlinkability has been addressed can be given through the case of the German "*neuer Personalauswess*" (*nPA*). The nPA is a federated eID scheme, based around a national eID card that is provided to every citizen. The scheme was built around Privacy by Design principles, supporting advanced privacy controls for the users.[49] The implementation does not depend on an Identity Provider,[50] identification of the user happens through the eID card and a user-controlled middleware software [42].

The nPA incorporates data minimisation through selective disclosure and pseudonymisation. The card has a pre-defined set of attributes stored inside a local RFID chip (refer to Table 1). When performing an electronic identification, the service provider requests the attributes necessary for the identification. The user can then select which of the requested identifiers they wish to disclose to the service provider (selective disclosure). Additionally the nPA employs pairwise-persistent pseudonyms in lieu of an identifier, which are different for every pair of user–service [23,24]. Notably, if a service decided to sub-let their eID infrastructure to other services, all services under the same infrastructure would receive the same pseudonym, therefore increasing the potential to infer the associated identity of the user by combining data. To eliminate this risk, Germany has

[47] This is an issue with 'pairwise persistent' pseudonyms. In a case where two or more services merge together, pairwise persistent pseudonyms can potentially allow linkability depending on the existence of other common identifiers in the dataset.

[48] Privacy-aware eID schemes have started to deploy alternative architectures to sidestep the privacy concerns of pairwise-persistent pseudonyms. See, for example, the implementation of Gov.UK Verify, where a hub in between the Identity and Service Provider mediates all communication in order to obscure the one from the other [27] (cf. though [46] on potential risks); in contrast, the approach taken by the German nPA scheme is to generate pseudonyms locally in the user's eID token.

[49] The basic premise behind the system's design is that the identifying set of information, referred to as a *"sovereign data set"*, has greater value after validation as trustworthy by an official source and therefore deserves greater protection.

[50] Although strictly speaking there is a central Identity Provider operated under the Federal Ministry of the Interior; however its role is to authenticate the service providers, not the users.

put policies in place (soft privacy) that forbid linking of data.[51] The nPA also supports advanced calculations, providing a Yes/No answer about a user's age or location eligibility – without disclosing therefore the user's date of birth or address [24].

Table 1. Minimum data set provided by the German eID scheme [26]

Opt.[a]	eIDAS MDS	German eID
M	Uniqueness identifier	Pseudonym[b]
M	Current family name(s)	Family name
M	Current first name(s)	First name
M	Date of birth	Date of birth
O	First name(s) and family name(s) at birth	Birth name (if present on the eID card)
O	Place of birth	Place of birth
O	Current Address	Address
O	Gender	N/A

[a] *M = Mandatory attribute, O = Optional attribute*
[b] The pseudonym of the German eID scheme is specific to each eID card and each receiving Member State (for public-sector bodies) or each service provider (for private-sector bodies).

On 22 August 2017 Germany pre-notified the nPA under the process of eIDAS Article 9 [25], with the notification published on 26 September 2017.[52] Since the nPA is the first notified scheme, it is an excellent example to highlight potential issues with unlinkability within the eIDAS framework. Of note, the nPA is not the only national scheme to feature unlinkability for privacy protection: the Austrian and the UK's schemes also feature a form of pairwise persistent pseudonymisation, whereas Austria plans to also introduce a form of selective disclosure.[53] The Belgian scheme is also exploring pseudonymisation solutions.[54]

[51] German law is rich in privacy-enhancing principles. At the core is the *'right to information self-determination'* which is a German inception. It confers the right to decide when and within what limits information about one's self should be communicated to others [31]. The right stemmed from a decision of the German Constitutional Court: Volkszählungsurteil 1 BvR 209/83, BVerfGE 65 E 40 1ff. The Court further prohibited any future creation of a persistent unique identifier, ibid s 1. Public authorities operate under a *'separation of informational powers'* – they are not allowed to collate data, as the state should not operate as a single entity, and all data transfers have to be justified against the principles of 'purpose specification' and 'proportionality' [8].

[52] CEF Digital, Overview of pre-notified and notified schemes under eIDAS (2018) https://ec.europa.eu/cefdigital/wiki/display/EIDCOMMUNITY/Overview+of+pre-notified+and+notified+eID+schemes+under+eIDAS.

[53] See for more [47] pp. 48–64.

[54] ibid, Appendix.

Effectively, one third of the schemes currently undergoing a notification procedure for eIDAS is deploying some level of unlinkability.[55] However, this paper will largely refer to the German scheme since it has already undergone the notification process.

5 Unlinkability in the Interoperability Framework

The aspiration of eIDAS to set up a *"technology-neutral"* Interoperability Framework should indicate that advanced privacy designs are supported. This is supported by the explicit mention that the Interoperability Framework will facilitate Privacy-by-Design[56] and eIDAS will not prejudice against the use of pseudonyms.[57] At the very least, it should denote that where national systems support such features, they can be integrated in the Interoperability Framework. However, it appears that the necessity of a common denominator, which is considered essential for transactions with public-sector services, hampers the extend to which Privacy by Design can be used. The main obstacle is the Minimum Dataset and its mandatory attributes[58].

The Minimum Dataset was devised in order to ensure that public-sector service providers, who are obliged to accept other EU Member State's notified eIDs, will have enough information to uniquely identify a foreign citizen. The Minimum Dataset, in other words, is based on the assumption that public-sector services are dependent on successful unique identification of a person in order to provide a service. This is true for a lot of public-sector services eIDAS targets: filing taxes, proving residence status, using student services, opening bank accounts.[59] In addition, some e-government services depend upon a degree of linkability, that the Minimum Dataset provides, in order to satisfy the *'once-only principle'*.[60] However, not all services benefit from a degree of linkability: this is certainly true for providers of the private-sector who rarely require identification in order

[55] For the full list of national schemes undergoing notification, see https://ec.europa.eu/cefdigital/wiki/display/EIDCOMMUNITY/Overview+of+pre-notified+and+notified+eID+schemes+under+eIDAS.

[56] eIDAS Art. 12(3)(c).

[57] eIDAS Art. 5(2).

[58] This is also the position of the ABC4Trust project in [2], which was published before the GDPR, and hence before Data Protection by Design was elevated to a requirement.

[59] The four use cases are indicative examples about the benefits of eIDAS by the eGovernment and Trust team: https://ec.europa.eu/digital-single-market/en/trust-services-and-eid.

[60] An e-government concept that citizens and businesses provide diverse data only once in contact with public administrations, while public administration bodies take actions to internally share and reuse these data. The *'once-only principle'* was one of the targets of the EU's 'eGovernment Action Plan 2016–2018' [21] and the reason behind the EU's 'Single Digital Gateway': http://www.europarl.europa.eu/news/en/headlines/economy/20180911STO13153/single-digital-gateway-a-one-stop-shop-for-all-your-online-paperwork.

to provide a service, such as for example online social platforms, but also for a number of public-sector providers who either operate services where identification is not necessary (i.e. where age verification suffices) or operate sensitive services, like national health services (i.e. drug rehabilitation services). In such cases, linkability could damage the reliability of the provided service, increasing the risk of profiling or data misuse.

Looking back at Germany's notification, the adaptation of the nPA's characteristics in order to conform to eIDAS' requirements already excludes its full pseudonymisation and selective disclosure capabilities. Germany will be deploying the nPA as middleware instances – an instance located at and operated by each receiving Member State. They have also provided a mapping against the attributes required by eIDAS (Table 1). The optional attribute of the gender is not present in the nPA dataset; the optional attribute of name at birth can be provided but only where the attribute has been included in the eID card. All mandatory attributes, nonetheless, are supported. Since eIDAS mandates that the mandatory part of the Minimum Dataset shall in any case be transmitted, and, depending on the receiving service, might be enriched by optional attributes, a user of the nPA will not be able to (de)select attributes for transmission without resulting in an unsuccessful authentication.

In addition, in absence of unique identifiers in Germany[61] the nPA will substitute the mandatory *'Uniqueness identifier'* with a pseudonym. As explained above, the card is capable of producing a persistent unique pseudonym for each pair of user–service, which provides a basic protection against linkability of data between services. However, in cross-border authentications, all of the public-sector services of the receiving Member State will be considered as one service. The receiving Member State will be assigned a pseudonym unique for the pair user–Member State, which will function as the Minimum Dataset's *'uniqueness identifier'* [26].[62] As a result, all public-sector services of the receiving Member State will be receiving the same unique identifier (along with at least the remaining mandatory attributes) thereby raising the question of how linkability of data and uses within a receiving Member State can be prevented. Note that, as abovementioned,[63] under the GDPR in order for a dataset to be considered pseudonymised all other attributes aside from the pseudonyms have to be such that identification of the data subject is not possible. That would be the case, for example, when the only attributes in a dataset are a pseudonym and a date of birth. Seeing as, even when a pseudonym is used in place of a unique identifier, it will always be accompanied by identifying information (the rest of the

[61] See prohibition of the German Constitutional Court above Footnote 51.

[62] The decision might be related to how services in Germany are authorised to access the eID data: services have to file an application with the Federal Office of Administration, listing all the attributes they wish to have access to along with how the attributes relate to the processing purposes [51]. The decision to treat all public-sector services of a Member State as one, and therefore request a combined authorisation, might be in an attempt to make the process easier for the receiving Member State's authorities.

[63] In Footnote 42 and related discussion.

mandatory Minimum Dataset attributes) it is unlikely that an eIDAS dataset will ever meet the definition of GDPR's pseudonymisation.[64] In this sense, there can be no pseudonymisation in eIDAS without selective disclosure. Thus, use of pseudonyms in eIDAS might not be *'prohibited'* per se, but it certainly is restricted.

With selective disclosure and pseudonymisation restricted, *'facilitation'* of Privacy by Design is constrained. In the spirit of the GDPR, the measures afforded by a system should be proportionate to the levels of risk involved in the data processing [4]. Cross-border eID provision should be expected to involve high-risks of processing before any mitigating controls are put in place, in light of the guidance on Data Protection Impact Assessments [5].[65] It can be argued therefore that, by limiting the amount of unlinkability afforded by national systems, service providers that do not require all the attributes of the Minimum Dataset will face problems justifying its processing. Obviously this assertion is contextual. The capabilities of the national scheme providing the electronic identification have a clear impact, as not all systems support selective disclosure and pseudonymisation. However, at least when supported by the national system, the eIDAS Interoperability Framework should be able to support a higher level of unlinkability.

Acting otherwise can prove highly problematic for national schemes that support a high level of unlinkability, as these national schemes will not be able to guarantee such a level for cross-border transactions. In light of eIDAS Article 8, the description of the Levels of Assurance and their governing data protection goals, i.e. integrity, the Member States that offer a high degree of unlinkability would not be in a position to negotiate attributes with service providers that do not require the full Minimum Dataset. As a result, there is an argument that eIDAS Article 12(c) would not be met in the sense that the Interoperability Framework would undermine rather than facilitate Privacy by Design. Going further, national data controllers enabling and operating eID and authentication cross-border would be prevented from offering to their users a high level of data protection in cases where the services requesting eID do not need the complete Minimum Dataset. This could have implications in terms of liability as eIDAS Article 11 should be read in combination with GDPR Articles 82 and 83.

6 Reinforcing the Level of Data Protection by Design in eIDAS

Better incorporation of selective disclosure and pseudonymisation into the Interoperability Framework could reinforce Data Protection by Design in the eIDAS Interoperability Framework. It is true that modifying the Framework to accept different capabilities depending on the features of every national system might

[64] See further analysis in [48].

[65] Among others: processing that affects a significant proportion of the population, using data items in high volumes or on a wide scale, with a significant processing duration and in a large geographical extent.

be impossible, as it would require an upfront insight into the design of all EU systems – whose participation in the Framework is after all voluntary and, hence, not guaranteed. A potential practical way out however would be through an extension of the supported SAML exchanges.[66] Currently the SAML profile specifies that *"at least all attributes defined as mandatory within this minimum data set MUST be requested. At least one minimum data set MUST be requested in each* `<saml2p:AuthnRequest>`*"* [19].[67] The SAML exchanges could be enriched to be able to distinguish and accept requests for a smaller amount of attributes than the ones present in the Minimum Dataset, depending on the requirements of the service provider. The extension would be similar to the proposed scenario in [32]. In this scenario, the service provider would have to specify the required attributes in its request for authentication (see Listing 1 in [32]). The Minimum Dataset would still be sent to the eIDAS node, so as to satisfy the design of systems that do not natively support selective disclosure or pseudonymisation. However, the eIDAS node would then be able to extract only the attributes specified in the request, repackage them into a set under a different pseudonym in place of a unique identifier and transmit them back to the service provider. A similar architecture has been proposed in [43], when the FutureID broker acts in a *'claims transformer mode'*. However, the eIDAS node would not perform the authentication itself (at least when functioning in a proxy mode) but it would simply transform the SAML assertion received by the national eID scheme. Such a functionality is supported, for example, by the eID component (based on [34]) in the FutureTrust project currently under way [33].

If the notified scheme is deployed in a proxy mode [17], and therefore operated by the sending Member State, a solution like that would ensure that no excessive personal data leave the territory of the notified scheme. In cases where the national system is deployed in and operated by the receiving Member State in a middleware configuration, the transmitting Member State has significantly less control over the amount of attributes used. In a middleware configuration it seems likely that the Minimum Dataset will always have to be transmitted to the receiving Member State. However, instead of forwarding the whole Minimum Dataset to the service provider, the eIDAS node could then be able to selectively transmit attributes. The ability to select which attributes to disclose and package them under different pseudonyms would strengthen the level of privacy by reducing the amount of information service providers receive and, effectively, the risk of data collusion. Additionally, selective disclosure at the receiving Member State level would guarantee that in a case of dispute, i.e. in cases of fraud or a law enforcement investigation, the receiving Member State would be able to backtrack the pseudonymisation to identify the affected citizens. This extension

[66] The national systems, the deployed eIDAS nodes and the service providers communicate through defined queries and answers in Security Assertion Markup Language (SAML) [18].

[67] See 6.2 SAML AuthnRequest in [19]. Of note, the equivalent SAML profile of the STORK 2.0 project, which formed the basis of eIDAS, was capable of selective disclosure (see 4.1.4.8.1 in [45]).

of eIDAS constitutes an easy, low cost solution since it requires neither the alteration of eIDAS nor the modification of the architecture. Instead, it can be effected through the issuance of a Regulatory Technical Standard that will provided the added SAML elements to the current eIDAS SAML profile.

7 Conclusion

The risk-based approach of the GDPR in principle allows data controllers to tailor the protection of personal data in their systems as determined by the nature of data processing. The GDPR supports this relative freedom by refraining from specifying an explicit list of appropriate compliance measures. However in practice this might lead to protection that is sub-par to what technology can currently support. Such a case can be observed in relation to eIDAS and its requirement for a Minimum Dataset of mandatory attributes.

Modern electronic identity technology recognises that the amount of information needed for successful authentication varies depending on the service. It also accepts that for better protection of personal data, linkability of datasets should be prevented as far as possible. This paper argues that, on par with the GDPR's risk-based approach, data minimisation should vary subject to the needs of the accessed service and the implemented technical and organisational measures in the Interoperability Framework should provide the same level of data protection guaranteed by the Member States.

The adequate level of data protection should be judged based upon the data-protection goals, which systematise the obligations put forth by the GDPR. eIDAS has been diligent in satisfying most of these protection goals, through its provisions and related Implementing Acts and technical specifications. However, in an effort to define a common denominator for interoperability, the existence of the Minimum Dataset and its unique identifier put constrains into the degree of unlinkability that can be afforded by eIDAS' Interoperability Framework.

This is problematic for participating national schemes that provide a high degree of unlinkability through advanced selective disclosure and pseudonymisation. These schemes will be forced, when participating in eIDAS, to lower the level of protection they provide to their citizens.

This paper proposes that the way the eIDAS nodes operate should be altered so that selective disclosure and pseudonymisation can be possible for the national schemes that support them. Selective disclosure and pseudonymisation, and consequently a greater level of data minimisation, will significantly improve the amount of data that data controllers in electronic identification, residing either in the sending Member State or the receiving Member State, are processing. Thus, such a solution would reduce the associated risks, offering easier ways to demonstrate compliance with the GDPR. We demonstrate how such a solution could be achieved through alterations to the eIDAS SAML profile by way of a Regulatory Technical Standard so that its implementation causes the minimum disruption possible.

Acknowledgement. This research was partly funded by the Research Councils UK Digital Economy Programme, Web Science Doctoral Training Centre, University of Southampton, EP/L016117/1 and partly funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 700542. The authors would like to thank the DG CONNECT, the DG DIGIT and the eIDAS Technical sub-group for their invaluable feedback during the review process of this paper. This paper reflects only the authors' views; the Commission is not responsible for any use that may be made of the information it contains.

References

1. 32nd International Conference of Data Protection and Privacy Commissioners: Resolution on Privacy by Design. Approved in October 2010, Jerusalem, Israel (2010). https://icdppc.org/wp-content/uploads/2015/02/32-Conference-Israel-resolution-on-Privacy-by-Design.pdf
2. ABC4Trust: Privacy-ABCs and the eID Regulation. Position paper, ABC4Trust (2014). https://abc4trust.eu/download/documents/ABC4Trust-eID-Regulation.pdf
3. Arapinis, M., Chothia, T., Ritter, E., Ryan, M.: Analysing unlinkability and anonymity using the applied Pi calculus. In: 23rd IEEE Computer Security Foundations Symposium, pp. 107–121, July 2010. https://doi.org/10.1109/CSF.2010.15
4. Article 29 Data Protection Working Party: statement on the role of a risk-based approach in data protection legal frameworks. WP 218, 30 May 2014
5. Article 29 Data Protection Working Party: Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679. WP 248 rev 0.1, 4 April 2017
6. Bieker, F., Friedewald, M., Hansen, M., Obersteller, H., Rost, M.: A process for data protection impact assessment under the european general data protection regulation. In: Schiffner, S., Serna, J., Ikonomou, D., Rannenberg, K. (eds.) APF 2016. LNCS, vol. 9857, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-44760-5_2
7. Bishop, M.: Introduction to Computer Security. Addison-Wesley Professional, Boston (2004)
8. Burkert, H.: Balancing informational power by informational power or Rereading Montesquieu in the internet age. In: Brousseau, E., Marzouki, M., Méadel, C. (eds.) Governance, Regulation and Powers on the Internet, Book Section 4, pp. 93–111. Cambridge University Press, Cambridge (2012)
9. Castro, D.: Explaining international leadership: electronic identification systems. Technical report, ITIF (2011). http://www.itif.org/files/2011-e-id-report.pdf
10. Cavoukian, A.: 7 Laws of identity: the case for privacy-embedded laws of identity in the digital age. Information and Privacy Commissioner of Ontario (2006). http://www.ontla.on.ca/library/repository/mon/15000/267376.pdf
11. Cavoukian, A.: Privacy by design: the 7 foundational principles. Information and Privacy Commissioner of Ontario (2009). https://www.ipc.on.ca/wp-content/uploads/resources/7foundationalprinciples.pdf
12. Chaum, D., Fiat, A., Naor, M.: Untraceable electronic cash. In: Goldwasser, S. (ed.) CRYPTO 1988. LNCS, vol. 403, pp. 319–327. Springer, New York (1990). https://doi.org/10.1007/0-387-34799-2_25

13. Chaum, D.L.: Untraceable electronic mail, return addresses, and digital pseudonyms. Commun. ACM **24**(2), 84–90 (1981). https://doi.org/10.1145/358549.358563
14. CNIL: Privacy Impact Assessment (PIA): Methodology (How to Carry out a PIA). Commission Nationale de l'Informatique et des Libertés (2015). https://www.cnil.fr/sites/default/files/typo/document/CNIL-PIA-1-Methodology.pdf
15. Conference of the Independent Data Protection Authorities of the Bund and the Länder: the standard data protection model. V.1.0 - Trial version (2017). https://www.datenschutzzentrum.de/uploads/sdm/SDM-Methodology_V1.0.pdf
16. Dhamija, R., Dusseault, L.: The seven flaws of identity management: Usability and security challenges. IEEE Secur. Priv. **6**(2), 24–29 (2008). https://doi.org/10.1109/msp.2008.49
17. eIDAS Technical Sub-group: eIDAS - Interoperability Architecture (2015). https://joinup.ec.europa.eu/sites/default/files/document/2015-11/eidas_interoperability_architecture_v1.00.pdf
18. eIDAS Technical Sub-group: eIDAS SAML Attribute Profile, 20 June 2015. https://joinup.ec.europa.eu/sites/default/files/eidas_saml_attribute_profile_v1.0_2.pdf
19. eIDAS Technical Sub-group: eIDAS Message Format. v. 11.2 (2016). https://ec.europa.eu/cefdigital/wiki/download/attachments/46992719/eIDAS%20Message%20Format_v1.1-2.pdf?version=1&modificationDate=1497252919575&api=v2
20. ENISA: Privacy and Data Protection by Design (2015). https://www.enisa.europa.eu/activities/identity-and-trust/library/deliverables/privacy-and-data-protection-by-design
21. European Commission: Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: EU eGovernment Action Plan 2016–2020 - Accelerating the digital transformation of government. COM(2016) 179 final, Brussels, 19 May 2016. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52016DC0179
22. European Commission: eIDAS-Node Integration Package Service Offering Description (2018). https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eIDAS+Node+integration+package?preview=/46992716/59191417/CEF_eID_eIDAS-Node_Integration_Package_Service_Offering_Description.pdf
23. Federal Office for Information Security [BSI]: Innovations for an eID Architecture in Germany (2011). http://www.personalausweisportal.de/SharedDocs/Downloads/EN/Flyers-and-Brochures/Broschuere_BSI_innovations_eID_architecture.html?nn=6852820
24. Federal Office for Information Security [BSI]: Technical Guideline TR-03127: Architecture electronic Identity Card and electronic Resident Permit (2011). https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/Publications/TechGuidelines/TR03127/BSI-TR-03127_en.pdf
25. Federal Office for Information Security [BSI]: eIDAS Notification of the German eID, February 2017. https://www.bsi.bund.de/EN/Topics/ElectrIDDocuments/German-eID/eIDAS/notification/eIDAS_notification_node.html

26. Federal Office for Information Security [BSI]: German eID based on Extended Access Control v2: Overview of the German eID system. version 1.0, 20 February 2017. https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/EIDAS/German_eID_Whitepaper.pdf?__blob=publicationFile&v=7
27. Government Digital Service: GOV.UK Verify Technical Guide: Architecture Overview, October 2014. https://alphagov.github.io/rp-onboarding-tech-docs/pages/arch/arch.html
28. Hansen, M.: Marrying transparency tools with user-controlled identity management. In: Fischer-Hübner, S., Duquenoy, P., Zuccato, A., Martucci, L. (eds.) Privacy and Identity 2007. ITIFIP, vol. 262, pp. 199–220. Springer, Boston, MA (2008). https://doi.org/10.1007/978-0-387-79026-8_14
29. Hansen, M., Jensen, M., Rost, M.: Protection goals for privacy engineering. In: 2015 IEEE Security and Privacy Workshops, pp. 159–166. IEEE, San Jose (2015). https://doi.org/10.1109/SPW.2015.13
30. Hes, R., Borking, J. (eds.): Privacy-Enhancing Technologies: The Path to Anonymity, Revised edn. Registratiekamer, The Hague (2000)
31. Hornung, G., Schnabel, C.: Data protection in Germany I: the population census decision and the right to informational self-determination. Comput. Law Secur. Rev. **25**(1), 84–88 (2009). https://doi.org/10.1016/j.clsr.2008.11.002. http://www.sciencedirect.com/science/article/pii/S0267364908001660
32. Horsch, M., Tuengerthal, M., Wich, T.: SAML privacy-enhancing profile. In: Hühnlein, D., Roßnagel, H. (eds.) P237 - Open Identity Summit 2014, pp. 11–22. Gesellschaft für Informatik e.V, Bonn (2014)
33. Hühnlein, D., et al.: Futuretrust - future trust services for trustworthy global transactions. In: Hühnlein, D., Roßnagel, H., Schunck, C.H., Talamo, M. (eds.) P264 - Open Identity Summit 2016, pp. 27–41. Gesellschaft für Informatik eV, Bonn (2016)
34. Hühnlein, D., et al.: SkIDentity - Trusted Identities for the Cloud (2015). https://www.skidentity.de/fileadmin/Ecsec-files/pub/7_SkIDentity-final.pdf
35. ISO/IEC 15408–1:2009: Information technology - security techniques - evaluation criteria for it security - part 1: Introduction and general model, International Organization for Standardization, Geneva, CH (2009)
36. ISO/IEC 27002:2013: Information technology - security techniques - code of practice for information security controls, International Organization for Standardization, Geneva, CH (2013)
37. ISO/IEC 29134:2017: Information technology - security techniques - guidelines for privacy impact assessment, International Organization for Standardization, Geneva, CH (2017)
38. Khatchatourov, A., Laurent, M., Levallois-Barth, C.: Privacy in digital identity systems: models, assessment, and user adoption. In: Tambouris, E., et al. (eds.) EGOV 2015. LNCS, vol. 9248, pp. 273–290. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-22479-4_21
39. Koning, M., Korenhof, P., Alpár, G.: The ABC of ABC - an analysis of attribute-based credentials in the light of data protection, privacy and identity. In: Balcells, J. (ed.) Internet, Law & Politics : A Decade of Transformations. Proceedings of the 10th International Conference on Internet, Law & Politics, Universitat Oberta de Catalunya, Barcelona, 3–4 July, pp. 357–374. Huygens Editorial, Barcelona (2014). http://edcp.uoc.edu/proceedings_idp2014.pdf
40. Le Métayer, D.: Privacy by design: formal framework for the analysis of architectural choices. In: Proceedings of the Third ACM Conference on Data and Application Security and Privacy (CODASPY), San Antonio (2013)

41. Pfitzmann, A., Hansen, M.: Anonymity, Unlinkability, Unobservability, Pseudonymity and Identity Management - A Consolidated Proposal for Terminology. Version v0.34, 10 August 2010. https://www.kantarainitiative.org/confluence/download/attachments/45059055/terminology+for+talking+about+privacy.pdf
42. Poller, A., Waldmann, U., Vowe, S., Turpe, S.: Electronic identity cards for user authentication - promise and practice. IEEE Secur. Priv. **10**(1), 46–54 (2012). https://doi.org/10.1109/MSP.2011.148
43. Roßnagel, H., et al.: FutureID - shaping the future of electronic identity. In: Annual Privacy Forum 2012, Limassol, Cyprus, 10–11 October 2012
44. Servida, A.: Principles and guidance on eID interoperability for online platforms. Revised draft version of January 2018. https://ec.europa.eu/futurium/en/system/files/ged/draft_principles_eid_interoperability_and_guidance_for_online_platforms_1.pdf
45. STORK: D4.11 final version of technical specifications for the cross-border interface (2015). https://www.eid-stork2.eu/index.php?option=com&view=file&id=64:d411-final-version-of-technical-specifications-for-the-cross-border-interface&Itemid=174
46. Tsakalakis, N., O'Hara, K., Stalla-Bourdillon, S.: Identity assurance in the UK: technical implementation and legal implications under the eIDAS regulation. In: Proceedings of the 8th ACM Conference on Web Science. WebSci '16, pp. 55–65. ACM, New York (2016). https://doi.org/10.1145/2908131.2908152
47. Tsakalakis, N., Stalla-Bourdillon, S.: Documentation of the legal foundations of trust and trustworthiness. FutureTrust deliverable D2.8 v. 1.00, 29 June 2018. https://docs.wixstatic.com/ugd/2844e6_b441a5f255f94cf78a7d4c890e2fe6aa.pdf
48. Tsakalakis, N., Stalla-Bourdillon, S., O'hara, K.: What's in a name: the conflicting views of pseudonymisation under eIDAS and the general data protection regulation. In: Hühnlein, D., Roßnagel, H., Schunck, C.H., Talamo, M. (eds.) P264 - Open Identity Summit 2016, pp. 167–174. Gesellschaft für Informatik e.V., Bonn (2016)
49. Veeningen, M., de Weger, B., Zannone, N.: Data minimisation in communication protocols: a formal analysis framework and application to identity management. Int. J. Inf. Secur. **13**(6), 529–569 (2014). https://doi.org/10.1007/s10207-014-0235-z
50. Yee, G.O.M.: Privacy Protection Measures and Technologies in Business Organizations: Aspects and Standards. IGI Publishing, Hershey (2011)
51. Zwingelberg, H.: Necessary processing of personal data: the need-to-know principle and processing data from the new German identity card. In: Fischer-Hübner, S., Duquenoy, P., Hansen, M., Leenes, R., Zhang, G. (eds.) Privacy and Identity 2010. IAICT, vol. 352, pp. 151–163. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-20769-3_13
52. Zwingelberg, H., Hansen, M.: Privacy protection goals and their implications for eID systems. In: Camenisch, J., Crispo, B., Fischer-Hübner, S., Leenes, R., Russello, G. (eds.) Privacy and Identity 2011. IAICT, vol. 375, pp. 245–260. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-31668-5_19

Risk Profiling by Law Enforcement Agencies in the Big Data Era: Is There a Need for Transparency?

Sascha van Schendel(✉)

Tilburg Institute for Law, Technology, and Society (TILT), Tilburg University,
Tilburg, The Netherlands
s.vanschendel@tilburguniversity.edu

Abstract. This paper looks at the use of risk profiles by law enforcement in the age of Big Data. First, the paper discusses different use-types of risk profiling. Subsequently, the paper deals with the following three categories of challenges of risk profiling: (a) false positives (and to some extent false negatives) as well as incorrect data and erroneous analysis, (b) discrimination and stigmatization, (c) and maintaining appropriate procedural safeguards. Based on the hypothesis of risk profiling creating challenges, this paper addresses the question whether we need transparency of risk profiling by law enforcement actors, from the perspective of protecting fundamental rights of those affected by the use of risk profiles. The paper explores tackling these challenges from the angle of transparency, introducing Heald's varieties of transparency as a theoretical model.

Keywords: Risk profiling · Transparency · Law enforcement · Procedural safeguards · False positives · Discrimination · Data protection · Criminal law · Explanation

1 Introduction

Risk assessment has become very popular in all sectors of society, including in the prevention against crime. Over the last years, the term 'Big Data' has taken flight and has increasingly received much attention in government policies and practices [1]. The use of Big Data analysis is in part the reason for a strong emphasis on preventing and minimizing risk in society. Having the tools to analyze huge volumes of data and extract information from them, possibly completely by automated means, facilitates processes such as the creation and analysis of risk profiles [2]. The use of profiles grows as they can be constructed and applied more easily, while at the same time the construction, analysis and application of the profiles become more complicated and opaque.

The use of risk profiles to find suspects or determine if someone poses a risk to society has traditionally been an important tool to national law enforcement agencies to efficiently make use of their powers. Some scholars have described the emphasis on risk in criminal justice as entering into an era of actuarial justice [3–5] in which we focus on analyzing risk in a mathematical way, the rise of 'the logic of risk' [6], or 'the new paradigm of criminal law' [7]. While there are arguments to make in favor of law

Published by Springer Nature Switzerland AG 2019
E. Kosta et al. (Eds.): Privacy and Identity 2018, IFIP AICT 547, pp. 275–289, 2019.
https://doi.org/10.1007/978-3-030-16744-8_18

enforcement agencies making their practices more efficient by using risk profiles [8], this development is not without its issues and raises issues towards those affected by this practice. This leads to the first hypothesis of the paper: risk profiling in the Big Data era creates challenges. Based on this hypothesis of risk profiling creating challenges, this paper addresses the question whether we need transparency of risk profiling by law enforcement actors, from the perspective of protecting fundamental rights of those affected by the use of risk profiles. This research question also contains the second hypothesis of this paper, namely that transparency could be an interesting angle to tackle the challenges. The aim of this paper is to shed light on the challenges of risk profiling. Exploring whether transparency is a way to approach these challenges is intended as a starting point of a discussion. This paper does not provide an analysis of how transparency will solve the challenges of risk profiling, nor does the author outline what transparency should look like in this context. This is the topic of future research of the author.

Section 2 of this paper briefly maps risk profiling by law enforcement actors in practice. One specific example is taken as a case study to be explored in more detail. This example is SyRI (System Risk Indication), a Dutch risk profiling system. This example was chosen as it is currently under review in a national court case. SyRI is a good example of risk profiling that presents the starting point of a criminal investigation. Some parallels are drawn to the USA in Sect. 2, as some of the types of use of risk profiles are still very minimal in the European Union but might become more prominent following the USA's example. Section 3 describes the main challenges of the use of risk profiling, grouping them under three main, non-exhaustive headers: errors, discrimination and stigmatization, and lack of procedural safeguards or outdated safeguards. Section 4 describes why transparency might be an interesting angle to approach the challenges. For this purpose, Sect. 4 introduces and briefly describes Heald's 'varieties of transparency' [9] as a theoretical model. Subsequently, Sect. 4 narrows transparency down to foster further discussions, as transparency in itself is a very broad concept. For this purpose a bottom-up approach to the issues is chosen, focusing on explanations as a means of transparency. The focus on explanations is all the more relevant after the introduction of the General Data Protection Regulation [10] ('GDPR'), as it contains references to explanations in the context of automated decision making. Section 4 will therefore also briefly mention transparency and explanations under the GDPR and the Law Enforcement Directive [11] ('LED').

2 Risk Profiling in Practice

2.1 What Is Risk Profiling?

Risk profiling, for the purpose of this paper, is categorizing or ranking individuals or groups, sometimes including automated decision making, using correlations and probabilities drawn from combined and/or aggregated data, to determine the level of risk that is posed to the security of others or national security by those individuals or groups.[1] The most prominent type of risk here is the likelihood of an individual or group (re)committing crime.

Risk profiling can take many forms in the law enforcement context. Risk profiling can be used in concrete criminal investigations where there is already an identified suspect or perpetrator and a profile is applied to this person. A first instance is to make decisions about which police powers to employ. Brkan gives the example of automated decision making to determine whether to seize a mobile device [12].

Risk profiling of an identified individual can also be targeted towards future behavior. This can be risk profiling to determine whether someone is allowed bail or probation specifically whether that person is at risk of reoffending, or risk profiling in sentencing determining the duration of incarceration. The most famous example is from the USA, namely COMPAS. COMPAS is an algorithm used by judges and probation- and parole officers to assess a criminal defendant's likelihood of reoffending [13].

There are types of risk profiling where the target is a location. These are often types of predictive policing drawing from various sources of data, ranging from non-personal data such as the distance to the highway to different forms of personal data pertaining to inhabitants of that area such as the history of criminal records. Algorithms can in this way pinpoint the level of risk for areas, so that police officers can be deployed accordingly. This type of risk profiling is very popular in the USA, but also exists in Europe [14]. Such as in the Netherlands, where the Crime Anticipation System is used, creating a grid that is updated every 14 days which shows for each square what crime is likely to take place and on which time of day. This system was at first only applied in the capital, Amsterdam, but is now being used in various other cities. While such a system is targeted at the risk level of a location, it indirectly profiles the residents of that area. This is where discussions on stigmatization and self-fulfilling prophecies come in: by attaching a risk label to a certain area and sending police patrols there accordingly, this can impact the view residents and outsiders have of this area plus lead to an increase in crime detection further increasing patrols and measures taken against residents of this area. Indirectly the residents are also profiled as high risk. Of course this means that there is an assumption that the suspects or perpetrators would reside in this area, while this does not have to be reality.

Besides the above described type where law enforcement applies profiles to an already identified individual or area, risk profiles are also used to detect individuals-or groups-that fit the profile. In these cases an algorithm finds individuals that fit the risk profile in a haystack of data. These individuals are likely to commit a crime or are likely to have committed an undetected crime. This type of profiling does not take place within the boundaries of a specific criminal investigation but rather leads to the starting point of one. Risk profiling to detect individuals can take the form of 'heatlists', similar to the heatmapping or area profiling described above. An example from the USA is the system Intrado Beware, which is a mobile, cloud-based application, sold to the police, that gathers contextual information from social media, commercial data and criminal data, creating a risk score–green, yellow, red-for individuals [14]. Intrado Beware is slightly different from the standard model of detecting people who have committed a

[1] This definition of risk profiling is the author's own and is a working definition.

crime, as it is more targeted towards providing police information about the person they are about to encounter and identifying whether they are a risk in the sense of posing a risk to the security of the police officer. Another example of finding individuals that match the risk profile comes from the Netherlands, which is described in the section below.

2.2 The SyRI Case

An example of risk profiling can be found in the Netherlands in the SyRI ('System Risk Indication') program. SyRI was officially launched in 2014 and is employed by the Dutch Ministry of Social Welfare & Employment. It is a system in which many databases are combined–ranging from tax data and data about social benefits to data about integrating in Dutch society and education-, creating a large data pool to detect fraud [15]. SyRI targets three types of fraud: unlawful use of social benefits, taxation fraud, and fraud with labor laws [15]. Due to the broad scope and large governmental database, almost every citizen of the Netherlands is present in the database. Using a predetermined risk model, the system searches for correlations in the database flagging a potential case of fraud based on the model used for that specific search [16]. The individual is given a risk indication, which is forwarded to the Dutch National Police and/or prosecuting office, who then decide whether to investigate further. The risk indication is stored in a register which relevant public bodies can access [15]. So even though SyRI is not a specific risk profiling program of law enforcement solely, law enforcement is one of the parties that can be included in a cooperation to use SyRI and the risk score of SyRI can be the data point that starts a criminal investigation.

Even though SyRI has been used for a couple of years now, its use has not been without resistance. There have been parliamentary debates centered on the question whether SyRI met proportionality demands and whether its legal basis was not too broad. The program raises issues of transparency, mainly awareness and contestability. Most citizens are not aware that their data is in this system nor that they might be flagged. Most people are confronted with the existence of the system when they receive an administrative fine or encounter another negative consequence. Besides possible privacy and data protection issues that follow from a system that uses so much data, there are serious issues with possibilities to contest the system and correct errors. In March 2017, several NGOs and two citizens took up the initiative to launch a court case, which is still ongoing, to test whether SyRI is compliant with EU data protection legislation, the fundamental right to privacy and the right to fair trial under article 6 of the European Convention on Human Rights [17]. One of the points that is debated is the secrecy of the risk models, but also the lawfulness of the automated decision making and the broadness of the legal basis [17]. In this sense the problematic aspects of SyRI illustrate the challenges following from data driven policing or policing in the Big Data era, such as risk profiling.

3 Risk Profiling: Challenges

This section groups the challenges of risk profiling under three main headings: errors, discrimination and stigmatization, and lack of procedural safeguards or outdated safeguards. This is a non-exhaustive list but aims to give an oversight of the main challenges based on literature about profiling, algorithms, predictive analysis and data analysis in the law enforcement domain.

3.1 Errors: Relying on Statistics and Probabilities

Most profiles are probabilistic, describing the chance that a certain correlation will occur [18]. In most cases the individuals included under the profile do not share all the attributes or characteristics of the group profile [18]. This is especially true for non-distributive profiles, which are framed in terms of probabilities and averages, comparing members within a group or category, or comparing those groups or categories to each other [19]. This means that there is always an inherent risk of errors in the use of profiles, as it might include people erroneously within a profile or might miss certain individuals, leaving them out of scope. The first category is false positives, the second situation is false negatives [20]. In case of false positives, people would be incorrectly classified in a group or profile. This in turn could have consequences for decisions taken to the disadvantage of these persons, or they could be erroneously subjected to police powers. In the case of a false negative, we encounter the more traditional problem of law enforcement, namely overlooking someone who should be a suspect or miscalculating the risk of recidivism. Especially in the context of terrorism threats, risk profiles aim at minimizing false negatives, as the societal consequences are a lot graver when allowing for a false negative than a false positive [21]. Mittelstadt et al. talk about these issues in terms of ‘inconclusive evidence’, meaning that algorithms often draw from statistics and in doing so create only probable outcomes that are focused more on actionable insights than causal relations [22]. Algorithms become increasingly complex and autonomous, which makes it harder for law enforcement to be transparent about why they receive a certain outcome. Mittelstadt et al. refer to this complexity and opaqueness as ‘inscrutable evidence’, where humans have trouble interpreting which data points lead to the conclusion [22]. Risk profiling in the Big Data era relies heavily on algorithms and statistics. Statistics offer insight into numbers, for example how many people re-offend within an amount of years. Algorithms can be used to combine statistics, mine them for patterns, and make a prediction about an individual’s behavior by applying this information to their situation. This does not mean however that this person acts according to the statistics nor that the conclusion based on combining statistics is right. If the process becomes more complex and opaque it can become harder for law enforcement agencies to demonstrate why they received this outcome.

3.2 Discrimination and Stigmatization

The trend of risk management combined with the strong focus in politics on terrorism prevention can push law enforcement to target specific groups, especially with the pressure to fully use technologies such as algorithms and Big Data analysis. The technology, to a large extent, takes over tasks that were not fully automated before. Now algorithms take over the task of detecting the patterns, creating the profiles and finding correlations [7]. As these technologies are not foolproof–just as police officers' instincts and human observation and logic are not foolproof-this does pose a threat of discrimination and stigmatization of certain groups. The technology might 'over target' specific groups. It has been shown already that risk-based policing targets certain societal groups within different EU countries, such as North African youths, soccer supporters, Roma, and Muslims [21]. The technology might increase racial or ethnical profiling especially. For example, in the Netherlands, the existence and possible condoning of ethnic profiling by police officers has been a topic of societal debate for years [23]. While these types of debates were mainly targeted at racial profiling based on 'police instinct', automated profiling possibly increases racial profiling [21, 22]. As Van Brakel explains: "*Predictive mapping can potentially lead to ethnic profiling. If arrest rates are a measure for predicting in which areas most crime occurs, for instance, and if it is clear that arrest rates are disproportionately higher in particular population groups as a result of ethnic profiling there is a clear bias in the prediction, and the mapping can lead to even more ethnic profiling*" [14]. Referring back to the example of the Dutch predictive policing application CAS, ethnic profiling has already been demonstrated to be an issue [23]. When using automated means, all the data analysis is scaled up, increasing the scale of the problematic aspects. Profiling in itself is a discriminatory process, which is not illegal in itself, but can become illegal discrimination if based on factors such as race or religion [24]. Article 11 of the Law Enforcement Directive prohibits the use of sensitive data–officially called 'special categories of data'[2]- unless suitable safeguards are in place to protect the interests of the data subject. So when using sensitive data such as ethnicity or religion extra safeguards might need to be put into place. However, provisions that forbid the use of these types of factors or require extra safeguards, do not prevent the use of proxies. The use of proxies could nonetheless be discriminatory, such as using zipcodes or income as a proxy for ethnicity. Discrimination following directly from automated decision making is forbidden as far as the special categories of data go. Profiles that focus on other characteristics–or proxies for those characteristics-, such as age can also be deemed illegal. Recently, a court in the Netherlands ruled that the use of a risk profile–of single men of 55 years or older– was in violation of the right not to be discriminated against [25]. It is extremely hard, however, to tackle illegal discriminatory profiling if the impacted individuals are not aware that they are placed in a certain profile. As Leese states: "*as datadriven profiles produce artificial and non-representational categories rather than actual real-life social groups, the individual is likely to not even notice when he or she becomes part of a 'risky' category*" [21]. Besides individuals not being aware, the actors operating the algorithm might also be unaware of illegal discrimination happening in their dataset or algorithm, or they might be unaware that their use

[2] Special categories of data under the GDPR and LED are data that are deemed especially sensitive and therefore receive more protection. The set categories are: data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation.

of proxies has the same result as the illegal discrimination based on certain characteristics. These problems are only more difficult to detect and address as systems get more complex.

3.3 Procedural Safeguards

Automated risk profiling works in a different way than the more traditional policing, creating challenges in the way in which safeguards are set up. First, risk profiling is a form of proactive or even preventive policing. This forms a contrast to the more traditional reactive policing. Koops has referred to a shift in paradigm in criminal law which for example contains a focus on prevention, risk, groups, profiling, and statistics [7]. An issue is that safeguards might be linked to the prosecution phase, leaving out the opaque pre-investigation practices where a lot of data is already analyzed [7]. In reactive policing the focus for checks and balances is traditionally on the judge, who comes in at the later investigation stages or only at the trial. However with risk profiling someone might be arrested erroneously and released shortly after. Similarly, with risk profiling used in general policing, a lot of data is analyzed and privacy infringements could take place there as a consequence, but go undetected because there is no criminal investigation of a specific suspect yet. Second, in the information society decisions are increasingly made based on group profiles [20]. In literature on data protection and privacy there are increasingly more debates on the possibilities for collective procedures to address types of data processing such as Big Data analytics and group profiling [26, 27]. Vedder, in his work on KDD (Knowledge Discovery in databases), already signaled a tendency of treating people based on group characteristics [19]. This tendency has only grown with modern risk profiling, as risk profiling requires statistics and categorizing or ranking of people. Vedder discusses data that for example used to be personal data but during time has become part of a broader set of anonymous data, at some stage the data became part of aggregate data and individual identifiers were replaced with group identifiers [19]. As Vedder precisely states, using generalizations and categorizations based on profiles can be highly problematic when they are used as a basis for policy and people are treated as a member of a group instead of on their own merits [19]. However, safeguards and rights are often linked to individual decision making. Automated decision making under article 11 of the Law Enforcement Directive, which produces an adverse legal effect concerning the data subject or significantly affects him or her, is prohibited unless authorized by national law and provided with appropriate safeguards. Profiling is concerned with creating a set of correlations on the aggregate level and subsequently applying it to individuals or groups. One could argue that only the application of a profile to an individual situation is regulated here. Brkan gives the example of a group being the target of profiling by making an automated decision to patrol certain areas, affecting the lives of the people who live in such an area [12]. Again, reference could be made to the Dutch predictive policing system, CAS, indicating where and when which crimes are likely to take place. Based on those risk indications police officers are deployed, but it is not clear whether the decision to target areas as high risk areas meets the criteria of article 11 of the Law Enforcement Directive to require further safeguards.

4 Transparency

4.1 Using Transparency to Address Challenges

Having presented the most prominent challenges of risk profiling by law enforcement agencies, the issue is how to address these challenges. I propose to look at the concept of transparency for solutions to these challenges.

Transparency has possibilities to expose flaws or give insight into decision making. A lot of the challenges relate to processes being opaque. For example, maybe it is not visible that someone is placed in the wrong category or that there is illegal discrimination taking place. Or, because of a lack of procedural safeguards in the early investigation, mistakes do not come to light. In that sense transparency also increases possibilities of awareness. A lot of people are simply not aware that they are being profiled or that a decision about them, for example concerning arrest or deploying investigative measures, is based on a risk profile. A lack of awareness makes it difficult for those affected by risk profiling to check for compliance with their rights when necessary, such as the right to fair trial, equality of arms, privacy, or the principle of non-discrimination. Therefore transparency might be interesting to look further into. However, transparency is a very broad concept and has different meanings even within one discipline. Several authors have already described the relation between transparency and a concept that is often connected to it, namely ‘openness’. For example Birkinshaw proposes that transparency and openness are close in meaning but are both broader than merely access to (government) information [28]. Larsson also does not consider openness and transparency to be the same concept, as according to Larsson transparency goes beyond openness and also includes simplicity and comprehensibility [29]. Heald remarks that transparency has become ‘the contemporary term of choice’ for describing an openness of public actors about actions and decisions they make [9]. However, Heald makes various distinctions within the concept of transparency [9]. These distinctions are helpful to dismantle the broad concept and distinguish which functions or solutions transparency actually offers. Therefore Heald’s work on transparency is briefly discussed here as a theoretical framework.

First, Heald makes explicit different directions of transparency. There are two directions of vertical transparency: upwards and downwards. Upwards transparency can be seen in hierarchical terms of allowing the superior to observe behavior or results. Downward transparency can be seen in terms of democracy, allowing the ruled to observe behavior or results of their rulers [9, p. 27]. Second, Heald discusses the two directions of horizontal transparency: outwards and inwards. Transparency outwards occurs when the hierarchical agent can observe behavior outside of its organization or institution, so as to understand the domain it is operating in and observe the behaviour of peers. Transparency inwards occurs when those outside of the organization can observe what is happening within the organization [9, p. 28].

Next Heald distinguishes different varieties of transparency in general using three dichotomies: event transparency versus process transparency; transparency in retrospect versus transparency in real-time; nominal transparency versus effective transparency. When distinguishing between events and processes, an event can for example be the input or output data. When providing process transparency one can be transparent about the procedural factors–which rules are followed- or operational aspects–how are the rules applied in this situation- [9, pp. 29–32]. Another dichotomy is the temporal one, so one can allow for transparency after the fact–in retrospect- or one can continuously allow for transparency so that transparency takes place in real-time [9, pp. 32–33]. For the last dichotomy Heald states that there can be a gap between nominal and effective transparency, which he labels the 'transparency illusion' [9, p. 34]. Allowing for transparency does not always mean that it is effective: "*For transparency to be effective, there must be receptors capable of processing, digesting, and using the information*" [9, p. 35]. Also, transparency is not effective when it creates an information overload [9, p. 35].

After having some more insight into the concept of transparency, it is interesting to see how this theory relates to the problem at hand. First, concerning the vertical transparency: in the context of data processing by law enforcement actors, upwards transparency is concerned with transparency towards oversight authorities such as Data Protection Authorities or (investigatory) judges. Downwards transparency is directed towards the people that are the subject of the process, in the case of automated decision making this concerns for example the data subjects. When looking at horizontal transparency, inwards transparency can be offered to oversight authorities, the people affected by the data processing, the democracy or people at large, and so forth. In the context of law enforcement outwards transparency is not so relevant. When distinguishing between events and processes it becomes clear that in the case of risk profiling there is a large variety in what transparency could be given about. Transparency can for example concern events such as the input of new data, or the outcome that the algorithm gives. On the other hand transparency could be given about the process, such as procedural aspects like the decision rules, which in this case could be the algorithm itself. Concerning the process, transparency could also be provided about the operational aspects, focusing on a specific situation, explaining why the decision rules have in this case led to this outcome. With regard to the temporal dimension transparency could be given in retrospect, for example notifying oversight authorities or individuals that a decision has been made based on a risk profile. Or transparency could be offered in real-time, which in the case of law enforcement seems complicated, as this might pose difficulties for ongoing investigations. With regards to the dichotomy between nominal and effective transparency, a lot of issues are left open. To determine the effectiveness of transparency of risk profiling would be quite difficult.

Based on the description above, there is still a lot of variation possible in to whom transparency is offered, about what elements of risk profiling transparency is given, and what constitutes effective transparency. Transparency in risk profiling could have varying functions. However, going back to the research question of this paper, -to determine whether transparency could help with the challenges from the perspective of protecting the rights of those affected by the risk profiling-transparency needs to be narrowed down further along Heald's varieties of transparency. In focusing on those affected by risk profiling, downwards-inwards transparency is the relevant variety. When targeting transparency towards data subjects, and others that might be affected, three steps could be distinguished. The first step is to make data subjects aware that data processing and risk profiling is taking place. The second step is to explain to data subjects what is going on and how certain decisions are made. These two steps enable the third step, being able to contest profiles and automated decisions and receive due process. Perceiving transparency in this bottom-up way makes it easier to grasp the overall concept of transparency and connects to the challenges. It is after all important in safeguarding the rights of individuals affected that they are not erroneously profiled, illegally discriminated against, or undergoing a process without enough procedural safeguards to protect fundamental rights such as the right to a fair trial.

Alternatively, it is interesting to assess in the context of upwards transparency how law enforcement actors will explain their profiling practices and decisions to judges, or other competent authorities, when the analysis becomes more intricate and decisions more data driven. This is, however, a dimension of transparency that will largely take place behind closed doors and very difficult to analyze as researchers.

4.2 Food for Thought: Explanations as a Means of Transparency?

One aspect or means of offering transparency to data subjects, and others affected by risk profiling, is that of providing explanations of the profiling. Explanations of profiling and automated decision making have become very relevant with the reform of EU data protection legislation. To go further into this, a brief description of EU data protection legislation in the context of transparency is needed.

EU data protection legislation consist of several pieces of law. In 2016 the reform package for Data Protection legislation on the European Union level was adopted, introducing the General Data Protection Regulation [10] ('GDPR') and the Law Enforcement Directive [11] ('LED'). Before the introduction of the LED, data protection in this area was left in part to national legislation, partly standardized by Convention 108 of the Council of Europe [30], and in part regulated by a variety of specialist and sector specific instruments, creating a very fragmented landscape [31]. The LED repeals the Council Framework Decision 2008/977/JHA [32], which was very narrow in scope, only applying to cross-border transfers and exchanges of personal data, excluding domestic processing of personal data [33]. As the regulation of the processing of personal data by national law enforcement agencies has been left out of harmonization so far, a wide margin is left to the criminal procedural law of Member States to lay down requirements and safeguards. For data processing in the private sector it is logical to look for requirements and safeguards in the GDPR, but for data processing by national law enforcement agencies the LED needs to be seen together with the safeguards and requirements following from Member States' legislation that arranges the competencies of these actors. The current Law Enforcement Directive does not contain a general principle of transparent processing. Under relevant Council of Europe law this is different. The newest version of Convention 108, which also applies to the law enforcement domain, does contain a principle of transparent data processing under article 5.[3] When comparing the GDPR and the LED, a fundamentally different approach with regard to transparency becomes visible. Transparency takes a

[3] The Convention 108 has recently been modernized. The amending Protocol (CETS No. 223) to Convention 108 was adopted by the Committee of Ministers of the Council of Europe on 18 May 2018.

predominant place within the GDPR in the form of transparent processing[4], combined with various rights towards the data subject. Transparent processing in this sense can mean that data subjects are informed about processing before it takes place, during the processing itself and upon request of the data subject. Besides the principle of transparent processing that applies throughout all types of processing, articles 12 until 14 of the GDPR impose obligations on the side of data controllers as well as rights upon data subjects to request information. In the context of profiling especially article 13 is relevant where, in the context of providing information, it states: "(…)*the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject*." Recital 39 of the GDPR also pays specific attention to transparency, it states: "*The principle of transparency requires that any information and communication relating to the processing of those personal data be easily accessible and easy to understand, and that clear and plain language be used*". Thus this principle already implies; first, that information about the processing should be available to the data subject; second, that this information should be easy to access; third, the information itself should be easily understandable. The same aspects of transparency are highlighted in recital 58. Recital 60 underlines the importance of awareness as a component of transparency, by stating that transparency requires the data subject being informed of the *existence* of the processing. In contrast, in the LED the principle of transparent processing is not present. The only relevant reference to transparency is in recital 26. Recital 26 merely mentions that processing should be done in a transparent manner with regard to the persons concerned, while at the same time acknowledging the necessity of some covert operations and surveillance measures. The critical component is "*provided by law and constitute a necessary and proportionate measure in a democratic society with due regard for the legitimate interests of the natural person concerned"*. So the requirements for transparency and limits of opaqueness are determined on a case to case basis where practices differ much from country to country. Again, national criminal procedural law also has a role to play here, as regulation is left to the Member States in this area. Some countries might have more provisions on transparency than others.

Transparency in the context of data processing has also become a much debated topic in literature. On the one hand arguments are presented in favor of more transparency of algorithms and algorithmic decision making [34], on the other hand there is a continuously increasing awareness that 'transparency' as such is not an all-encompassing answer for issues with algorithms [22, 35]. Transparency more often than not, requires balancing of transparency as a value and other values such as protecting trade secrets, national security, and privacy of others [21, 22]. Especially in literature on the law enforcement sector, transparency is discussed in the context of a trade-off or balance between security and transparency–sometimes as an aspect of the right to privacy-[12].

[4] Under article 5 of the GDPR.

However, that there are arguments in favor of law enforcement agencies operating under a certain level of secrecy, does not mean that there is no room for transparency at all. Requiring law enforcement to explain why someone is profiled in a certain way puts up a safeguard in general against illegal discrimination and errors, as requiring an explanation stimulates checking the analysis to see whether the proper data was used and how the data was weighted to come to this result, as well as the level of probability. Providing explanations also serves a more specific purpose. While the process of profiling becomes more automated and technically complicated, it is important that law enforcement actors can still understand how a profile or decision came about, putting up a safeguard against algorithms that become so opaque and complex that humans cannot understand or justify the outcomes anymore. Law enforcement agencies need to be able to explain their decisions to a judge that checks the legality of, for example, searching a phone or computer; public prosecution needs to be able to explain during a trial why the prosecution authorities started the investigation, meaning why the person in case was suspect according to the risk profiling system. This requirement is inherent in criminal justice systems [36], if law enforcement cannot explain a decision, the judge will probably not accept it. However, giving these sort of explanations might be more challenging in automated processes, or processes with minimal human intervention. Therefore, it would be good to lay down an explicit requirement in national law, whether in data protection legislation or criminal law, for explaining profiling and automated decision making [36]. While the actors using a risk profiling system need to maintain a certain understanding of how it works so that they can be accountable for their decisions, it is equally important that the human actors involved do not over rely on the technology. In literature this has been discussed as ‘automation bias’, meaning that humans have a tendency to over rely on the accuracy of automated analysis and decisions, the result is assumed to be correct and no counterfactual evidence is sought out [37]. With this risk in mind, explaining the decision also ensures that human actors do not take the outcome for granted but investigate how it came about.

As stated in the introduction, this paper is not the place to develop what explanations of risk profiling in the law enforcement sector should or could look like exactly. It does offer food for thought though, especially with all the new transparency provisions under the GDPR.

5 Conclusion

Preventive and risk based policing is increasingly becoming the new form of policing. However, safeguards might be attuned to more traditional, less data-driven, policing and criminal procedures. This means that now and in the future there will be challenges on this front, such as dealing with probabilities, discrimination, effects on groups and shifting to the pre-investigation phase. This paper proposed to look at transparency for dealing with these challenges. Making the broad notion of transparency more feasible to grasp using Heald’s varieties of transparency, it becomes clear that there are a lot of different options regarding to whom transparency could be offered and what the object of this transparency would be. Basing decisions on these risk profiles can have very serious consequences from the perspective of those affected by the risk profiles, for example when it comes to their rights not to be discriminated against or the right to fair trial and equality of arms in being subjected to complex data analysis that might even concern future behavior. While transparency is prominent in the GDPR and in literature concerning the GDPR, the debate about transparency is not really taking place yet in the literature about the LED and literature about profiling in the law enforcement sector. The increasing use of Big Data and algorithms in policing and in prosecution, such as in the form of profiling and automated decision making will, however, only make transparency more important to discuss. This paper made a first step in discussing why transparency is important, by examining the challenges of risk profiling and the different options of transparency, and why we especially need explanations in the law enforcement domain as a means of transparency. The time has now come to also talk about explanations of profiling in the law enforcement sector to assess what role they could play and what they could look like.

References

1. van der Sloot, B., van Schendel, S.: International and Comparative Study on Big Data, Working Paper no. 20, Dutch Scientific Council for Government Policy (WRR) (2016)
2. Marks, A., Bowling, B., Keenan, C.: Automatic justice? Technology, crime and social control. In: Brownsword, R., Scotford, E., Yeung, K. (eds.) The Oxford Handbook of the Law and Regulation of Technology. OUP (2017)
3. Kemshall, H.: Understanding Risk in Criminal Justice. Crime and Justice Series. Open University Press, London (2003)
4. Reichman, N.: Managing crime risk: towards an insurance based model of social control. Res. Law Soc. Control **8**, 151–172 (1986)
5. Harcourt, B.E.: Against Prediction Profiling, Policing, and Punishing in an Actuarial Age. The University of Chicago Press (2007)
6. Ericson, R.V., Haggerty, E.: Policing the Risk Society. Clarendon Press, Oxford (1997)
7. Koops, E.J.: Technology and the crime society: rethinking legal protection. Law Innov. Technol. **1**, 93–124 (2009)
8. Zouave, E.T., Marquenie, T.: An inconvenient truth: algorithmic transparency & accountability in criminal intelligence profiling. In: 2017 European Intelligence and Security Informatics Conference (2017)
9. Heald, D.: Varieties of transparency. In: Hood, C., Heald, D. (eds.) Transparency: The Key to Better Governance? OUP/British Academy (Proceedings of the British Academy) (2006)
10. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) L 119/1
11. Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA, L 119/89
12. Brkan, M.: Do algorithms rule the world? Algorithmic decision-making in the framework of the GDPR and beyond, 1 August 2017. SSRN: https://ssrn.com/abstract=3124901
13. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

14. van Brakel, R.: Pre-emptive big data surveillance and its (dis)empowering consequences: the case of predictive policing. In: van der Sloot, B., et al. (eds.) Exploring the Boundaries of Big Data. Amsterdam University Press, Amsterdam (2016)
15. Besluit SUWI: Staatsblad, 320 (2014). https://zoek.officielebekendmakingen.nl/stb-2014-320.html
16. https://algorithmwatch.org/en/high-risk-citizens/
17. https://pilpnjcm.nl/en/dossiers/profiling-and-syri/
18. Hildebrandt, M.: Defining profiling: a new type of knowledge? In: Hildebrandt, M., Gutwirth, S. (eds.) Profiling the European Citizen, pp. 17–45. Springer, Dordrecht (2008). https://doi.org/10.1007/978-1-4020-6914-7_2
19. Vedder, A.: KDD: the challenge to individualism. Ethics Inf. Technol. **1**, 275–281 (1999)
20. Hildebrandt, M., Koops, E.J.: The challenges of ambient law and legal protection in the profiling era. Modern Law Rev. **73**(3), 428–460 (2010)
21. Leese, M.: The new profiling: algorithms, black boxes, and the failure of anti-discriminatory safeguards in the European Union'. Secur. Dialogue **45**(5), 494–511 (2014)
22. Mittelstadt, B.D., et al.: The ethics of algorithms: mapping the debate. Big Data Soc. **3**, 1–21 (2016)
23. van der Leun, J.P., van der Woude, M.A.H.: Ethnic profiling in The Netherlands? a reflection on expanding preventive powers, ethnic profiling and a changing social and political context. Policing Soc. **21**(4), 444–455 (2013). www.opensocietyfoundations.org/sites/default/files/equalityunder-pressure-the-impact-of-ethnic-profiling-netherlands-20131128_1.pdf. Open Society Initiative (2013) Equality under Pressure: The Impact of Ethnic Profiling
24. Schermer, B.: The limits of privacy in automated profiling and data mining. Comput. Law Secur. Rev. **27**, 45–52 (2011)
25. Centrale Raad van Beroep: 21 November 2017, ECLI:NL:CRVB:2017:4068
26. Taylor, L., Floridi, L., van der Sloot, B. (eds.): Group Privacy: New Challenges of Data Technologies. Springer, Heidelberg (2017)
27. Mantelero, A.: Personal data for decisional purposes in the age of analytics: from an individual to a collective dimension of data protection. Comput. Law Secur. Rev. **32**(2), 238–255 (2016)
28. Birkinshaw, P.J.: Freedom of information and openness: fundamental human rights. Adm. Law Rev. **58**(1), 177–218 (2006)
29. Larsson, T.: How open can a government be? The swedish experience'. In: Deckmyn, V., Thomson, I. (eds.) Openness and Transparency in the European Union. European Institute of Public Administration, Maastricht (1998)
30. Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (ETS No. 108, 28.01.1981)
31. De Hert, P., Papakonstantinou, V.: The police and criminal justice data protection directive: comment and analysis. Comput. Law Mag. SCL 2012 **22**(6), 1–5 (2012)
32. Council Framework Decision 2008/977/JHA of 27 November 2008 on the protection of personal data processed in the framework of police and judicial cooperation in criminal matters, L 350/60
33. Marquenie, T.: The police and criminal justice authorities directive: data protection standards and impact on the legal framework. Comput. Law Secur. Rev. **33**, 324–340 (2017)
34. Zarsky, T.: Transparent predictions. Univ. Ill. Law Rev. **4**, 1503 (2013)
35. Annany, M., Crawford, K.: Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability. New Med. Soc. **20**(3), 973–989 (2018)

36. Commissie modernisering opsporingsonderzoek in het digitale tijdperk, Regulering van opsporingsbevoegdheden in een digitale omgeving, June 2018. https://www.rijksoverheid.nl/documenten/rapporten/2018/06/26/rapport-commissie-koops—regulering-van-opsporingsbevoegdheden-in-een-digitale-omgeving. This Committee, that reviewed Dutch criminal law in the light of digital developments, also concluded that an explicit requirement for explaining automated data analysis is necessary in national criminal procedural law

37. Cummings, M.L.: Automation bias in intelligent time critical decision support systems. In: AIAA 3rd Intelligent Systems Conference (2004)

Author Index