UNIVERSITÀ
DI TRENTO
Department of
Information Engineering and Computer Science

Doctoral School in
Information and Communication Technology

Privacy in the small

Francesco Ciclosi

Advisor
Prof. Fabio Massacci
Università di Trento

Co-Advisor
Prof. Giovanna Paola Varni
Università degli Studi di Trento

Reviewer
Prof. Antonio Barili
Università di Pavia

Reviewer
Prof. Giampaolo Bella
Università di Catania

January 2025

Abstract

The advent of the General Data Protection Regulation (GDPR) and analogous global data protection laws has profoundly influenced organizations' socio-technical structures, mandating compliance with stringent personal data processing standards. These laws compel entities to critically understand their socio-technical systems, encompassing the complex interplay between legal, managerial, and technical components.

This thesis addresses the challenge of ensuring compliance through empirical methodologies and field studies, enhancing both theoretical understanding and practical application. Chapter 2 explores the multifaceted role of Data Protection Officers (DPOs) as mediators between compliance auditors and organizational management. It highlights the tension inherent in their dual role and the socio-technical risks DPOs navigate in diverse operational contexts. Chapter 3 focuses on user understanding of privacy policies across linguistic boundaries, proposing a methodology for creating cross-language comparable corpora. Using English and Italian privacy policies, it showcases how language and cultural adaptations influence user comprehension of technical terms and offers a replicable approach for cross-language research. Chapter 4 extends this work by refining tools for analyzing cross-language privacy policies. By mapping technical terms and assessing their frequency and relevance, it identifies the limitations of automated methods and underscores the importance of manual intervention for nuanced cross-lingual analyses. Chapter 5 examines GDPR implementation in resource-constrained settings, such as schools, revealing gaps between theoretical compliance and practical execution. A risk-based approach is proposed, advocating feasible and continuously improvable data protection practices over rigid adherence to legal stipulations. The conclusions (Chapter 6) summarizes the findings for cross-language privacy research and practical insights for improving compliance in socio-technical systems.

Keywords

Field Study, GDPR, Privacy procedures, Privacy violations, Usable privacy.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The introduction of the Regulation (EU) 2016/679 (known as the GDPR) [55] in the EU and other analogous data protection laws adopted in other countries heavily impact institutions and organizations that process personal data. These organizations must comply with relevant data protection laws when determining the purposes and means of processing personal data (acting as a controller) or processing personal data on the controller's (acting as a processor) behalf. These structures are a heterogeneous set of autonomous components, from employees to users, from suppliers to regulators. In any operations on personal data (processing), they act as a socio-technical system. Hence, an organization must know its socio-technical system deeply to comply with data protection law. Such socio-technical systems depend on human factors (user's knowledge, beliefs, and subjective perception). The user's knowledge may be of language, the specific reference domain, and the specialized technical and legal terminology of the reference domain. The user's beliefs pertain to specific cultural, social, philosophical, political, and religious convictions. At the same time, multiple factors influence the user's subjective perceptions [159] [169], such as his or her experience, level of schooling, and cognitive abilities. This knowledge is achievable through fieldwork with an empirical activity of gathering sources of evidence (from documents to interviews with relevant personnel), which is the purpose of this dissertation.

1.1 Historical Overview

Past studies focused on analyzing socio-technical systems (STS) but mainly on the general security aspects without paying attention to the specificity of data protection. Other studies have formally described a privacy-oriented STS's internal components

and relationships. However, they have yet to conduct a field study to investigate the challenges and violations that small entities encounter daily in their processing processes. For example, Dalpiaz et al. [46] proposed STS, a method for designing secure complex systems that also provides a formally defined language named STS-ml and handling security by considering the system as a whole as described in Paja et al. [115]. Later, Robol et al. [145] proposed a method to support the design of GDPR [55] compliant systems composed of both a modeling language and a reasoning framework. This method is based on the socio-technical approach proposed by Dalpiaz et al. [46] and focuses on designing a secure system without specializing in privacy. Giorgini et al. [65] proposed another socio-technical approach to security, in which a system is considered a composition of autonomous actors, each with its own goals. This approach handles security aspects but does not focus on privacy.

There are currently several tools [116], [45] to support entities in designing their STS. However, they focus on designing systems rather than auditing and describing them, and their profitable use requires a minimum level of subject knowledge that is not accessible to everyone, much less to educational institution staff.

Other tools and methodologies aim to support entities in structuring their organizational structure (and consequently their STS) to comply with the relevant laws and rules. These tools can simplify the construction of a path aimed at implementing cybersecurity and data protection in an organization and facilitate continuous monitoring and improvement activities. Frameworks for cybersecurity and data protection (e.g., the CINI's "*Framework Nazionale per la Cybersecurity e la Data Protection*" [105]) are some of these. Unfortunately, they are either projected toward medium- and large-sized entities or (in the case where they address organizations of any size) are too complex to apply autonomously in smaller entities that lack the necessary skills to use them. Hence, the poor attention these frameworks give to smaller institutions is a problem when a small entity tries to use them because they do not consider small entities' peculiarities and the lack of resources.

Key observation: Most studies approach the security aspects of a socio-technical system without focusing on privacy. There are cybersecurity and data protection frameworks, but they do not fit well in small entities.

In Figure 1.1, there is a visual representation of the positioning of research questions according to Guest et al. [69].

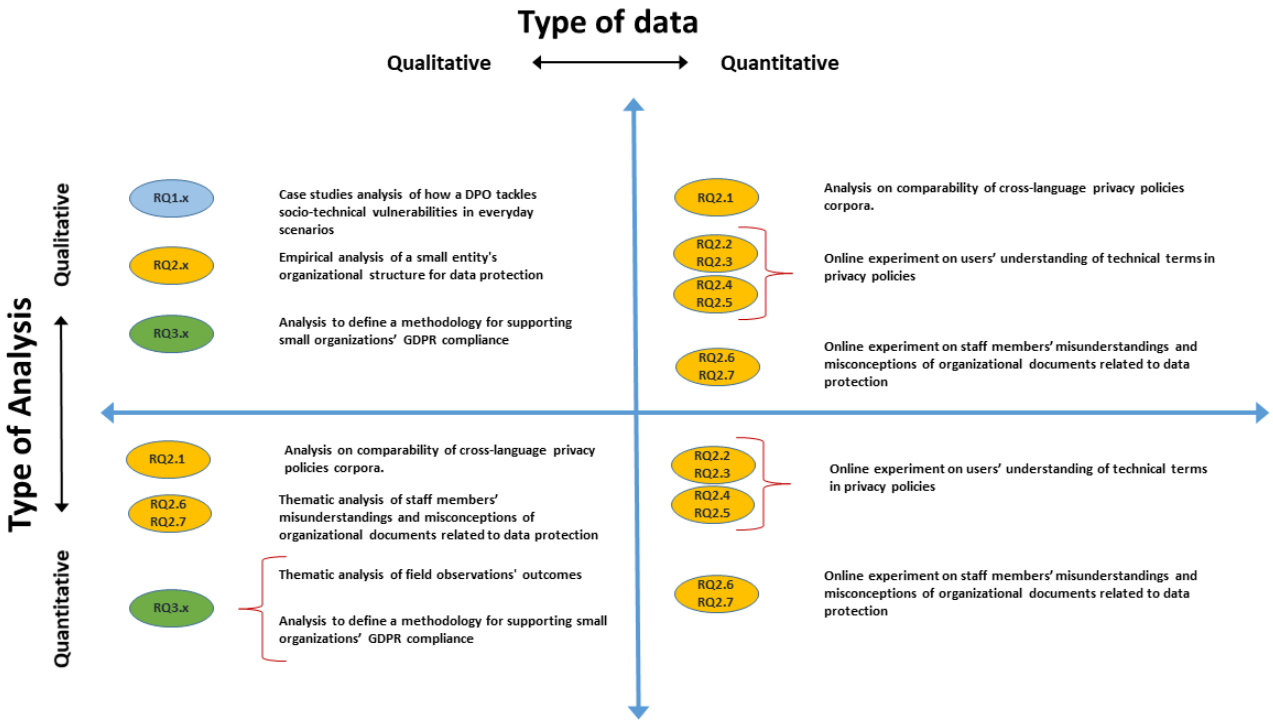

Figure 1.1: Visual framework of the research questions
Some research questions require the development of multiple contributions, shown in detail to their right. Suppose a research question occurs in more than one frame. In that case, different types of data (qualitative versus quantitative) and analysis (qualitative versus quantitative) were used to find an answer. This classification follows the one in Guest et al. [69].

1.2 Approach

In small entities (like educational institutions), compliance (not only formally) with data protection law requirements is challenging to achieve due to a lack of resources (such as time, financial, technical, and human). For example, Casutt et al. [23] empirical research discovered that one of the most common problems of the DPO is a lack of resources available.

In some cases, a misunderstanding of the role of data protection laws determines a significant difference between the formal definition of the privacy management processes and their actual implementation. Data protection management is sometimes perceived as unnecessary and, consequently, opposed by the employees or the decision makers of the company [72]. In this hostile climate for data protection, a controller could have insufficient knowledge of fundamental aspects of the processing carried out (e.g., in [129]).

Another difficulty encountered by smaller and less structured entities lies in the lack of knowledge and skills to manage business processes involved in data protection

issues. For example, a European study [51] revealed that programs of European master of science (M.Sc.) courses in cybersecurity do not cover knowledge units helpful for applying a privacy-by-design principle.

It is a scenario aggravated by the lack of economic and human resources, which prevents real, compliant management of personal data protection, as well as the creation and development of a corporate data protection culture [94]. In these smaller entities, the DPO's appointment does not produce the desired effects [23]. Sometimes, in an erroneous interpretation of the principle of accountability, appointing an external DPO could be conceived as mere outsourcing of the DPO's duties, not favoring the development and awareness of the corporate data protection culture.

In addition, in this scenario, entities likely tend to deal more with formal compliance with the rules on protecting personal data (such as the production of company regulations, information on processing, circulars, or the record of processing activities). This action neglects the development of a true and deep company culture of data protection, which is the only thing that guarantees compliance with the new accountability principle.

In summary, a structured approach to identifying, describing, and classifying these entities is needed to improve the understanding of the existent organization for data protection in small entities (like educational institutions).

1.3 Research Questions and Contributions

From Mendling et al. [112], it is known that process model understanding (accuracy and efficiency in comprehension) depends on the model of the process and the characteristics of the user who interprets it. A basis for developing such model is a thematic analysis, a systematic comparative technique to find themes and codes. It used grounded theory to build a theoretical model according to the Guest et al. approach [69], the Corbin and Strauss [41] grounded theory, and Yin [176] case studies analysis methodology. In this dissertation, Yin [176] case study approach was adjusted to consider aspects of Gioia's methodology for grounded theory [64, Appendix A]. Hence, I identified some general principles transferable to other scenarios. Indeed, according to Gioia's [64] vision, it is possible to generalize from a case study if it generates concepts or principles that have obvious relevance to another domain.

1.3.1 Understand what the DPO does

This work's first research question is related to empirically analyzing what a DPO does in its daily activity. The existing literature is surprisingly silent on this, except [23]. Considering the pervasiveness and importance of this role in an organization's data protection structure, knowledge of the daily challenges the DPO faces should be the starting point for further research. Knowing what a DPO does is also helpful for entities for which the GDPR does not apply: in many countries worldwide, there is data protection legislation in which a DPO role exists at some level [26]. The goal is to provide meaningful insight into the DPOs' daily challenges and the organizational weaknesses in personal data processing. I thus identified the following research questions:

RQ1.1: What does the DPO do in practice?
RQ1.2: Can the concrete activities of a DPO be enucleated in a few representative scenarios?

To answer **RQ1.1** and **RQ1.2**, I conducted a case study-based analysis to understand what the DPO actually does and what problems and challenges he or she faces.

This is discussed in Chapter 2. A finding of the chapter is that concrete activities of a DPO can be enucleated in a few representative scenarios. Using the perspective of the DPO in case studies to analyze the everyday activities and challenges of this security role, another result is providing practical insights that can be applied in real-world scenarios to support DPOs in what they do and what they may or must know.

1.3.2 Understand stakeholders' real practices

The final goal of this research question is to evaluate whether stakeholders involved in data protection correctly understand data protection-related documents (and, in this case, common misunderstandings and misconceptions). Indeed, ensuring users understand how data protection documents (like privacy policies) impact them is a key challenge for a GDPR deployment. This research question is comprehensive and complex, so I divided it into multiple subquestions.

Knowing what misunderstandings and misconceptions users may experience in understanding data protection documents provides indispensable elements to guide researchers in developing a simple methodology to improve data protection organization in small entities. Most studies on data protection-related documents, such as privacy policies, are based on US reality (e.g., [149], [159], [177], [178]), which has a differ-

ent culture and legislation than in other parts of the world (such as in the EU or China). It facilitates the comparison of papers and their results. Still, it may result in non-considering critical data because non-English-speaking users commonly read other privacy policies in their national language rather than in English. Hence, researchers need the availability of comparable cross-language privacy policy corpora aiming to replicate studies in different languages. I aim to investigate these topics in an EU national language and an EU context (e.g., focusing on the staff understanding of the Italian stakeholders). Thus, primarily, I identified the following research sub-question:

RQ2.1: Can a methodology be developed for generating comparable privacy policy corpora in different languages (e.g., English and Italian)?

To answer **RQ2.1** I conducted a study to investigate how to build a methodology to generate interlingual corpora of privacy policies usable from researchers to replicate in another language and to another target (like in the case of people who live in the EU Member States) a study in English and targeted to U.S. people.

This is discussed in Chapter 3. The findings of this chapter are a methodology for generating comparable cross-language corpora of privacy policies and a dataset of comparable multilingual corpora of privacy policies in Italian and English built to test the methodology and published on Zenodo.

The availability of comparable interlingual privacy policy corpora may not be enough to replicate a study in English in another language. This fact seriously limits the results' transferability to other contexts, such as the European Union. In fact, in the European Union, users who speak a national language other than English operate in the context of one of the 27 EU member states, interacting with documents written in their national language. Similar considerations also apply in other global contexts. Incidentally, the regulatory framework is also different in these contexts. Therefore, instead of blindly applying the results of original studies conducted in non-EU contexts, evaluating their transferability by replicating them in a European context is beneficial. This assessment of replicability goes beyond the use of parallel interlingual corpora. It involves checking for any changes in the documents written in different languages, such as variations in the numerosity of key technical terms or shifts in their meaning and significance across languages. Furthermore, given the substantial coding and analysis work required for such verification, it is prudent to explore the potential for streamlining and accelerating these tasks. Thus, secondarily, I identified the following research sub-questions:

RQ2.2: How is a tool-supported approach for key technical terms applicable in cross-language privacy policy corpora, such as English and Italian?

RQ2.3: Does the number of occurrences of technical terms and their meaningfulness change across policies written in different languages (i.e., English and Italian)?

To answer **RQ2.2** and **RQ2.3**, I conducted a study investigating the possibility of replicating a cross-national study on a privacy corpus in the Italian language using an auto-coding mechanism that considered the cross-lingual mapping of technical terms in privacy policies. In this study, I based the comparable corpora of privacy policy obtained as the outcome of *RQ2.1*, considering the example of English and Italian. Answering *RA2.2a* aimed to provide comparative analyses to determine (and refine) the semi-automated approach's reliability and the presence of any false positives and negatives. Answering *RQ2.3* aimed to code and analyze the considered interlinguistic corpora of the previous contribution to determine the key domain terms' occurrence frequency or meaningfulness. This aspect was crucial because, for example, some technical terms do not match and occur with significantly different frequencies from one linguistic version of the document to another.

These are discussed in Chapter 4. Specifically, the findings of *RQ2.2* are a codified algorithmic method for coding and analyzing interlinguistic corpora of privacy policies and a potential semi-automated approach for policies' coding. The findings of *RQ2.3* are some suggestions for replacing less interesting terms for users who speak the language of the replication study. The findings are the input for the contribution to answering the *RQ2.4*.

I applied the method of measuring how the frequency, meaning, and significance of the domain's key technical terms in parallel interlingual corpora changes in a real case of study replication to test its effectiveness and validity. I took Tang's study [159] on user misunderstanding and misconceptions of technical terms in privacy policies as a sample. Thus, I identified the following research sub-question:

RQ2.4: How can an English study be replicated using another language, such as Italian?

To answer **RQ2.4**, I conducted a study to determine how to modify an English-language survey to understand technical terms by adapting it for administration to Italian-speaking users.

This is discussed in Chapter 4. The finding of *RQ2.4* is a methodology to determine the composition of the Italian-language survey questions in replicating existing studies in another language.

I took the survey obtained as an outcome of the previous research question and

used it for a replication study in an online experiment with a sample of Italian-speaking users. Subsequent execution of statistical tests and qualitative analysis of the responses assessed how Italian-speaking users misunderstand and misconceive the privacy policies' technical terms and how this behavior differs from that of the English-speaking U.S. users analyzed in the original study. Thus, I identified the following research sub-question:

RQ2.5: Does understanding technical vs. non-technical terms change across policies written in different languages (i.e., English and Italian)?

To answer **RQ2.5**, I conducted an online experiment to assess the staff's understanding of technical terms in privacy policies in educational institutions.

This is discussed in Chapter 4. The finding of *RQ2.5* is a qualitative analysis of the staff's response to the survey to determine (i) their understanding of technical terms commonly used in privacy policies, (ii) their comfort with personal data processing practices outlined in the privacy policies or (iii) their comfort with data use practices outlined in the privacy policies, depending on these practices, are expressed using domain terminology (that of the GDPR) or non-technical but more explanatory and descriptive language.

The design of personal data processing that complies with the principles of the GDPR must consider the peculiarities of the human resources that interact and are part of the socio-technical system of educational institutions. It means investigating beforehand whether staff members of these institutions experience misunderstanding and misconception of organizational documents' content or processes related to data protection. Suppose this is the case; it means checking the most common misunderstandings and misconceptions, aiming to remove or mitigate them during system improvement activities. Thus, I identified the following research sub-questions:

RQ2.6: Do the educational institutions' staff members misunderstand and misconceptions organizational documents and processes related to data protection?

RQ2.7: What are common misunderstandings and misconceptions that educational institutions' staff members have in documents related to data protection?

To answer **RQ2.6** and **RQ2.7**, I conducted an online experiment to assess if and how the staff of educational institutions understand institutions' practices and procedures in personal data protection. To compose the survey questions, I considered several organizational documents that the educational institution used or should use, like data protection internal regulations, information on the processing of personal data, notices on the processing of personal data, instructions for data subjects' rights exer-

cise, contractual clauses for the transfer of personal data outside the EU, personal data breach procedures, data protection policies, cookie policies, and DPO appointments. I used appropriate statistical tests to make the online experiment comprehensible and qualitatively analyzed the evidence source.

This is discussed in Chapter 5. The findings of this chapter are that many staff members are unaware of data protection regulations, ignore the meaning of some data protection concepts, or do not know or are unaware of the data protection documents' contents adopted by their educational institutions. Another finding is that a possible solution to address these concerns is staff training on awareness of data protection processing principles, aiming to help the staff react better to procedures that fail and avoid designing processes that are great on paper but doomed to fail in the field.

1.3.3 Design a methodology to support small entities independently

This thesis's third research question investigates how to support complying with the GDPR principles of small entities that experienced a chronic lack of resources [23], including skills, to perform complex actions. The methodology building needs a complete and clear picture of the processes related to data protection. Thus, it is critical to ground it on the knowledge of the organizational structure for data protection of a sample of small entities belonging to the class of educational institutions, which is the one selected for analysis. This approach would allow for an accurate description of the organizational structure of small entities' class types. Prior knowledge derived from the outcomes of contributions *RQ1.1*, *RQ1.2*, *RQ2.1*, *RQ2.2*, *RQ2.3*, *RQ2.4*, *RQ2.5*, *RQ2.6*, and *RQ2.7* is required to respond to this research question.

Key challenge: How can we help small entities' socio-technical systems comply with GDPR requirements?
How can we help small entities improve the understanding of their organizational structure for data protection?

The initial method for this research was a literature review. According to Gioia's methodology [64], which enhances grounded theory development, I carried out this action by temporarily suspending judgment about existing conclusions to facilitate the discovery of new insights. Later, I included other methods to conduct this empirical research, such as qualitative analysis through case studies, survey handouts, recorded semi-structured interviews, and other sources of evidence (like documentation, archival records, or field notes). The data were first collected and analyzed to derive the theo-

retical construct. I used thematic analysis to build a theoretical model using grounded theory and case study analysis. The aim of this thesis was to understand what happens in the field and suggest concrete activities that educational institutions can accomplish to counter data protection-related breaches. Thus, I identified the following research questions:

RQ3.1: How can specific classes of small entities be helped to comply with their socio-technical systems with GDPR?
RQ3.2: How structural and high-risk violations can be identifiable and mitigable while balancing them with available resources and societal goals in the true spirit of the law (Recital 4 EU GDPR)?

To answer **RQ3.1** and **RQ3.2**, I conducted a field study on selected educational institutions to identify in real contexts the common violations in data protection processing and understand how to help these entities through the developing of a framework usable by staff members to improve data processing and reducing the potential violation in processing. I conducted the field observations according to a specific observation schema, minding all the daily activities related to every educational institution's roles and functions. I then qualitatively analyzed the field notes collected in field observation in a thematic analysis.

This is discussed in Chapter 5. The findings of this chapter are that a practical, risk-focused approach may be more effective than striving for perfect security, as they allow for continuous improvement while upholding the law's principles. Hence, by raising awareness of data protection principles rather than focusing solely on legal jargon, staff can better respond to procedural failures and avoid impractical processes. As an outcome, this chapter offers valuable insights for managing schools and small organizations, helping them identify and correct mistakes in personal data processing.

1.4 Author's contribution

Chapter 2 provides an overview of the Data Protection Officer (DPO) role. Among all cybersecurity and privacy workers, the DPO stands between those auditing a company's compliance and those acting as management advisors. A person who must be somehow versed in legal, management, and cybersecurity technical skills. I describe how this role tackles socio-technical risks in everyday scenarios.

The chapter is partially published in: [33] *F. Ciclosi and F. Massacci, "The Data Protection Officer: A Ubiquitous Role That No One Really Knows," in IEEE Security &*

Privacy, vol. 21, no. 1, pp. 66-77, Jan.-Feb. 2023, doi: 10.1109/MSEC.2022.3222115.

Chapter 3 proposes a methodology for building comparable cross-language in a national language and a reference study language. The methodology is then explained through its application to compare English and Italian, extending the corpus of one of the first studies about users' understanding of technical terms in privacy policies. Moreover, it investigates other open issues that can make replication harder.

The chapter is partially published in: [61] *Ciclosi, F., Vidor, S., Massacci, F. (2023). "Building Cross-language Corpora for Human Understanding of Privacy Policies." In: Skarmeta, A., Canavese, D., Lioy, A., Matheu, S. (eds) "Digital Sovereignty in Cyber Security: New Challenges in Future Vision." CyberSec4Europe 2022. Communications in Computer and Information Science, vol 1807. Springer, Cham.* `https://doi.org/10.1007/978-3-031-36096-1_8`. I have been in charge of conceptualization, methodology, software, validation, investigation, data curation, writing (both the original draft and review & editing), and visualization.

Chapter 4 conceives a methodological approach to investigating the different frequencies of occurrence and meaningfulness of technical terms that exist in different language versions of the same privacy policy, which belong to interlingual corpora of privacy policies. The proposed approach consists of the following steps: (i) a tool-supported match of technical terms in cross-language privacy policy corpora; (ii) a manual assessment of the meaningfulness of the translations of technical term change in cross-language privacy policy corpora; and (iii) the identification of the occurrences of mapped terms and, thus, the different relevance of the terms. The approach validity is tested by replicating in the Italian language the study by Tang et al. on users' misunderstanding and misconception of technical terms in privacy policies.

The chapter will be submitted to: IEEE Transactions on Privacy

Chapter 5 reports the results of a multi-site field study on educational institutions and the challenges they face in implementing the GDPR while running activities full of sensitive issues without adequate resources. It discusses how privacy violations happen on the field, from critical with potential impact on pupils' security and safety to 'formal' violations for which life is too short to bother, and how a risk-based approach would be more effective in addressing them.

An article partially based on the chapter has been accepted to the 46th IEEE Symposium on Security and Privacy.

Finally, Chapter 6 reflects on the main contributions of this work and provides a discussion on future work.

1.4.1 Acknoledgements

The European Union has partly supported this work under the H2020 Leadership in enabling and industrial technologies (LEIT) program under grant agreement 830929 (CyberSec4Europe).

Chapter 2

The Data Protection Officer, a ubiquitous role nobody really knows

Abstract: This chapter describes how the data protection officer (DPO) role tackles sociotechnical risks in everyday scenarios. The chapter explains the challenges in legal, management, and cybersecurity technical skills and how the person holding the office must balance between those auditing a company's compliance and those acting as management advisors.

2.1 Introduction

The recent application of Regulation (EU) 2016/679, best known to the world as the *General Data Protection Regulation* or GDPR, introduced the role of the data protection officer (DPO). While DPOs have been a key enabler of the GDPR [6], the role of this privacy worker is not a new concept: in several EU Member States, its appointment was already good practice for some years. Yet, the GDPR does not formally describe the DPO job profile, and most papers discuss how to support a DPO with algorithms without providing practical examples of what the DPO does.

For example, Diamantopoulou et al.[50] identify which ISO 27001/2 controls need to be extended to meet GDPR requirements and which of them the DPO is involved - but why in some and not in others? Ryan et al. [147] explain how their framework RegTech can be helpful to a DPO for checking GDPR compliance - but when and for what concretely? Chatzipolidis et al. [30] describe a readiness assessment tool for GDPR's compliance - but for solving social or technical issues? Other articles discuss

GDPR's compliance topics as if DPOs did not exist, from software engineering [107] to socio-technical management processes [104] or the GDPR's cost among cybersecurity investments [94].

The purpose is to introduce this legally required organizational role — this ubiquitous privacy worker — to the engineering community represented by Security & Privacy readers through concrete examples of what problems DPOs face, what they do, and what they may or must know. While the literature is surprisingly silent about this, we think that the knowledge of the everyday challenges that DPOs have is the starting point for all subsequent research activities. For example, if a researcher has no reference to the daily activities of *the* key privacy worker in charge of GDPR compliance, how can one design logic or a tool for checking this privacy compliance or any privacy-by-design technology with a practical impact? In summary, the key research question is:

- *Can we enucleate in a few representative scenarios the concrete activities of a DPO?*

The article focuses on the role of the DPO introduced by the GDPR, but the insights are valuable for readers outside the EU countries. The GDPR can apply to organizations that carry out their activities in the EU and organizations outside the EU that process the personal data of EU data subjects. Further, in many countries worldwide, there is data protection legislation in which a DPO role exists at some level. The International Association of Privacy Professionals (IAPP) lists the different roles in many countries worldwide that share some characteristics with the DPO legally defined in the EU [1].

2.2 Methodology

The insights described in this article are grounded in case studies along Yin's case study methodology [176, Ch.4] and the suggestions[2] by Glazer, the founder of grounded theory, to build core categories across field observations derived from the live experience.

First, we analyzed data protection laws and recommendations of relevant authorities. Secondly, we analyzed the seven functions of the DPO that the European Data Protection Supervisor (EDPS) identified in its paper on the role of DPO in compliance with Regulation (EC) 45/2001. Then, we looked at the summary of opinions of some

[1] `https://iapp.org/media/pdf/resource_center/dpo_requirements_by_country.pdf`

[2] In Glaser, B. G., & Holton, J. (2004, May). Remodeling grounded theory. In Forum qualitative sozialforschung/forum: qualitative social research (Vol. 5, No. 2). `https://www.qualitative-research.net/index.php/fqs/article/download/607/1316/`.

supervisory authorities (i.e., Bulgaria, Croatia, Italy, Poland, and Spain) on the DPO's activities involved with these functions (e.g., [88]) to have a perspective that was not restricted to a single country.

To make this chapter concrete as a use cases references, we selected only sources of information for which there was evidence that the activities carried out by DPO involve at least one of these seven functions. The starting point for the case study selection was the personal experience of the candidate, who has been a DPO in the Italian public administration for the past five years and is a member of the Italian Association of DPOs.

To make the results of this study accessible, we looked for some publicly available information (for example, court decisions, supervisory authorities' decisions, and newspaper articles) on case studies similar to the ones on which the candidate had first-hand experience. This approach allows us to go beyond the individual experience and provide shareable evidence. Out of over 90 public case studies, we finally distilled 12 scenarios with at least one analyzable practical example for each of the DPO's primary functions. In this chapter's main text, we made the scenario general by renaming the actors involved (but without altering their nature) to be resilient to potential requests on the "right to be forgotten" and be of general interest to the readers.

The scenarios we propose map well into the general literature. For example, the site `https://www.enforcementtracker.com/` reports 1475 occurrences of GDPR fines related to DPOs across 31 different EU Member or EEA States. The proposed scenarios cover more than 1400 cases.

The concrete references are listed in the supplemental material available in Appendix A.

2.3 How this role is born

In Europe, the concept of DPO comes from the German data protection law of 1977, the Bundesdatenschutzgesetz (BDSG), which introduced a precursor of the role (Figure 2.1).

The DPO role over time became widely adopted by the other European States until, in 1995, the European Community issued Directive 95/46/EC on the protection of individuals concerning the processing of personal data and on the free movement of such data. A patchwork of approaches followed: many Member States introduced the DPO role in their national law, as in Austria (where the appointment was mandatory) or in France (where the appointment was optional), but only in some of them (in Italy,

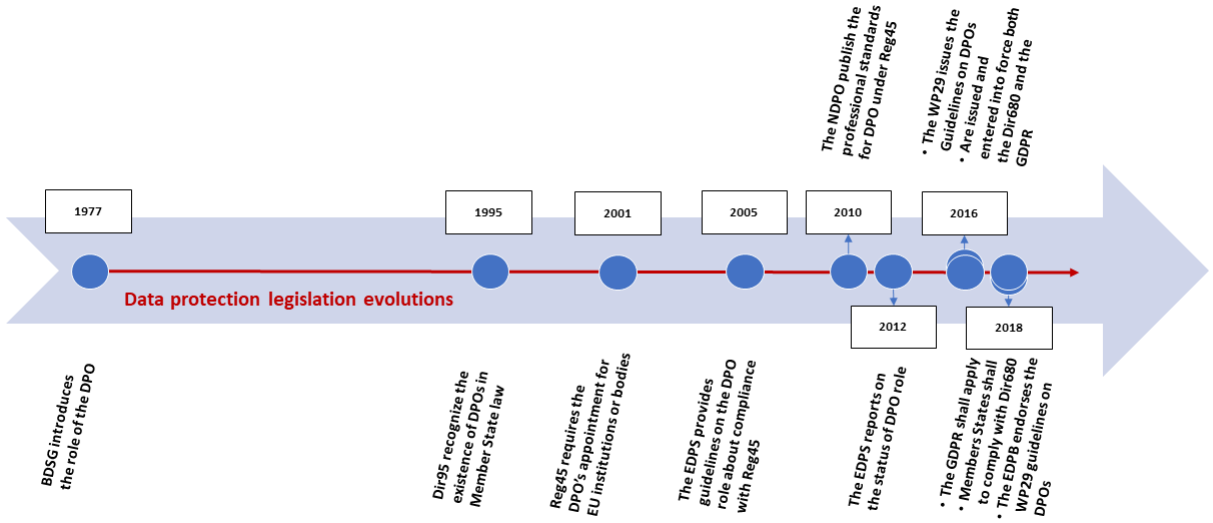

Figure 2.1: The evolution of the European role of Data Protection Officer (DPO)

the DPO role was absent in national law). Moreover, the DPO duties were limited to independently ensuring an organization's internal application of the national provisions taken according to the Directive and keeping a register of processing operations carried out by the controller.

For EU institutions, only the appointment of at least one DPO was mandated by Regulation (EC) 45/2001. These rules were very similar to the ones that would be introduced in later years. 2016 was a pivotal year for data protection as the European Parliament and the Council issued the Directive (EU) 2016/680 on criminal offenses or criminal penalties and the GDPR on personal data protection.

To provide an interpretation of the EU's data protection legislation, the Article 29 Working Party Committee issued the Guidelines on DPOs (WP243) [11], which was initially adopted on December 13, 2016, and later revised on April 5, 2017. After the GDPR adoption, the new European Data Protection Board (EDPB), an independent European body tasked with ensuring the consistent application of data protection rules throughout the European Union, endorsed these guidelines in its first plenary meeting on May 25, 2018.

Not all organizations are required to appoint a DPO, although doing it is a good practice. There are three specific cases in which a controller or a processor must appoint a DPO (GDPR article 37). In the first case, if the organization is a public institution, it must appoint a DPO. Otherwise, an organization must appoint a DPO if it requires regular and systematic monitoring of data subjects on a large scale (second case) or processes on a large scale personal data which belong to special categories or are related

Table 2.1: High-Level View of the Functions of the DPOs

This table describes the tasks of the DPO grouped according to the seven functions of the DPO that the European Data Protection Supervisor (EDPS) identified in its position paper on the role of DPO in compliance with Regulation (EC) 45/2001.

DPO's Functions	Summary Descriptions
Organizational function	Review or even directly organize a processing operations register on behalf of the controller, help both assess the related risks, and support the processing activities with high-risk value.
Monitoring of compliance	Investigate (on autonomous initiative) matters and occurrences directly relating to the GDPR and report back to the controller
Advisory function	Make recommendations for the practical improvement of data protection to the controller and advise it on matters concerning the related provisions.
Cooperative function	Facilitate cooperation (between the Supervisory Authority and the controller), especially in the frame of investigations, complaint handling, or prior checks.
Handle queries or complaints	The authorization to handle queries or complaints originated from the very possibility of autonomous investigations.
Information and raising awareness function	Prepare staff information notes, training sessions, privacy statements, and learning material.
Enforcement	Powers of enforcement are limited.

to criminal offenses (third case).

The GDPR Article 39 entrusts the DPO with various tasks:

1. to inform and advise one's employer on how to carry out processing about its obligations under the law;

2. to monitor compliance with the law;

3. to provide advice regards the data protection impact assessment (DPIA) and monitoring its execution;

4. to cooperate with the supervisory authority (SA) as a contact point between this and the organization.

Table 2.1 summarizes the seven functions of the DPO, identified by the EDPS in its position paper on this role. Appendix B provides further information and examples about these seven functions of the DPO.

In summary, DPOs must be fully cognizant of the controller's working environment to carry out their tasks. This awareness implies that a DPO must know the internal distribution and allocation of responsibilities and tasks related to every personal data processing. A DPO must also be familiar with any external links (between the controller and other organizations) and with legal frameworks in which these links take place. According to many supervisory authorities' opinions [88], a preliminary task in which a DPO scopes the controller's environment fulfills this requirement.

2.4 What problems does a DPO face?

To understand the daily problems a DPO faces, we have illustrated several real-life scenarios in Table 2.2. We edited them to obfuscate the original entity responsible for the privacy issues faced by the DPO. The supplemental material reports the sources of these scenarios.

Empirical research on the problems of the DPO [24] has shown that sometimes these could be very basic and relate to a lack of sufficient resources (time, finances, and humans) to carry out one's duties and to some issues in the operational interpretation of the law. Below in Table 2.2, we analyze the challenges that DPOs face even when adequately supported.

A first example: although maintaining a record of processing activities is formally a controller's duty, the DPO will most likely be in charge of this work or closely involved in its oversight activities. In quite some job adverts for DPO appointments, it is formally stated that the DPO is in charge of maintaining the record of processing activities. Without a regulatory constraint, the DPO may also be appointed to carry out some activities that are formally a duty of the controller. It is a free choice of the controller who pays the DPO. The involvement of the DPO is only sometimes an indicator that things are working well. A data controller could use a non-GDPR-compliant service because of a DPO's mistake (Tab. 2.2, Wrong-Advice). Unfortunately, the controller is solely responsible for the choice made, and even a DPO's evaluation errors expose it to administrative fines or penalties. The accountability principle constrains the controller to demonstrate having complied with the regulatory requirement.

In an opposite scenario, a controller operating a catering service implements a new data processing to control the EU Digital Covid Certificate of staff but neglects the DPO's advice without justifying in writing why that advice has not been taken into account (Ignored-DPO-Advice). This action is to blame because it is reasonable that the DPO gave his or her advice to ensure that the processing complied with the

Table 2.2: Scenarios and Privacy Issues daily faced by a DPO

Short name	Scenario description	What went wrong
Wrong-Advice	A controller wants to process data using a processor service and asks the DPO's advice to understand if the proposed contract complies with GDPR.	The advice of the DPO turned out to be wrong, but the controller uncritically trusted the DPO's advice.
Ignored-DPO-Advice	A controller implements a new data processing. The DPO advises that a prior execution of a DPIA is required.	The controller chooses to neglect the DPO's advice without justifying in writing why it did not take into account that advice.
DPO-Advice-Not-Sought	A controller implements a registration procedure of service without prior asking for DPO's advice about compliance with GDPR principles	The registration procedure does not carry out a check on the identity of the person who enrolls, so it is unknown who saw the data
Website-Forces-Choices	In a controller procedure, data processing for marketing purposes asks for the consent of the data subject.	The procedure forces a data subject to release consent, imposing him or her to select a box, otherwise preventing it from continuing.
Admin-Asks-For-Everything	A controller's staff member satisfies the law on access to data (e.g., FOIA) by publishing some documents about a data subject.	A controller processes personal data by asking the data subject and publishing all data without correctly applying the data minimization principle.
Neglected-Subject-Right	A data subject files a complaint with the competent supervisory authority because s/he has not received a response within the time frame set by law	The controller did not designate the DPO, or it did, but the designated DPO did not monitor the official address.
Subject-Right-Request	A data subject exercises one's rights of access according to article 15 of the GDPR, sending a formal request to the DPO's address.	When the controller drew up the record of processing activities, it did not correctly identify the actual processing activities, so the DPO could not answer the query.
No-Data-Protection-Principles	A controller implements a configuration in company equipment.	There is an incorrect configuration that allows unfettered access to personal data.
Unchecked-Remote-Monitoring	A controller implements remote assistance software tools on company workstations	The technicians use a remote assistance software solution that does not notify the user when remote access is performed.
Wrong-Public-Procurement	A controller issue a public tender for procuring products or services.	In the tender evaluation grid, there is no explicit checkpoint for applicants to "demonstrate" that their products or services fully comply with the GDPR.
Software-End-of-Life	The DPO of a company noticed some results about the software used for processing activities.	The software used in the company's workstations is obsolete, and the support provided by the vendor software house is expired or close to expiring.
Subcontractor-Violates-Privacy	A controller appoints its DPO to test a software procedure of a tender's winner	The tender winner has violated the contract for the supply of IT programs and services terms on GDPR compliance.

GDPR. Again, not documenting the reasons behind choosing to neglect this advice results in violating the accountability principle, exposing data subjects to risks and the controller itself to administrative fines or penalties.

The DPIA execution is a controller's responsibility, while the DPO task is limited to providing the advice requested. The WP29 [11] highlights that a controller should justify in black and white in the DPIA's documentation why it has not considered the DPO's advice. For properly handling the execution of a DPIA, a controller could define, in an internal regulation, the procedures for consulting the DPO about this topic. By way of example, this regulation may contain whether or not to carry out a DPIA, what methodology to follow, and whether to use internal resources or outsource it. Other helpful information to include in regulation is what safeguards to apply to mitigate any risks to the rights and freedoms of the data subjects.

The scenario DPO-Advice-Not-Sought describes a case in which a controller is exposed to mistakes in the processing's design because it did not ask the DPO's advice. The scenario Website-Forces-Choices exemplifies the vulnerability stemming from a wrong implementation of a seat reservation software procedure of a transport company. The software forced the data subject to consent to other forms of processing.

The Admin-Asks-For-Everything scenario is related to a failed application of the minimization principle. In it, the human resources office of a public institution (which must satisfy the law on publication obligation) publishes in the *"Transparent Administration"* section of the institution's website the unredacted CV of the winner of a public selection, thus exposing the data subject's personal data (e.g., home address, personal phone number, and others).

The relationship between the DPO and the supervisory authority (SA) is significant. The WP29 [11] highlights that the DPO must act as a "*facilitator*" by cooperating with the SA. Furthermore, the obligation of secrecy or confidentiality cannot prohibit the DPO from contacting and seeking advice from the SA. The controller who acts to weaken this relationship could be sanctioned. If the data controller does not appoint the DPO, problems will likely arise in scenarios Neglected-Subject-Right and Subject-Right-Request. In Neglected-Subject-Right, another violation is the missing communication to the SA of the DPO's contact details because, in that case, the SA does not know how to contact the DPO to handle the complaint. In Subject-Right-Request, an aggravating factor would occur if, when the controller draws up the record of processing activities, it does not correctly identify them. In such a case, even if appointed, the DPO may find it challenging to identify the processing details and promptly respond to the data subject's requests.

2.5 Risk management

DPOs face many challenges that we can classify into two categories: technical risks and socio-organizational risks. The first challenges stem from faulty technical or technological solutions which do not fully guarantee the protection of the subjects whose personal data is processed. The second type of challenge is due to incorrect organizational procedures or incorrect human behavior.

We can subdivide technical risks into two subgroups. The first type relates to design problems when the DPO supports and advises the controller on technical choices. For example, the DPO may be asked to choose between configurations that comply with the data protection by default principle and configurations that seem to do so (Tab. 2.2, No-Data-Protection-Principles and Unchecked-Remote-Monitoring). S/he might find (or fail to find) insecure technical solutions in a call for tenders (Wrong-Public-Procurement) that can cause kick-start litigation between contractors. The second type of risk involves misconfigurations or errors which appear during the implementation (Subcontractor-Violates-Privacy and Software-End-of-Life).

Socio-organizational risk can appear daily while the DPO supports the controller in processing activities. We can divide them into four additional categories, such as auditing (Tab. 2.2, Ignored-DPO-Advice and Subcontractor-Violates-Privacy), communication (No-Data-Protection-Principles and Unchecked-Remote-Monitoring), processing designing (Software-End-of-Life) and relationship (Neglected-Subject-Right). Conflicting requirements are a particular instance of these risks.

No matter how sophisticated the technical protection measures are, DPOs may experience side-channel attacks from the most unexpected human behavior. A well-designed processing activity may become unlawful because a staff member operates differently from what is prescribed by the procedure and from the instructions received by the controller. For example, an operator recorded the screen of the closed-circuit video surveillance cameras, only accessible to the local police control station, with a smartphone and disseminated a video of a traffic accident on social channels. Appendix C provides further information about technical and socio-organizational risks.

Table 2.3 shows some of the possible consequences of damaging the freedom and rights of natural persons. It is difficult to determine the impact grade without an in-depth analysis of the processing and the technical and organizational means used to mitigate their risks. In some circumstances, the case study's high-level description (e.g.,

in the case of the GDPR accountability principle's violation) is sufficient to conclude that the consequences are actual and potentially disastrous for an organization. Because personal data breach is a broad concept, we split this class into three classes representing the specific breach. Finally, Table 2.3 includes some more specific classes (e.g., "*GDPR non-compliant processing*" is a particular case of "*unlawful processing*").

Some DPO mistakes (such as WRONG-ADVICE) may also cause a controller's wrong assessment, which subsequently induces wrong organizational choices.

If a threat exploits even one vulnerability, it will determine a GDPR's principles violation, exposing the controller to fines or penalties. Of course, this might depend on somebody tipping the supervisory authority or the SA coming to investigate its initiative or after a data breach.

2.6 What does a DPO actually do?

The boundary between the DPO's functions sometimes is fuzzy, especially in complex scenarios where more of them are involved. In Table 2.4, we summarized some examples, presenting the role of DPO in mitigating vulnerabilities (namely: *prevention*, *detection*, *response*, and *investigation*) for each scenario from Table 2.2.

In DPO-ADVICE-NOT-SOUGHT, the DPO's activities correspond to a detection (D1 and D2) and a response scenario (R2, R5, R7, and R10). The DPO primarily exercises the monitoring of compliance. However, in the response part, there is a combination of the advisory, organizational, and cooperative functions, performed secondarily, as well as the enforcement one.

The case WEBSITE-FORCES-CHOICES corresponds to two different scenarios. The first is a detection scenario in which the DPO exercises the function monitoring of compliance (D1 and D2), while the second is a response one where the Enforcement function is applied (R5).

In the scenario ADMIN-ASKS-FOR-EVERYTHING, the DPO carries out his or her activities in a response scenario. Some (R3 and R9) combine enforcement and handling queries or complaints functions. At the same time, another (R11) involves primarily the handling queries or complaints function and secondarily the monitoring of compliance one.

The importance of the DPO's cooperation function comes to light from the scenarios NEGLECTED-SUBJECT-RIGHT and SUBJECT-RIGHT-REQUEST. These cases are linked to two response scenarios. The first is primarily involved in the DPO's handling queries or complaints function, followed by the function monitoring of compliance

Table 2.3: Some possible consequences of risks faced by a DPO may deal with

This table summarizes some examples of the consequences of technical or social risks that a DPO may face while carrying out his or her duties. These consequences are grouped into classes and associated with a possible reference scenario. Any such infringement exposes the controller to administrative fines of up to 20 000 000 EUR, or up to 4% of the total worldwide annual turnover, whichever is higher.

Impact's class	**Impact example**	**Scenario example**
Attack on customers	Identity theft of the data subject could happen (e.g., an attacker steals from a misconfigured database a data subject's personal data and later uses them pretending to be the data subject).	No-Data-Protection-Principles, DPO-Advice-Not-Sought, Ignored-DPO-Advice
Attack on employees	Unauthorized persons could capture the private data of employees	Unchecked-Remote-Monitoring
Contract termination	It is possible that the tender be subject to litigation due to a violation of the terms indicated in the contract for the supply of IT programs and services	Subcontractor-Violates-Privacy
Data breach (access or disclosure)	A personal data breach may happen because unauthorized access to (or disclosure of) personal data transmitted, stored, or otherwise processed (e.g., a software's misconfiguration allows technicians to connect to a workstation without the user's consent stealthily).	Unchecked-Remote-Monitoring, Subcontractor-Violates-Privacy, No-Data-Protection-Principles
Data breach (destruction or loss)	There is unlawful destruction or an accidental loss of personal data held by the controller (e.g., a malware of type ransomware has ciphered all office's files stored in a file server)	Software-End-of-Life
Data breach (alteration)	There is an unlawful alteration to personal data stored (e.g., exploiting a system's vulnerability, a cracker modified some personal data stored in a database)	Software-End-of-Life, No-Data-Protection-Principles
Data Disclosure	Excess and irrelevant personal data are disseminated over the Internet by the controller.	Admin-Asks-For-Everything, Neglected-Subject-Right
Inadequate identification	Anyone could pretend to be another data subject and access his or her personal data.	DPO-Advice-Not-Sought
GDPR non-compliant processing	A controller processes personal data without providing information to the data subject (e.g., a controller issues a loyalty card to a customer without providing the customer with a privacy policy)	Wrong-Advice, Ignored-DPO-Advice
Risky pro-	The processing could be risky for the rights and	Ignored-DPO-

Table 2.4: Some examples of activities that a DPO carries out to mitigate consequences of realized risks

ID	Illustrative mitigation by the DPO	Applicable Scenarios (or DPO's function)
Prevention - Controller initiated		
P1	A group of joint controllers (two or more controllers who jointly determine the purposes and means of processing) asks their DPOs for advice on the technical and organizational aspects of periodic or new processing. For example, processing will build on an integrated territorial video surveillance system using OCR cameras.	*Advisory function*, *Cooperative function*, *Organizational function* and *Monitoring of compliance*
P2	The DPO helps the controller to do a DPIA before starting new high-risk processing (e.g., one relating to a Covid-19 screening data acquisition and management system).	*Organizational function*
P3	The DPO supports the controller in choosing the configuration of a company's telecommunication equipment or a software tool that guarantees the protection of personal data by default.	No-Data-Protection-Principles, Unchecked-Remote-Monitoring
Prevention - DPO initiated		
P4	A DPO monitors the data protection laws changes and the indications of the bodies in charge, and after that, he prepares short information pills or notes for an SME's staff.	*Information and raising awareness function*
P5	The DPO advises the controller that in issuing public tenders, it should expressly call for applicants that can "demonstrate" that their product or service fully complies with GDPR.	Wrong-Public-Procurement
P6	The DPO advises the controller to launch a census of all the PCs in the organization that have a Microsoft operating system version 7 or older. The DPO interacts with the ICT Department to develop an updated operating software plan.	Software-End-of-Life
P7	The DPO of a National central bank illustrates to some banks' DPOs the controller's obligations.	*Advisory function*
P8	The DPO consult the competent SA about implementing a new data processing.	*Cooperative function*
Investigate - DPO initiated		
I1	The DPO initiate an investigative activity to verify compliance with contract terms.	Subcontractor-Violates-Privacy
I2	The DPO immediately advise the controller about the existence of a violation of the contract terms. The controller, in turn, immediately formally warns the processor about this violation, ordering it to stop the infringement at once (e.g., by returning the data encryption key).	Subcontractor-Violates-Privacy
Investigate - Data subject initiated		

(R11). The second relates to handling queries or complaints and monitoring compliance functions (R4 and R12). Moreover, related to NEGLECTED-SUBJECT-RIGHT, there is an additional response scenario in which the DPO's cooperative function is the primary involved, which follows up handling queries or complaints function (R6 and R8). The cases NO-DATA-PROTECTION-PRINCIPLES and UNCHECKED-REMOTE-MONITORING highlight the importance of the DPO's organizational function (P3), primarily in a prevention scenario. The advisory function follows up as second in the same.

In WRONG-PUBLIC-PROCUREMENT, there is an example of a prevention scenario in which the DPOs exercise their advisory function (P5). While in SOFTWARE-END-OF-LIFE, there is a prevention scenario where DPOs exercise a combination of their organizational and advisory functions (P6). The SUBCONTRACTOR-VIOLATES-PRIVACY shows the DPO's monitoring of compliance use in an investigative scenario (I1). Then, in a combination of response and investigative scenarios, the DPO exercises the monitoring of compliance function, followed by a mix of the advisory, organizational and cooperative functions performed secondarily (R2, R7, R10, and I2).

Regarding the DPO's advisory function, when DPOs are involved in new data processing, they can consult the SA if necessary (P8). In WEBSITE-FORCES-CHOICES, when the DPO becomes aware of a process that is not entirely compliant with data protection policies (D2), he or she advises the controller, making recommendations for practically improving it (D1 and then R3 or R5).

Finally, the DPO should perform the investigative activity even if the controller does not involve him or her. For example, in SUBCONTRACTOR-VIOLATES-PRIVACY, the DPO can independently carry out this activity (I1). If he or she finds a violation of data protection by design and default principles, the DPO acts accordingly (I2).

2.7 What does a DPO need to know?

GDPR does not specify the professional qualities required at a DPO, but only that the needed expert knowledge must be adequate for the data processing operations and their protection rank. The WP29 [11] suggests that it must be commensurate with the sensitivity, complexity, and amount of data the organization processes. Table 2.5 summarizes the expertise and skills that a DPO should have. They consist of qualities, expertise in law and practices, ability, and educational qualification.

In the absence of relevant bodies' specific guidance, from a legal perspective, it is challenging to define DPOs' selection criteria that can truly measure the adequacy of their level of knowledge. While technical and management skills are essential, there

Table 2.5: Expertise and skills of the DPO

This table describes the expertise and skills a DPO should have to carry out his or her tasks well.

Criteria	Description	Examples
Qualities	DPOs must possess specific professional qualities	*Supervisory authorities* provide continuous training courses reserved for DPOs (e.g., the T4Data international project and the SME Data project). *A controller* required in the call for DPO's appointment that the candidates must have: in-depth knowledge of the organizational structure, the information systems present, and the specific sector of activity of the controller, as well as being familiar with the data processing operations carried out by the latter. *EDPS* asserts that it is better to recruit the DPOs of EU institutions/bodies/agencies (EUI) within the EUI. These people usually ensure a better knowledge of the organization, structure, and functioning of the EUI itself.
Expert in law	DPOs must have expert knowledge of data protection law	*The EDPS* asserts that the expert knowledge of data protection law is a prerequisite to the EUI's DPOs function. *A controller* required in the call for DPO's appointment that the candidates must know the legislation and practices on data protection both from a legal and IT point of view, including in-depth knowledge of the GDPR.
Expert in IT practice	DPOs must have expert knowledge of IT, security, and organization	According to *EPDS* for EUI's DPOs, one of the professional qualities is knowledge of IT, including security aspects and organizational and communication skills. *The Network of Data Protection Officers* of the EU institutions and bodies recommends that the EUI's DPOs should have at least three years of relevant experience/-maturity, to serve as DPO in a body where data protection is not related to the core business. Otherwise, this period grows to at least seven years. Similarly happens if the DPO will serve in an EU institution or which has an essential volume of processing operations.
Ability	DPOs must have the ability to fulfill the tasks listed in GDPR, Article 39	*A controller* explicitly required in the call for DPO's appointment that the candidates must have personal qualities, including integrity and high professional ethics. According to *EPDS* for EUI's DPOs, the DPOs' ability to fulfill their tasks should be referred to their personal qualities and knowledge and their position within the organization.
Educational qualification	DPOs could have a variety of qualifications in law and computer science, security and privacy	However, they cannot be uniquely determined. *An Italian court ruling* asserts that holding ISO27001 certification cannot be a binding prerequisite in the selection procedure for a DPO's appointment.

is no consensus on "specific" certifications that guarantees an adequate expert knowledge level of the DPO. As a result, it is complex to establish the absolute value of specific qualifications (e.g., university master's), professional certifications (e.g., UNI 11697: 2017 certification), and being an author of books, articles, papers, or research products. Courts reverted, as unfair requirements, several attempts to mandate this or that certification (e.g., BS-7799 or ISO 27001).

In my experience, a DPO can achieve a good knowledge of data protection practices by studying documents arranged by EDPB, EDPS, ENISA, and supervisory authorities of EU member states.

Another critical point is finding appropriate training for the professional profile of the DPO, from both a technical and a legal point of view: while knowledge of data protection law is a crucial requirement, a DPO may not have a law degree. A DPO may have a cybersecurity degree, but a European study [51] found that European master of science (M.Sc.) programs in cybersecurity practically do not cover the knowledge units on component procurement. Knowing this unit is critical to guarantee compliance with the privacy-by-design principle because third-party components and contracts with IT providers are the norms for any administration because they rarely have in-house developers.

Some training support can also come from internal and external information sources. An example of internal ones may be information that a SOC (Security Operation Center) or the IT staff provides to DPO about events and incidents of security. The national CSIRT (Computer Security Incident Response Team) is an external source that provides pre-alerts, alerts, bulletins, and information regarding risks and incidents. For instance, in case of a data breach, a DPO should collaborate with the internal IT department (if available) and later refer to the national CSIRT. As a further example, considering the interaction between a DPO and a SOC, the DPO will not have direct access to a security incident and event management (SIEM) system for an independent analysis of the inputs collected from the connected security devices and sensors [19]. Instead, the DPO will refer to summary reports prepared by security analysts. In the case of a data breach uncovering, direct contact with security analysts could help obtain more information about the incident and better determine its impact and extent. Unfortunately, in many places (e.g., small public administrations), the "IT security department" is just *one* IT person who, among other duties, knows something about security. The DPO may end up being *the* security expert.

The currently available technology can help DPOs make it easier to carry out their tasks. For example, using requirement analysis tools in software development or pro-

curement could improve compliance with the data protection by design principle.

Additional support comes from research. Research findings can provide DPOs with insights into where to focus their efforts. According to this vision, the DPO advising task is "driven by research." For example, Tang et al. [159] find that users have difficulties understanding the technical terms used in privacy policies because they misunderstand and misconstrue them. As a result, also the privacy policies themselves are misunderstood and misinterpreted. Considering the results of this and other similar studies helps the DPO understand the training gaps in the firm's workforce.

2.8 Conclusions

The goal of compliance with the GDPR makes the DPO's role dual. This data protection specialist is both the person who controls the processing activities in the organization and the person who acts as a wise advisor to the management. This tension could be problematic as the DPO needs information to carry out his or her duties. At the same time, the manager wants to have support in determining the purposes and means of processing personal data without giving the DPO too much information.

The duties assigned to the DPO role can quickly put this person in conflict within the organization for which she or he works. This case could happen especially in organizations with a negative attitude toward data protection. For example, Hadar et al. [72] reported a qualitative study where 17 developers out of 27 declared that the climate of the organization they worked for was averse to data protection. Developers reported that they must comply with organizational norms against data protection laws contradictory to the company's formally stated policies. In these circumstances, DPOs who carry out their duties will likely experience conflicts with the management. Casutt et al. [24] found a similar outcome. They showed that DPOs experienced an inherent conflict between complying with the law and realizing the organization's project whenever there is a gap between privacy requirements and those of the organization for which the DPO works.

A potential limitation of this research is that we based the proposed scenarios on a detailed analysis of 90 specific cases, mostly of Italian origin, and court decisions may differ across EU Member States. Several factors mitigate this issue. At first, the relevant legislation (the GDPR) is a single regulation for all EU Member States and is directly applicable to them, regardless of the national legislation, and the *European Data Protection Board* is tasked to facilitate the consistent application of data protection rules throughout the European Union and promote cooperation

among the supervisory authorities of individual EU Member States. Second, we reviewed supervisory decisions of several countries [88], including over 1400 cases from `https://www.enforcementtracker.com/`. So we are reasonably confident that this scenarios will stand the test of cross-border analysis.

While we do not have an engineering solution for the DPO's problems, at least being aware of the concrete problems is the first step toward a solution.

Chapter 3

Building Cross-language Corpora for Human Understanding of Privacy Policies

Abstract: Ensuring users understand how privacy policies impact them is a key challenge for a real GDPR deployment. Research studies are mostly conducted in English, but in Europe and elsewhere, users speak a language other than English. Replicating studies in different languages requires the availability of comparable cross-language privacy policies corpora. This chapter provides a methodology for building comparable cross-language privacy policies corpora in a national language and a reference study language. It also supplies an application example of this methodology comparing English and Italian, extending the corpus of one of the first studies about users understanding of technical terms in privacy policies. Other open issues that can make replication harder are also discussed.

3.1 Introduction

A well-developed literature exists in relation to the analysis of privacy policies (e.g., [173, 177, 183]), particularly concerning the development of tools to improve the writing [79] and clarity of the policies themselves [57].

Specific studies on users' understanding of privacy policies are instead less frequent (e.g.[95, 142, 166]). While the assumption that existing privacy policies are excessively long and complex is at the foundation of the majority of works on the topic [7, 144, 177,

181], the details of users' misunderstanding of policies are less studied (e.g. [160, 169]). For example, Tang et al. [160] are the first to focus specifically on misconceptions related to technical terms used in privacy policies.

Yet, the ability of citizens to understand what they accept (e.g., 'are they waving their rights without even understanding it?') is a critical issue to gauge the actual success of privacy legislation initiatives such as the European Union General Data Protection Regulation (GDPR) [55]. The introduction of the GDPR in the EU has increased organizations' focus on complying with data protection principles. This privacy-driven approach progressively has grown the complexity of cybersecurity implementation and consequently their costs [94], encouraging the development of GDPR's compliance tools [29], and introducing security workers such as the Data Protection Officer [33] (already existing in the past, but now well-defined and mandatory in all the EU member states). Because an organization must write and comply with privacy policies, it is crucial to assess whether the alleged beneficiaries of the protection granted by these policies (i.e., citizens) can actually understand what they are protected from (and what they are *not* protected from).

A critical issue in this respect is that most studies use English and are based on the U.S. reality (e.g. [149, 160, 173, 178]), which has a different culture as well as differently constructed legislation.

European Union policies are typically written in the national language of a country, and any study that is based on the English language would not accurately reflect the proper understanding of the users: we would be probing at the same time their understanding of English *and* their understanding of the privacy issues. Still, using English as the reference language has advantages, particularly in ensuring that researchers from different countries can access and compare the work.

Given these premises, the main goal is to provide a methodology on how to generate comparable privacy corpora when dealing, on the one hand, with the English language and, on the other hand, with a National language (in this case, Italian).

In this work, we describe how we created such corpora and how to quantify the diversity within the corpora as well as a number of other open issues that makes replication studies way harder than one can initially think of.

3.2 Related Work

Privacy policies have long been the subject of detailed scientific studies; with the advent of the Internet, online privacy policies have proliferated [25, 178] – often in the form of

long, complex and, easily misunderstandable statements – in tandem with the necessity to explain users' data treatment for each online service. This, situation has led, to the necessity (that under some legislation, as in the EU, also became an obligation) to identify and develop methods to write privacy policies more easily, but also to make them more accessible to users of all backgrounds. Within the literature on privacy policies, two major issues are at the forefront of the discussion: the construction of adequate corpora for future analyses, and the development of automated tools for the analysis of existing policies and the drafting of future policies on the basis of a continuously evolving legal environment.

Privacy Policies' Corpora Selection. One of the main issues identified by the literature in proceeding with the (automated) analysis of existing privacy policies for their overall improvement is the lack of appropriate datasets from which to start such analysis [7, 149]. Specifically, the problem for practitioners lies in the selection of policies that represent adequately the great variety in length, complexity and service coverage present among online privacy policies. The situation is further complicated by the presence of different legal backgrounds concerning privacy, for example between the European Union (where the content of privacy policies is mainly determined by the GDPR [55]) and China (where the Personal Information Protection Law (PIPL) sets similar requirements [114]) or the United States (where applicable laws vary between different States or circumstances). These differences can result in different policy structures and contents. As a consequence, the privacy policy of the same company can vary significantly depending on the country from which it is read. An additional factor of variety is the dimension of the company which the privacy policy is referred to, an element which generally impacts both the length of the policy and the frequency with which it is updated [173].

The methods used for policy selection vary between different works. Two main approaches emerge from the literature: the identification of criteria for the manual selection of representative policies (e.g. [173]); and the development of web crawlers to extract the highest possible number of policies available online (e.g. [7]). The two approaches seem to reflect the distinct priorities of different studies conducted on privacy policies, which tend to focus either on the characteristics of the selected policies (a preference for "quality") or on the sheer amount of considered policies (a preference for "quantity"). The second approach seems to be the most common in the literature, though a combined method – establishing a set of quality criteria, often starting from a manual selection and analysis of a few policies generally selected on the basis of

popularity, and using such criteria as a basis of action for a crawler – is also employed often.

Tools for Analysis. A thriving strand of literature is then dedicated to automated tools aimed at, among others, analysing, creating, synthesizing, verifying the compliance or extracting specific elements of interest from privacy policies. The automated analysis of privacy policies has become a necessity both for the organizations writing the policies and for supervising authorities, but also for users requiring new manners to identify and understand the highlights of such policies [49]. Situated at the intersection between legal and technical domains, such analysis has recently turned to machine learning and text mining in order to automate and potentially improve a process that still relies heavily on human contribution [149].

Automatic privacy policy analysis is generally performed through natural language processing (NLP) techniques, which remain the dominant approach in the area. As reported by Del Alamo et al. [49], NLP techniques can be divided into symbolic (or classic) and statistical (or empirical). The first starts from human-developed rules to process the policy's text and model natural language; the second, instead, applies mathematical techniques to previously-created corpora of policies to develop generalized linguistic models. The attention of practitioners is mostly directed to the statistical approach and to the creation of tools enabling its execution in an automatic manner [183, 181].

Readability and Users' Misconceptions. The academic interest over privacy policies is not limited to the manner in which they can be composed. In fact, the efficacy of a privacy policy does not depend only on its adherence to the legal requirements established by the countries of reference – though that is indeed an important and necessary element – but also on how understandable it is by the final recipient of the policy: the concerned website's user.
The capability of users to understand the content of privacy policies is analyzable in more than one way. The majority of the literature focuses on the measure of so-called "readability" [54, 90], which can be calculated according to a series of mathematical formulas or characteristics such as length, language complexity and univocal meaning.

Though definitely less studied throughout the years, a significant portion of scholarly investigation related to privacy policies has concerned users' misconceptions about the policies' meaning and function [142, 166].

3.3 A Reference Corpus in English

To start a reference corpus, we used the study by Tang et al. [160], who tried to investigate the understandability of privacy policies from the users' perspective, focusing in particular on specific technical terms commonly used in data use policies in the context of the USA.

To achieve such an aim, the authors of [160] ran three different studies:

1. A qualitative pilot study to identify commonly misunderstood technical terms;
2. A large-scale main study to test the respondents' understanding of the selected terms and their comfort with some data use practices;
3. A small-scale follow-up study to support the main study's results.

To select the 22 terms to be included in the main study, the authors of [160] created a preliminary list of 57 terms obtained from manual analysis of the Alexa top 10 U.S. websites as of June 2020 (a common practice in studies on privacy policies [142]) and some selected apps from the Android Play Store. The list was then validated through an automated analysis of a 3609 English-language policies corpus to verify the frequency of use of the selected terms. The 57 terms were also divided into 11 categories based on the macro-area they belonged to (e.g., crypto, storage, tracking). From the original study's authors, we have obtained a spreadsheet including the technical terms (58 since one of them was missing from the original appendix) and the categories used in the pilot study. We also received the authors' original notes, where the policies in which they found the technical terms, the contexts of use, and a link to (the current version of) the relevant privacy policies are specified.

In the study by Tang et al. [160], the 57 terms, shown randomly based on their category, were included as part of the pilot study. In the pilot study, respondents from Amazon Mechanical Turk were asked to define the term they were shown. Based on the pilot results, 20 technical terms that were misunderstood the most (of which ten belong to the set of high-frequency terms commonly misdefined, while the others belong to the set of terms that the participants significantly misunderstood) were selected for the main study, together with two mostly well-understood terms. The main study was divided into two sections, asking respondents to answer multiple-choice questions defining some of the 22 terms and rate their comfort with some data use practices on a five-point Likert scale. The ratio of misunderstood to understood technical terms used for the survey is highly unbalanced (20-to-2), which may affect Tang et al. [160] study's

validity. From this perspective, we use them only as control elements to make the study comparable.

Due to the uncertainty connected to the definition of comfort, the follow-up study was then used to verify whether users' attitudes to policies changed using technical versus descriptive terms.

3.4 Methodology

This section describes a usable methodology to build comparable privacy policies' cross-language corpora by mapping the policies of a corpus taken from a reference study into a new one that adopts the language of a future replication study. Figure 3.1 represents this methodology's diagram summarizing inputs, outputs, main characteristics considered, and transformations steps involved. Indeed, a cross-national replication study needs to consider a new corpus of privacy policies comparable with the original one because it will probably not be available or existent in a hypothetical parallel corpus. Hence, in this case [86], verifying the comparability of the two corpora is a precondition for using the new one.

3.4.1 Practical steps

The proposed methodology consists of five steps (see Figure 3.1).

1. **Consider input elements**. The privacy policies corpus of the original study and the replication study language should be considered input.

2. **Checking the comparability of privacy policies**. Consider the individual policies of the study's privacy policy corpus in the original language and see if comparable policies exist in the language of the replication study.

3. **Building of the two new privacy policies corpora**. Build the two new privacy policies corpora (in the language of the original study and that of the replication study) following the high-level process summarized in Algorithm 1 and described in Subsection 3.6. The selection of the policies should be made respecting the criteria of representativeness, homogeneity, and homoscedasticity.

4. **Comparison of corpora and calculation of indicators**. Compare each corpus in pairs (three different comparisons for three corpora: the original corpus and the two new corpora) involving every possible combination of corpora. For each

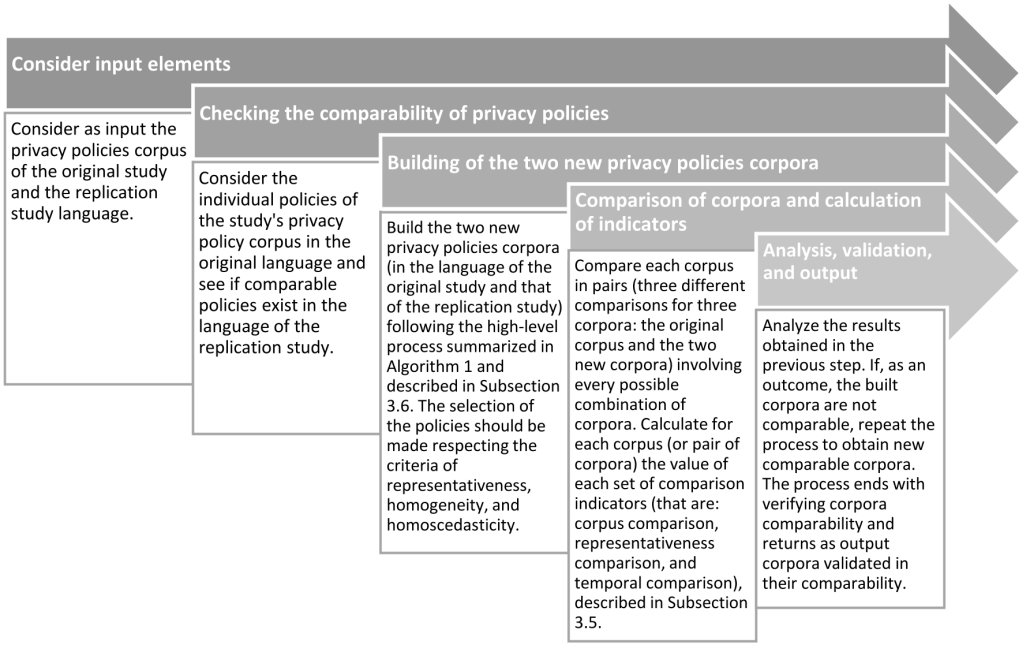

Figure 3.1: The methodology to build comparable privacy policies' cross-language corpora

corpus (or pair of corpora), calculate the value of each set of comparison indicators (corpus comparison, representativeness comparison, and temporal comparison) described in Subsection 3.5.

5. **Analysis, validation, and output**. Analyze the results obtained in the previous step. If, as an outcome, the built corpora are not comparable, repeat the process to obtain new comparable corpora. The process ends with verifying corpora comparability and returns as output corpora validated in their comparability.

3.4.2 Discussion and justification

The comparability check requires at least three different arguments. The first is the object to be compared, the second is another object to compare the first with, and the third is the respect used to compare these two objects. Hence, the third element is crucial, and its choice influences the comparison's result. Therefore, it is necessary to identify some properties concerning which they compare the corpora in comparing two corpora. There is the need to assess the comparability of a corpus with another that is not available since it will be composed of privacy policies used to replace some that are not available in the language of the future replication study. Therefore, before

selecting policies based on the appropriate and specific criteria described below, it will be necessary to consider three data features [86]. These are representativeness, homogeneity, and homoscedasticity. The first feature relates to sharing a property between the two privacy policy corpora. The second feature requires that the data be homogeneous concerning the variables relevant to the purpose of the corpus. The third feature is related to the equal variance of the data across all classes.

Before a researcher can replicate a considered study, it is necessary to ensure that the corpora are comparable. Indeed, constructing multilingual corpora of privacy policies requires considering three different corpora: the *original corpus* in the original study and in the original study's language, a *source corpus* in the original study language and *replication corpus* in the replication language. Hence, it is necessary to make three different comparisons between three distinct corpora:

1. between the original study's corpus and the new source corpus in the language of the original study;

2. between the original study's corpus and the new replication corpus in the language of the replication study;

3. between the new source corpus in the original study's language and the new replication corpus in the languages of the replication study.

Identifying a new source corpus in the original language study is necessary because not all policies are usable/present in the replication language, and therefore, there might not be a correspondence between some/several policies in the original language and some policies in the replication language. As we illustrate here, this may happen because a company or institution in the original corpus does not exist in the replication country, or the company does exist but only have a policy in some languages and not the replication language. Only in very particular cases the original corpus and the new source corpus will coincide.

Usually, the availability of comparable corpora is more significant than parallel ones because they have the only requirement that their component documents cover related content in different languages [97]. Many works in the literature have demonstrated the usability of comparable corpora in various areas; for example, in improving CLIR systems [158] or in bilingual vocabulary extraction [31]. [96] and [152] have shown that an increase in the quality of a comparable corpus offers better performance to applications that refer to it. Moreover, [97] highlighted the need to adopt methods that can qualitatively assess the quality of a corpus. This method allows moving beyond a

naive approach based only on reasoning, which can lead to an a priori unpredictable level of performance.

Because of the problems highlighted in comparing different corpora [86], we make use of solutions for measuring the distance between comparable corpora across languages derived from the one described in [118].

In this instance, we built new corpora of privacy policies (in Italian and English) from the privacy policies considered in the study taken as an example. These policies are accessible from the sites of the organizations that make them available through an interlanguage link (i.e., a link that relates documents on the same topic but written in different languages). This reflection allows us to extend to that case the considerations made in [118] regarding retrieving comparable Wikipedia documents. Furthermore, a manual evaluation of the privacy policies retrieved and included in this new corpora (source and replication) showed that, in many cases, the privacy policies in the language of replication were created from a simple translation of the analogous source policies written in English. Indeed, as is evident from Table 3.1, most companies come from states whose native language is English.

Note that in this case, because the privacy policies are retrieved from the same website (simply choosing the correct language) or from different regional websites of the same company, it is probable that the considered corpora are parallel. However, this is not sure because, due to different legislation (from U.S. and EU), it is not taken for granted that a policy text is the translation of another policy text. Hence, we must assume the only certain fact that, in this situation, the two individually considered corpora are similar in some aspects; that is, they are comparable.

In this case, as suggested in [118], we focused on finding policies with a similar length (difference in word count less than 20%) to increase the probability that they are comparable; that is, they are similar in structure and content. Although most of the policies found have a similar length, there are some exceptions where this condition is invalid. Examples are in the policies of Yahoo, Amazon (IT and U.S., while the state is verified between the IT and NL-EN versions), and the USPS-Poste and Teamsystem-Force pairs (of which there is only one language version of the policies).

A replication gap between the referred example study and the replication study will always exist because some time has passed. This gap has to be manageable and acceptable. To this extent, we identified three groups of indicators to quantify such a gap and which we discuss in more detail in the next section. Then, we analyzed the outcomes provided by these three groups of indicators and qualitatively discussed the real comparability of the corpora.

Finally, having verified the corpora's comparability, we focused on mapping the policies' terms. This activity was made manually and revealed some criticalities described in Section 3.8 which is the ground for future work.

3.5 Comparison Indicators

Based on the discussion in [118], [86], and [97], we propose to use three groups of indicators: *corpus comparison*, *representativeness comparison*, and *temporal comparison*, to qualitatively measure the privacy policies' corpora comparability.

The first indicators' group (*corpus comparison*) measures whether the policies of the original study corpus are comparable with the corresponding (if they exist) policies in the cross-national study language. To compare corpora, we are mainly interested in quantifying how many policies are comparable, replacement (or complementary), and destructured. The indicator *Number of Comparable Policies* measures the number of privacy policies referred to the same company for which there is both a text in the original and replication language. These texts must have a similar (comparable) structure.

The *Number of Replacement Policies* indicator measures the number of policies in which the company from which the original study's policy is extracted does not exist in the replication country language. In contrast, the *Number of Complementary Policies* indicator measures the number of policies for which does not exist a localized version of the policy the original study considered. In the first case, if a product or service of a comparable company exists from the same industrial sector, the policy can be replaced with a new policy of these comparable companies. In the second case, if a localized privacy policy text of a *comparable* company's product or services exists, this policy will be used.

A privacy policy is considered complementary when it is added to a corpus to reflect any updates to a website rankings list (e.g., Alexa TOP10 U.S. or IT) that have occurred since the date of the original study.

The *Number of Destructured Policies* indicator estimates how many policies where a policy in the target language exists; still, this text is not immediately comparable with the text in the original study language due to a significantly different structure. The more destructured policies, the harder any replication study is. To ensure better replicability across languages of a study, the number of comparable policies must be as high as possible.

The second set of indicators measures the representativeness of the corpora. It

provides qualitative criteria to follow in determining how to replace policies in the original corpus that are not usable in the replication study corpora (source corpus and replication corpus).

To measure the corpora's representativeness, we propose to do first a qualitative comparison using a table to compare the rank sources for each corpus. Indeed, it is crucial that the process used to build the new corpora be the same. Analogously, the rank sources used by the different corpora must be distinct from each other (because they are based on the constituent policies' language) and homogeneous (they must use the same source as, for example, Alexa Top10). One possible replacement policy criterion is to investigate if a TOP web ranking used in a study carried out in a particular language also exists in a different language, such that both policies talk about the same terms.

For example, it is possible to consider the Alexa TOP10 website list in both languages. Another example is considering the ranking of top companies in a specific industry (e.g., the top banks) for each state to which the language used in the studies refers.

Finally, to consider the replicability of the original study, it needs to assess the temporal comparability of the corpora (third indicators' group). We evaluated qualitatively three indicators to measure this comparability: the *Temporal Internal Consistency*, the *Temporal Replication Gap*, and the *Qualitative Replication Gap*.

We used the *Temporal Internal Consistency* indicator to ensure that the source and replication corpora policies are all in the same narrow interval. This indicator measures within the same body of privacy policies the number of months between the publication date of the most recently updated policy and the one of the least recently updated policy. The *Temporal Replication Gap* indicator aims to highlight if the policy has changed. It measures the average number of months of the update time of policies in the source and replication corpora vs. the ones in the original corpus. This value is calculated only for the comparable or destructured policies because it has no sense for the other ones (complementary or replacement). Indeed, these policies are added contextually to the new source and replication corpora and are absent from the original one. Since the study may give different results because time has passed and many things have happened (e.g., changes in laws), the *Qualitative Replication Gap* indicator is used to mark if major events happened from the original study and its replication.

3.6 Reconstruction of policy corpus

In the initial phase, we focused on looking for the correspondence between the Italian and English policies of the new cross-national source and replication corpora and the English policies of the original corpus we took from the example study of Tang. et al.[160]. Hence, excluding the case in which the policies are comparable, we identified three cases that require attention.

1. Missing policy: there is no Italian policy; however, there is a corresponding Italian app or website.

2. Missing company: there is no corresponding Italian app or website.

3. Destructured policy: there is a corresponding Italian version of the original privacy policy, but we found that this has an entirely different structure.

Henceforth, a team of two researchers analyzed the original corpus of policies and found correspondence rules to build comparable replication and source corpora of Italian and English ones. After defining the policy replication corpus for the Italian study, we concentrated on analyzing whether its policies were structurally similar to those in English of the original corpus. The high-level process to build the new replication and source corpora is presented in Algorithm 1. All the operations listed in the algorithm are done manually by humans. When in Algorithm 1, the condition *ItalianCompanyMissing* is verified, there is a case of a replacement policy. Analogously, if the condition *ItalianPolicyMissing* is satisfied, there is no Italian policy, but there is a corresponding Italian app or website. Hence, this is a case of a complementary policy. Instead, if the condition *ItalianPolicyDestructured* is verified, an Italian policy exists for a specific company. However, it has an entirely different structure from the original English policy (case of destructured policy).

If we need to substitute the English company with another of the same type (for example, another bank will replace a bank), we choose one that operates globally. Analogously, in case we need to substitute the missing privacy policy, we look for another related to a comparable company for the business sector, market segment, and type of product or service. With these choices, we have favored companies operating in the USA and Italy.

We performed a manual analysis of the Alexa top 10 Italy websites as of November 2021, and analogously, we analyzed selected apps that, in the same period, had ranked better in the "most profitable games" category of the Play Store for Italy. After that,

Algorithm 1 Look for the corresponding Italian privacy policy entry

```
1:  procedure ITALIAN POLICY CHECK(EnglishPolicyURL)
2:      Retrieve English privacy policy text from EnglishPolicyURL
3:      Read English privacy policy text
4:      Retrieve English company from EnglishPolicyURL
5:      Read English company
6:      Search for the correspondent Italian company
7:      if ItalianCompanyMissing = True then
8:          Choose a new company with the same type as the original company and a privacy policy in Italian and English
9:          Retrieve URL and text of Italian policy of new company
10:         Retrieve URL and text of English policy of new company
11:         Add new company's English policy to new English corpus as replacement
12:         Add new company's Italian policy to Italian corpus as replacement
13:     else
14:         Search for the corresponding Italian privacy policy of the original Company
15:         if ItalianPolicyMissing = True then
16:             Choose another privacy policy in Italian related to a company comparable to the original company
17:             Retrieve the URL and the text of the Italian policy of the new comparable company
18:             Add the original company's English policy to the new English corpus
19:             Add the new comparable company's Italian policy to the Italian corpus
20:         else
21:             Retrieve the URL and the text of the original company's Italian policy
22:             Add the original company's English policy to the new English corpus
23:             Add the original company's Italian policy to the Italian corpus
24:             Analyze the Italian privacy policy structure
25:             Compare the Italian policy structure with the English one
26:             if ItalianPolicyDestructured = True then
27:                 Operate a manual measures activity of the privacy policy
28:             else
29:                 Classify the Italian privacy policy as comparable
```

we compared these lists with the analogous ones (that refer to June 2020) in the original study for the U.S. privacy policies. To overcome the 17-month time gap between the establishment of the original corpus and the replication corpus, we made a further comparison by considering the content of the Alexa Top 10 U.S. ranking in the language of the original study as of November 2021.

From the original study, we considered 58 URLs obtained from 17 different companies' privacy policies. After this review phase, we removed six companies' privacy policies; however, we added twelve. This addition allowed us to improve the size of the corpus of policies, especially its specialization in the Italian reality and its focus on EU GDPR [55] concepts. We summarized these activities in Table 3.1. The process in detail was as follows

1. the privacy policies of the apps *Infinite Word Search* and *Woody Puzzle* have been replaced by those of the apps *Coin Master* and *Empires & Puzzles* that, in addition to having an Italian privacy policy, respectively (on November 23, 2021) ranked first and fourth in the "most profitable games" category of the Play Store for Italy;

2. the privacy policy of the app *Signal* is replaced respectively by the policy and by

the FAQ documentation of the apps *Whatsapp* and *Telegram*. The *Signal* app has an Italian version but not an Italian privacy policy, so we have chosen to replace it with other messaging apps with similar characteristics, such as support for end-to-end encryption. Moreover, because for the app *Telegram* there is not an Italian privacy policy but only an Italian FAQ documentation section in which the considered technical terms are present, we referred to this documentation;

3. the privacy policy of the website *Bank of America* is replaced by the *Unicredit* one, that is a company of the same category and operating worldwide;

4. the privacy policies of the *Reddit* and *Verizon* websites are replaced by the *Vodafone* one, that is a company that provides analogous services;

5. starting from one of the top 10 Alexa U.S. websites on November 2021, we added both the privacy policy of the related company and the one connected to a comparable Italian company (in particular the *Force* U.S. company and the *Teamsystem* Italian one);

6. starting from one of the top 10 Alexa Italian websites on November 2021, we added both the privacy policy of the related company and the one connected to a U.S. comparable company (in particular the *Poste* Italian company and the *USPS* U.S. one);

7. we added the privacy policies of the companies *Zoom* and *Microsoft* which have a website that appears both in Alexa top 10 Italian and in Alexa top 10 U.S. We chose policies from websites that appeared in the Alexa top 10 rankings for IT and U.S. to maintain equivalence between the original corpus, the new source corpus of English policies, and the new replication corpus of Italian policies. The original study's authors built their English technical terms corpus by manually analyzing the Alexa top 10 U.S. websites.

3.7 Results and Comparison

In this example, we made three different comparisons between three different corpora to determine their comparability level. The first corpus is the English-language corpus, the original used in the study by Tang et al. [160]. The other two are cross-language corpora built (one, the source corpus, in English, and the other, the replication corpus,

Table 3.1: Composition of privacy policies corpus

This table describes the compositions of the original privacy policies corpus and the new English and Italian source and replication corpora. Columns named *Orig. Tang et al.*, *New US*, and *New IT* refer to different privacy policies corpora. The first entry refers to the English-language privacy policies corpus analyzed in the original Tang. et al. [160]study, while the second and third entries refer to the new privacy policies source and replication corpora in English or Italian, respectively. The *Referred company* field shows the company (and eventually their product) whose privacy policy we considered. Finally, the *Notes* field summarizes the rationale behind choices in selecting the privacy policies.

Referred company	**Orig. Tang et al.**	**New US**	**New IT**	**Notes**
Yahoo	√	√	√	
King Games (app Candy Crush)	√	√	√	
LinkedIn	√	√	√	
Amazon	√	√	√	
Google	√	√	√	We maintained it from initial study.
Apple	√	√	√	
Facebook	√	√	√	
Wikipedia	√	√	√	
Twitter	√	√	√	
Ebay	√	√	√	
Firefox (documentation)	√	√	√	
Random Logic Games (app Infinite Word Search)	√			We substituted policies because they do not exist in their Italian version.
Athena FZE (app Woody Puzzle)	√			
Moon Active (app Coin Master)		√	√	Companies added from the PlayStore's rank on Nov. 2021.
Zinga (app Empires & Puzzles)		√	√	
Signal	√			We substituted the policy because it does not exist in its Italian version.
Telegram (FAQ doc.)		√	√	Companies added to replace the *Signal* one.
WhatsApp		√	√	
Bank of America	√			We replaced the policy because the respective Italian companies do not exist.
Unicredit		√	√	Company added to replace the *Bank of America* one.
Reddit	√			We substituted companies with new ones focused on Italian reality.
Verizon (cookie policy)	√			
Vodafone		√	√	Company added to replace the *Reddit* and *Verizon* ones.
Zoom		√	√	Policies added from the top 10 Alexa Italian and U.S. on Nov. 2021.
Microsoft		√	√	
USPS (only US)		√	√	It was added because it is a U.S. company with similar features and products to the Italian *Poste*.
Poste (only IT)		√	√	Company added from the top 10 Alexa Italian on Nov. 2021.
Teamsystem (only IT)		√	√	It was added because it is an Italian company with similar features and products to the U.S. *Force*.
Force (only US)		√	√	Company added from the top 10 Alexa U.S. on Nov. 2021.

in Italian, which is the language of a potential replication study) from the first corpus. We aim to use them for a potential replication study on how humans understand privacy policies.

Comparing the English-language original corpus from Tang et al. [160] with the source corpus built by us in the same language, we have 52,38% comparable policies, 38,10% complementary policies, 9,52% replacement policies, and no destructured policies. The results are worst when comparing the original English-language corpus and the new replication one in Italian. In that case, we have 47,62% comparable policies, 38,10% complementary policies, 9,52% replacement policies, and 4,76% destructured policies.

In contrast, comparing the two new source and replication corpora in Italian and English, we find these are well aligned. We have 85,72% comparable policies, 9,52% complementary policies, 4,76% destructured policies, and no replacement policies. In all comparisons, the number of destructured policies is low, if not wholly absent.

Hence, rather than using Tang et al.'s [160] privacy policies' source corpus for a cross-language comparison, we compare the two new source and replication corpora (built from the original one) that are more closely aligned.

Moreover, internal temporal consistency exists in the new source and replication corpora policies. Indeed, these policies are all in the same narrow interval because their release varied from June 2020 to September 2021. Furthermore, with few exceptions, the Italian and English versions of these policies were published simultaneously. The exceptions to this are Amazon (where the Italian version is December 2020 while the English version is February 2021), Facebook (with versions varying from August 2020 to January 2021), WhatsApp (varying from January 2021 to November 2021), or Vodafone (varying from July 2021 to November 2021).

For a correct calculation of the value of the "*temporal replication gap*" indicator, it is essential to know the version and date of each privacy policy considered (both in the original corpus in the original study and those in the source and replication corpora in the replication study). Knowing the number of intermediate versions of privacy policies for each company is also helpful. In this example, there are many problems, and this indicator is not helpful. The original study [160] did not specify the version and date of policy updates. We deduced the date of the considered policies by cross-referencing the updating date of the policy with the date of consultation of the TOP 10 of Alexa U.S. (June 2020). We know that this date is incorrect, but it still gives us an indicative idea. Moreover, considering that not all companies expose the historical archive of policy versions, we can obtain a policy date to calculate the indicator in only

7 out of 11 cases. Other companies have made some intermediate versions of the privacy policy disappear from their website (for example, on the Facebook website, the previous versions are no longer available between September 9, 2016, and January 4, 2022). This fact generates another interesting problem because users have given their consent based on a no longer existing policy. We have not considered the case of Facebook because it presents an outlier value. Therefore, we used 6 out of 11 policies in calculating the indicator. Interestingly, with only one exception (Google, for which there were four different versions in the face of a 17-month time gap), between the two studies, there was at most one version gap in the privacy policies considered. We calculated the temporal replication gap value only for the comparable or destructured policies because it has no sense for the other ones (complementary or replacement). Indeed, these policies are added contextually to the new source and replication corpora and are absent from the original one. The final value obtained for this indicator shows that the average number of months of the update time of the policy (comparable or destructured) is 16,5. This result is good because it shows that from the policies version available in the example study's original corpus and the new ones (source and replication corpora) built by us, there is only one version difference and a time lag between updates of just over a year.

Considering the qualitative replication gap indicator, it shows that in the time between the construction of the original corpus and of the pair source and replication corpora (June 2020 - November 2021), there have been no significant regulatory developments. For example, both the *GDPR* and the *California Consumer Privacy Act (CCPA)* predate the construction of the original corpus, while other laws, such as the *Connecticut Data Privacy Act (CDPA)*, the *Utah Consumer Privacy Act (UCPA)*, and the *Virginia Consumer Data Protection Act (VCDPA)* was not yet in effect. Despite this, there have been some changes at the corporate level (e.g., the corporate change from *Facebook* to *Meta*) or at the policy level (e.g., additional information added regarding data retention in the Yahoo policy).

3.8 Open Problems: Technical Terms May Not Match

After building a privacy corpus in a different language, the next step is replicating different studies. For example, the authors in [160] arranged a survey based on 22 key terms derived from the pilot study's results to investigate the understanding of technical terms.

We experienced many cases of technical terms that do not match and occur with extremely different frequencies.

The most egregious case are the English terms "*personal information*" and "*personally identifiable information (PII)*" (that were distinct in the considered study). Hence, we used a syntactic criterion both to be sure of their overlapping and overcoming it in a single Italian term, "*dati personali*." In practice, we used a syntactic rule to identify the possible syntactic distinction of these terms. Hence, they are distinguishable if there is no correspondence between the two policies (English and Italian) or if there are orphans or widows. Otherwise, it is possible to conclude that they are the same thing and are codifiable with the same term in Italian.

To understand if the terms PII@US (*personally identifiable information*) and PI@IT (*informazioni personali*, *personal information*) are equivalent, we manually looked for widows' or orphans' presence related to these technical terms. An *orphan* is a term that appeared in the Italian policy without a match in the corresponding clauses or sections of the English policy. A *widow* is any occurrences in the English policy in which the term PII@US (or the term SPI - *sensitive personal information* -) appears without the term PI@IT in the Italian policy.

We also looked for implicit terms or pronouns that refer to technical terms. We found very few occurrences of those cases.

Other critical issues occur when technical terms do not appear or are rare in these corpora. For example, the term *fingerprinting* does not appear in the Italian corpus and is useless for an Italian language survey. Instead, *public information* and *browser web storage* terms are rare in these corpora, but for an Italian survey, they are significant and can be used.

We illustrate these issues by considering the 22 terms from [160]:

- two high-frequency terms that most users in the pilot study correctly defined;
- ten high-frequency terms that the study's participants commonly misdefined;
- ten additional terms for which the study's participants exhibited significant misunderstandings.

Using the proposed new source and replication corpora of Italian and English policies, we investigate the differences between the frequency of occurrences in privacy policies of the most common Italian and English technical terms (the TOP 10 and the TOP 22 lists). More details are listed in Table 3.2.

The correspondence between the TOP 10 English list and the TOP 10 Italian one is generally reasonable. 80% of the terms included in the TOP 10 English list also belong to the TOP 22 Italian one (with a peak of 70% of terms also included in the Italian

Table 3.2: TOP 10 and TOP 22 lists of more frequent terms in the new Italian source corpus of privacy policies

This table describes the TOP10 and TOP22 lists of terms that appear most frequently in the new Italian source corpus of privacy policies. We used boldface to represent the terms belonging to the TOP 10 list. The order (*Rank IT* column) of technical terms in the table follows the terms that appear most frequently in Italian language privacy policies. The table also shows the ranking of the most frequent technical terms in English language privacy policies (*Rank U.S.* column). Corresponding to the Italian-language term (*Term IT*) is the corresponding English-language term (*Term US*). Finally, the table shows the cardinality with which we detected the term in the new replication and source corpora of Italian-language (*#Freq IT*) and English-language (*#Freq US*) privacy policies, respectively.

Rank IT	**Rank US**	**Term IT**	**Term US**	**#Freq IT**	**#Freq US**
1	1	**Dati Personali**	**Personal information**	1523	1430
2	12	of which	**Personally identifiable inform.**	1012	79
3	2	**Cookie / Marcatori**	**Cookies**	645	784
4	4	**Terzi / Terze parti**	**Third parties**	618	360
5	3	**Informativa sul trattamento dei dati personali**	**Privacy policy**	471	540
6	14	**Rettificare**	**Correct**	170	60
7	7	**Indirizzo IP**	**IP address**	136	153
8	26	**Disattivare**	**Deactivate**	128	25
9	9	**Cifratura / Crittografia**	**Encryption**	86	113
10	5	**Dati sull'account**	**Account information**	60	324
11	56	Hash crittografico	Cryptographically hashed	52	2
12	36	Transferimento di dati	Data transfer	50	15
13	10	Affiliate	Affiliates	48	87
14	6	Pubblicità mirata	Targeted ads	47	157
15	15	Identificatori univoci	Unique identifiers	45	54
16	20	API/SDK	API/SDK	44	44
17	16	Dati relativi all'ubicazione	Location-related information	42	50
18	18	Informazioni sui pagamenti	Payment information	41	47
19	40	Resi anonimi	Anonymize	40	13
20	46	Operazioni del dispositivo	Device operations	39	6
21	29	Categorie particolari di dati personali / Dati sensibili	Sensitive personal information	34	19
22	19	Dati di utilizzo	Usage data	28	45

TOP 10). Hence, only 20% of terms in the TOP 10 English list do not belong to the TOP 22 Italian one. The correspondence between the TOP 22 English list and the TOP 22 Italian list is acceptable. 58,33% of terms in the TOP22 English list belong to the Italian one, and 16,66% of them also belong to the TOP10 Italian list. Hence, only 41,67% of terms in the TOP22 English list are outside the TOP22 Italian one.

Out of 22 terms used in the Tang et al. [160] example study's privacy policies English original corpus, ten of these (45,46% of the total) appear in one of the TOP English lists of the new English source corpus (specifically, five in the TOP 10 and the other five in the TOP 22). While the remaining 12 terms (54,54% of the total) are outside the source corpus TOP 22 English list. This outcome means that referring to the source corpus of privacy policies, more than half of the technical terms considered in the study taken as an example are rare. This result is crucial because the proposed source and replication corpora are more focused on EU reality.

The situation is worse, considering the terms used in Italian policies. Only 7 (31,82% of the total) of these terms appear in one of the TOP Italian lists (in particular, five in the TOP 10 and the other two in the TOP 22). The remaining 15 terms (68,18% of the total) are outside the TOP 22 Italian list.

Hence, the technical terms used in the example study's survey by Tang et al. [160] are rarely used in the privacy policies in Italian. This consideration is a crucial point that shows the need first to map technical terms and, secondly, to adapt the survey for replicating it to Italian-speaking respondents. These activities will be the subject of future research work.

3.9 Conclusion and Future Works

The first conclusion of this work is that a strict replication study in a different language is possible only up to a certain extent because some elements do not transfer from the original study to the new one.

Therefore, this contribution is to discover and describe how elements (for example, the original corpus used in the survey) can be adapted from an original study to a new one to understand what is needed to maintain relative comparability of the components of the study, such as the privacy policy corpora.

Still, the results obtained in this case study confirm that by taking due care, it is possible to build cross-language source and replication corpora of privacy policies usable for research purposes, such as investigating how humans understand their content.

The outcomes of this work will be the base for further research activities. Future interesting work is to investigate the validity of automated analysis approaches, i.e. whether just looking for term occurences in a policy would yield a correct analysis. Precision and recall values achievable in the automated coding of technical terms should be compared from those derived by manual coding. A second aspect is mapping technical terms in different language corpora privacy policies. The preliminary analysis has already shown that this is not so obvious. Future work could define some indicators to identify the irrelevant, infrequent, overlap, and unreliable technical terms that must be removed or replaced according to the case. This information will help, for example, to replicate the understandability questions from Tang et al. [160] to Italian-speaking people who interact with Italian privacy policies.

Dataset Availability

The datasets generated and analyzed during the study are archived on Zenodo at `https://doi.org/10.5281/zenodo.7729546` [35], while Appendix H displays a description of the dataset characteristics and composition.

Chapter 4

My terms are (not always) your terms: technical terms in privacy policies in Italian and English

Abstract: This chapter proposes a methodological approach consisting of the following steps: (i) a tool-supported match of technical terms in cross-language privacy policy corpora; (ii) a manual assessment of the meaningfulness of the translations of technical term change in cross-language privacy policy corpora; and (iii) the identification of the occurrences of mapped terms and, thus, the different relevance of the terms. The final result is a cross-mapping of 58 technical terms identified in previous work on 81 comparable publicly available Italian and English privacy policy corpora, obtaining over 4000 occurrences of these terms. While it is broadly possible to compare the terms of privacy policies across English and Italian languages, there are also not negligible differences, as technical terms have different frequencies of occurrence and meaningfulness across languages. Relying on the identification of terms by regular expressions produces only a partial picture. Although precision can be improved by adjusting the mapping of terms with appropriate synonyms (as opposed to a dictionary translation), recall is harder to improve.

4.1 Introduction

The study of how well privacy policies successfully inform users about the processing of their personal data depends on whether and to what extent people can understand

the content of such policies [99, 168]. Although previous work (e.g. [99, 168]) reveals that different cultural, social, and linguistic backgrounds may lead to a different understanding of privacy policy, the adopted technical terms and their perception play a relevant role too. Technical terms are words or locutions denoting concepts and tools that are specific to a given domain, here privacy. As reported for example in [62] e [159], such a terminology is the natural target of privacy studies.

Most studies on privacy policies understanding routinely use English (e.g., [149, 159, 177, 178]). Choosing English facilitates, indeed, the comparison of analytical findings by offering, however, only a partial view of reality. In their work, starting from a set of 35102 unique URLs related to privacy policies, Linden et al. [102] obtained 4909 English privacy policy URLs, which means a loss of 92% of the available data. According to Mhaidli et al. [113] only a few works focused on non-English written policies. Among them, Arora et al. [9] provides a bilingual corpus of app privacy policies in German and English, whereas Ciclosi et al. [34] built a cross-policy corpus in Italian and English. Mhaidli et al. [113] also pointed out that the lack of tools and resources to study non-English written policies is one of the major open issues that researchers are facing nowadays.

This chapter investigates the interplay between language and technical terms, as relevant information might be lost in translation.

In particular, given English as the *source language* and Italian as the *replication language*, we focus on how English technical terms are mapped onto Italian technical terms. The research questions addressed by this chapter are as follows:

RQ2.2 How is a tool-supported approach for key technical terms applicable in cross-language privacy policy corpora, such as English and Italian?

RQ2.3 Does the number of occurrences of technical terms and their meaningfulness change across policies written in different languages (i.e., English and Italian)??

RQ2.4 How can an English study be replicated using another language, such as Italian?

RQ2.5 Does understanding technical vs. non-technical terms change across policies written in different languages (i.e., English and Italian)?

We adopted the corpora in [34] whose data are publicly available at the URL `https://doi.org/10.5281/zenodo.7729546`. For the technical terms, we contacted the authors of [159] and we obtained the full list of the technical terms, the categories used in the pilot study, the specification of the privacy policies in which they found the

technical terms, the contexts of use, and a link to (the current version of) the relevant policies.

4.1.1 Artefact availability

All material used in this chapter (i.e., manual and automated annotations of the privacy policies, lists of the regular expressions used in the "auto-coding" activity, classifications of terms as true or false positives, locutions or synonyms) is anonymously available for reviewers at `https://anonymous.4open.science`. After acceptance it will be made available on Zenodo. Some tables illustrating detailed examples of the most frequent technical terms are in the appendix of the chapter.

4.2 State of the art

In a nutshell, understanding a policy is hard because several factors (from legislation to choice of wordings) may introduce ambiguities [143]. In the following, a short description of these factors is reported.

GDPR's induced complexity. The introduction of the GDPR has resulted in a significant overhaul of privacy policies, which have become significantly longer and, therefore, more difficult to understand [102, 25, 178]. Even when automated tools are used for policy generations (e.g., [182]), they are not necessarily easy to read nor they do correspond to the actual practices of the systems they describe (e.g. [92, 67]).

Natural language structural ambiguity. Even without GDPR's added complexity, privacy policies are drafted using natural language that can lead to users' misunderstanding [62]. Natural language defects are also issues in other fields. Dalpiaz et al. [47] discovered that they lead to misunderstanding in requirement engineering [15, 17]: ambiguity is present if there are different terms referring to the same denotation (*correspondence*) or if the same term is used for different denotations (*conflict state*). Incompleteness occurs when an item refers to something not appearing in the policy (*missing requirements*). In the privacy domain, the incompleteness of policies could induce users to make inaccurate predictions about personal data processing, resulting in a potential wrong estimation of personal data risks and processing acceptability [18, 16]. To address this problem, Guaman et al. [67] proposed to use annotations that identify different statements referring to a single concept.

Privacy policy choice of words. In privacy policies sometimes different technical terms are used to describe the same kind of processing. For instance, in [18], *collection*

is described using the following terms: *logging*, *submission*, and *gathering*. The choice of a specific term is not neutral because subtleties in meaning can affect users' understanding. For instance, *submission* relates to a user's active action, while *logging* relates to an automatic activity carried out during a user's action. Bhatia et al. [18] found that many privacy policies do not specify to users the purpose of data operations, and this absence is a concrete compliance issue for some jurisdictions like the EU GDPR.

Users understanding of privacy policies. As a result, users studies (e.g. [54, 90]) show that privacy policies are generally affected by bad readability [157, 22] and rarely read by users. Kohavi et al. [85] shows that less than 1% of users click on privacy policies. Analogously, Steinfeld et al. [153] discovered that users tend not to read a privacy policy if it is possible to accept the website's terms and conditions without reading it. As an alternative to reading, Tesfay et al. [163] introduce the summarization tool *PrivacyGuide* to support privacy policies' readability. Two under-investigated fields are: i) the users' misconceptions of policies' meaning and function (i.e., [142, 165]); ii) the users' understanding of the technical terms used in policies (i.e., [159]).

Consequences of Privacy policies' multilingualism. EU Member States have a common data protection regulation but 24 official languages and many micro-enterprises only write privacy policies in their local language. This gap is significant since, for example, in the EU, these small entities employ about 90 million people (about 2/3 of the EU-27 workforce) [62]. The absence of non-English policies might not be felt in a first instance, as many studies (e.g. [173] and [162]) rely on Alexa ranking considering only the most popular websites. Popular sites are then likely to also have an English policy that could be used.

Yet, each language has its syntactic structures (and characters) that impact the policies' text [113]. Indeed, some works [58] [18] show that the meaning of a technical term could be understood only in conjunction with the information that relates to it. As a mitigation factor, automatic translation systems exist, as well as training resources (e.g., the *Europarl Corpus* [84], which is based on a multi-language text collection from the proceeding of the European Parliament). While analyzing the policies of all countries under the GDPR's jurisdiction is clearly an important task, it is unclear how to systematically do it [113]. A possible solution to this issue is using comparable cross-language privacy policy corpora. For instance, Ciclosi et al. [34] provide an example in Italian and English, Arora et al. [9] show another one in German and English.

Analysis of Privacy Corpora. Three principal approaches are commonly used. The first approach consists in the definition of criteria for manually selecting the most

representative policies [173, 34]. The second approach develops web crawlers to extract and analyze online privacy policies automatically [7, 78], and use the results for automated analysis of them [49, 149]. The last approach, such as [102], is a combined method that first defines quality criteria and then use them as a basis for an automatic crawling activity (balancing *quality* and *quantity*). In this study, we adopted this third approach.

4.3 Privacy policy corpora

Several privacy corpora exist. Some of them being annotated offer concrete opportunities for semi-automated privacy policy processing [62]. Among these annotated corpora, the work in [102] concerns EU and worldwide privacy policies; Guaman et al. collected 100 English policies ([67], [2]) derived from the APP-350 corpus [184] and from popular Android applications. This annotated corpus indicated cross-border transfer practices [68]. Another example is For instance,the PriBot application [73], a free-form question-answering system, is based on the annotated dataset OPP-115 [173]. Other works provide meta-data related to policies. For example, the works in [100, 101, 83] include a JSON hierarchical list of third-party domains and their owners.

Unfortunately, the outcomes of some works [7, 149] have discovered a lack of policy corpora helpful to conducting these analyses.

Timeliness is also an issue that must be considered to account for the impact of the GDPR: Degeling et al. [48] measured such difference in websites' privacy policies before and after the GDPR enforcement date. Also Gallé et al. [62] highlighted that using algorithms trained on corpora existing before the GDPR adoption could introduce some significant bias because GDPR presents new specific elements not commonly addressed by previous corpora. For instance, the OPP-115 [173], and the Tesfay et al. [162] datasets do not include annotation about the data protection officer, a mandatory crucial and ubiquitous security role [33].

In this work, we adopted the policy corpora described in Ciclosi et al. [34] and the technical terms reported in Tang et al. [159]. Such corpora are, indeed, publicly available and interrelated as depicted in Figure 4.1.

Tang et al. investigated the understandability of privacy policies from the users' perspective, focusing on specific technical terms commonly used in U.S. policies. The outcome of their work is a list of 58 technical terms clustered into 11 categories (e.g., crypto, storage, tracking). This list was validated by automatically looking for unigrams, bigrams, and trigrams frequency counts in a corpus of 3609 English-language

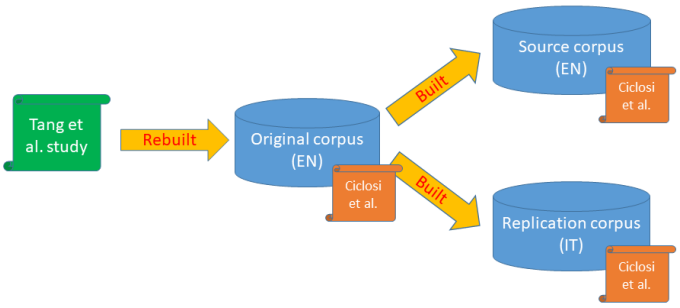

Figure 4.1: Relationships and genesis of the three corpora of privacy policies from Ciclosi et al. [34]

privacy policies [102] Then, this list was used to build a users survey with questions on 26 technical terms.

The source and replication corpora of Ciclosi et al. are built from the corpus of Tang et al. They are composed of 81 paired policy texts in Italian and English from 2018 to 2021, and they address some relevant problems issues for cross-language comparisons. More specifically: (i) a policy might have multiple versions in time and the versions in the different languages must be aligned in time as well; (ii) part of the text of a policy might be automatically generated with Javascript and or be presented in different format and these inconsistencies must be first resolved; (iii) some companies (and the related policies) might be very popular in the international rankings but might be obscure in a country and must be replaced by a nationally popular company (with its induced policy).

Table 4.1 summarizes some characteristics of the Ciclosi's corpora and of two restricted sets of policies from the source and replication corpora we used to build the ground truth (see Section 4.5).

From now on, we will refer to the English and the Italian policies in Ciclosi et al. as the *source* and the *replication* corpus, respectively . The English policies in the corpus from Tang et al. are part of the *original* corpus.

To make this chapter self-contained, we report in the Appendix, Table J.1 the composition of privacy policy corpora of Ciclosi et al. and the technical terms by Tang et al.

Table 4.1: Some digits of corpora of privacy policies

The policies written in English in Ciclosi et al. [34] make the *source* corpus and the corresponding policies written in Italian make the *replication* corpus. The English policies in the corpus from Tang et al. [159] from which the first two were extended are part of the *original* corpus. The Table also describes the two sets of policies used for the initial ground truth. Since six privacy policies belong simultaneously to the original and source corpus, the sum of the distinct policies of the three corpora is 81.

Analyzed corpus	**#Policies occ.**	**#Terms occ.**	**Language**
Source Corpus	32	4339	English
Replication Corpus	28	5841	Italian
Original Corpus	29	3223	English
Source Corpus (ground truth)	6	662	English
Replication Corpus (ground truth)	3	460	Italian

4.4 General overview

As a takeaway, based on findings, we used the corpora of Ciclosi et al. [34] to replicate the Tang et al. [159] study. Before administering surveys, we adjusted the questions according to the outcomes described in Section 4.6.2.

4.5 Methodology: Finding the Terms

Once the policies were selected, we set up a *ground truth* to compare the outcomes of manual coding (as described in Section 4.5.3) with those of automated coding (as described in Section 4.5.6)

4.5.1 Understanding structural similarity

To build the ground truth, two researchers selected the following two subsets of policies from the source and replication corpora to study the structural similarity between the privacy policies in English and those in Italian. The similarity analysis was carried out manually and was aimed at comparing the structure of policies drawn up by the same company but written for people who spoke different languages and lived or resided in other countries. The manual coding, described in the following, allowed us to compare policies drawn up by companies operating in the same sector (i.e., banking) but in different countries.

Appendix K contains more details regarding the composition of the two subsets.

4.5.2 Understanding terms similarity

The next step was to determine the most suitable Italian translation of each English technical term appearing in the work of Tang. For example, the term *key* in the source corpus assumes the following meanings in the replication corpus: cryptographic key, database key, and the adjectival meaning of crucial. Hence, we chose not to use the Italian translation *chiave* in all the cases, because it could result in polysemous misunderstandings. Instead, we adopted this term only when it was referring to the cryptography's semantic area (as *scambio di una chiave di sicurezza* (i.e. *exchange of a security key*), in Unicredit's Italian privacy policy). Further issues come from the different legal regimes on personal data protection existing in the U.S., to which the privacy policies of the source corpus comply, and in the EU, to which the privacy policies of the replication corpus comply.

The fundamental intuition behind the process is to map *both* the structure and wordings of the paired policies (written one in the source and another in replication language), looking for evidence that reveals the same meaning between the terms. An example of this evidence is the presence of orphans and widows.

Moreover, a technical term may also include a notion that is a specialization of a root concept. For example, the technical term *sensitive personal information* is a subset (sometimes used as a descriptor) of the special categories of personal data, according to the GDPR terminology.

Appendix L details the correspondence mapping process and the orphan's and widows' recognitions.

4.5.3 Coding procedure

Following the guidelines for applied thematic analysis reported in Guest et al. [69], two researchers manually coded the two chosen subsets of privacy policies independently. A third researcher helped the colleagues to solve possible disagreements on the annotations. This work was carried out using the Atlas.ti tool [1]. Finally, the researchers built two synonyms' tables: one in the source and the other in the replication language. These tables were constructed by listing all the synonyms and locutions found during the manual coding. A single list was built for each technical term.

Appendix M describes the algorithms and code table used.

[1] `https://atlasti.com/`

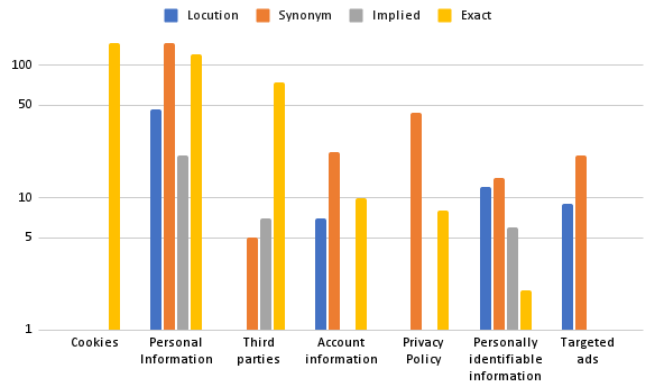

Figure 4.2: Number of occurrences (in logarithmic scale) of the most frequent technical terms appearing as exact, synonym, locution and implies in the two sets extracted from the source corpus as described in Section 4.5, 4.5.1

.

4.5.4 Coding outcomes

Two synonym tables, listing terms, synonyms, and locutions, were obtained. The entries of these tables are terms that appear in both language versions of policies. Table N.1 of Appendix N shows an excerpt of the synonym tables. The synonym tables will be used later to build regular expressions (both in source and replication language) to carry out the "auto-coding" step.

Figure 4.2 shows a graphical representation of the coding results in the source language onto privacy policy belonging to the two validation sets built from the source corpus of Ciclosi et al. [34]. In particular, it shows excerpts of the technical terms' occurrences, specifying whether each is *exact*, a *synonym*, a *locution*, or only *implied*. We chose the technical terms to show in the graph by sorting the list of terms in descending order according to each of these four dimensions and then selecting the top three most frequently. For some terms, synonyms may occur more frequently than the actual term (e.g., *targeted ads* or *account information*). Other terms (like *cookies*) occur almost always in an exact form.

Interestingly, some Tang's technical terms do not explicitly appear in the analyzed privacy policies. Instead, they do so implicitly using different synonyms or locution. For example, no unique term is used in the analyzed privacy policies of cross-language corpora from Ciclosi et al. [34] to express the concept related to the technical term *device signal*. However, it is referred to implicitly by using expressions such as: "*connectivity data*," "*Wi-Fi information*," and "*Bluetooth signals*." In other cases, it refers to this term using a locution, like "*information about your Internet service provider*."

4.5.5 Constructing regular expressions

We used regular expressions built from the synonym tables' content (see Appendix N, Table N.1 for an excerpt). These expressions aim to automatically detect all occurrences of a single term in the analyzed privacy policies. Appendix N describes the regular expressions.

We carried out the terms' coding activity utilizing the Atlas.ti software to run queries using regular expressions. This process is called "auto-coding" in the Atlas.ti software and regular expressions are called "smart codes". We will use these terminologies interchangeably throughout this section.

4.5.6 FP/FN coding activity

Also, in this case, two researchers who worked independently of each other carried out the "auto-coding" activity while a third researcher arbitrated conflicts along the procedure recommended by Guest et al. [69]. The coding with regular expressions was structured through the following steps:

1. to use regular expressions for "auto-coding" on the sentence level all policies (both in the source and replication language) belonging to the two policy sets (see Subsection 4.5.1 of Section 4.5);

2. to check by hand the differences between the codes added manually in the pilot coding phase (see Section 4.5) and the ones added in this "auto-coding" step;

3. to manually add two additional codes (namely: *false positive* and *false negative*) according to the following rules;

   (a) if the auto-code assigned by the regular expression marks the exact sentence manually tagged in the pilot coding phase, then this auto-code is correct, and nothing more needs to be added;

   (b) if the auto-code assigned by the regular expression marks a sentence that was not tagged manually in the previous pilot coding phase, then this auto-code generates a false positive, and we add a *false positive* code to keep track of this occurrence;

(c) if the auto-code assigned by the regular expression does not mark a manually marked sentence in the previous pilot coding phase, then this auto-code generates a false negative, and we add a *false negative* code to keep track of this occurrence.

4.5.7 Refinements for FP/FN identification

Because we chose coding at the sentence level to mark the full semantic value that the technical term takes on in the context, we needed to do an additional refinement to isolate the terms that were false positives or false negatives in each sentence. Therefore, at a later stage, we used the Atlas.ti query tool to generate additional smart codes related to false positives and negatives. Appendix O describes the refinement process for identification of false positive and false negative items.

4.5.8 Evaluating the mapping of technical terms

We used a syntactic criterion based on the presence of widows or orphans, to check whether a syntactic distinction exists between technical terms. Appendix O shows an example related to the mapping of technical terms PI and PII. The idea is that technical terms (eg. PI and PII) are distinguishable if there is no correspondence in their position in the text of the two linguistic policies versions belonging to source and replication corpora. Otherwise, the technical terms PI and PII are indistinguishable. Hence, they are the same thing and can be codified in the replication language with the same technical term. We executed all the queries on the ground truth corpus described in Section 4.5), first by searching for the presence of widows and then orphans. Appendix R provides further information on how to search for orphans and widows.

4.5.9 Precision and recall

To assess the validity of an automated approach, we analyzed the precision and recall values achievable by using the "auto-coding" technique (see Section 4.5.6) against those derived from manual coding used to construct the ground truth as described in Section 4.5.

We used the formula (4.1) to calculate the precision value and formula (4.2) to calculate the recall one.

$$P = \frac{(\#Autocoded - \#FP)}{\#Autocoded} \tag{4.1}$$

Table 4.2: Precision and recall values in the "auto-coding" of source corpus policies

Refined precision and recall values obtained by the source corpus' restricted set of privacy policies of Ciclosi et al. [34]. Values refer to the following companies: *Bank of America (BoA)*, *Amazon*, and *Unicredit*. The abbreviations *pp* and *cp* are privacy policy and cookies policy, respectively.

	BoA US (pp)	**BoA US (cp)**	**Unicredit US (cp)**	**Unicredit US (pp)**	**Amazon NL-EN (pp)**	**Amazon US (pp)**	**Total**
Auto-coded	125	67	85	45	167	133	**622**
False positives	5	0	2	5	2	4	**18**
False negatives	14	15	41	52	66	56	**244**
Precision	96.0%	100.0%	97.6%	88.9%	98.8%	97.0%	**97.1%**
Recall	89.6%	81.7%	66.9%	43.5%	71.4%	69.7%	**71.2%**

$$R = \frac{(\#Autocoded - \#FP)}{(\#Autocoded - \#FP + \#FN)} \tag{4.2}$$

4.6 Results Cross Policy Comparison

4.6.1 Result RQ2.2: Where precision and recall were calculated by comparing values from auto-coding technique and manual coding (4.5)

After completing the "auto-coding" and analyzing the assigned codes, many false positives emerged.

A paradigmatic example is the mapping of the technical terms onto the source language *personal information (PI)* and *personally identifiable information (PII)* that were distinct in Tang et al. but collided in the replication language.

In summary, for these purposes, the technical terms *personally identifiable information* (PII) and *personal information* (PI) are equivalent.

Another example is that we have detected a ubiquitous type of false positive occurrence on replication language policies when the words *dati personali* are used to denote the name of the Italian supervisory authority (the *Garante per la protezione dei dati personali*) instead of the technical term. Many false negatives result from using the abbreviation *dati* in the replication language privacy policies to refer to *dati personali*. In the opposite direction, if we try to automatically look for the term *dati personali* by searching the term *dati*, we would return many false positives.

Table 4.2 shows the precision and recall values obtained from the automatic coding with regular expressions. The analysis was carried out by looking for technical terms by Tang et al. on the restricted set of privacy policies built from the source corpus. These values are refined, not including the four technical terms: *cookies analytics*, *cookies profiling*, *personally identifiable information*, and *sensitive personal information*. Ciclosi

Table 4.3: Precision and recall values in the "auto-coding" of replication corpus policies

Refined precision and recall values obtained by replication corpus' restricted set of privacy policies of Ciclosi et al. [34]. Values refer to the following companies: *Amazon* and *Unicredit*. The abbreviations *pp* and *cp* are privacy policy and cookies policy, respectively.

	Amazon IT (pp)	**Unicredit IT (cp)**	**Unicredit IT (pp)**	**Total**
Auto-coded	204	102	49	**355**
False positives	14	11	7	**32**
False negatives	56	61	52	**169**
Precision	93.1%	89.2%	85.7%	**91.0%**
Recall	77.2%	59.9%	44.7%	**65.7%**

et al. introduced the first two terms to intercept specific peculiarities of GDPR. The last two were in Tang et al. However, *personally identifiable information* is equivalent to another term (*personal information*). Whereas, *sensitive personal information* always appears with the term (*personal information*) because it represents a subset of the former.

Analogously, Table 4.3 summarizes the refined precision and recall values we found in the in "auto-coding" technical terms of Tang et al. [159] related to the replication corpus' restricted set of privacy policies of Ciclosi et al. [34]. In that case, we excluded the term *personally identifiable information* because it is equivalent to the technical term *personal information*, as shown in Section 4.5.8.

Key outcome: *The automated analysis approach for technical terms in cross-language privacy policy corpora may only produces a partial picture. While precision can be made good, recall is less so.*

4.6.2 Result RQ2.3: Where we evaluated the occurrences of the 58 technical terms by Tang et al. obtained by auto-coding (4.5)

We conducted an analysis to determine if the frequency of technical terms changes as the privacy policy version changes from the source to the replication corpus. To do this, we compared the frequency of technical terms used by Tang et al. in the privacy policies of Ciclosi et al. in both the replication and source corpora. We considered the top 10 and 26 lists of technical terms as an outcome of the "auto-coding" activity conducted on the entire two source and replication corpora. We carried out the checks related to the Italian version of the privacy policies on the replication corpus. Whereas, those involving the English version of the privacy policies used the source corpus. These

Table 4.4: The ten most frequent terms in the source corpus

The Tang's technical terms in the top 10 list of the source corpus of Ciclosi et al. [34]. These terms correspond to the terms belonging to the top 10 of the replication corpus except *Targeted ads*, *Track/Tracking*, and *Keys*. In parallel, the technical terms *Correct*, *Deactivate*, and *Personally identifiable information (PII)* who belong to the top list of replication corpus do not belong to this top 10 list of the source corpus.

Rank	Source Language	#Occurr.	Cum. Frequency
1	**Personal information**	1219	28,09%
2	**Cookies**	686	15,81%
3	**Privacy policy**	437	10,07%
4	**Account information**	280	6,45%
5	**Third parties**	278	6,41%
6	**IP address**	131	3,02%
7	**Targeted ads**	130	3,00%
8	**Encryption**	104	2,40%
9	**Track/tracking**	95	2,19%
10	**Keys**	71	1,64%

checks aim to verify that the list of technical terms includes terms frequently used in English and Italian privacy policies. Table 4.4 shows the top 10 lists of English technical terms that appear most frequently in the source corpus. Table 4.5 shows the same lists but refers to the Italian version of these technical terms in privacy policies belonging to the replication corpus. Table P.1 of Appendix P represent extended versions of these lists related to the first 26 terms.

Table 4.6 compares the position of technical terms of Tang et al. in the top 10 rankings of the most frequent technical terms of English and Italian privacy policies, belonging, respectively, to the source and replication corpora. Appendix S shows the comparison details.

In addition, we conducted a comparative analysis of the frequency of technical terms in the three corpora (original, source, and replication) to find out the share (in percentage) that the most frequent terms have in the total occurrences of the terms detected in each corpus. An exciting result of this analysis was that the technical terms appearing most frequently in each corpus represent the most occurrences of terms detected overall. Appendix S shows the occurrences details.

These outcomes suggest the importance of the frequency of occurrences of the most frequent technical terms being similar in each corpus, as the remaining terms are less relevant.

We checked the difference between the frequency of use of the 26 technical terms used in the main survey of Tang et al. in replication and source corpora of privacy policies of Ciclosi et al. In the study by Tang et al., these 26 technical terms are significant for investigating people's understanding of technical terms and are used in a

Table 4.5: The ten most frequent terms in the replication corpus

The Tang's technical terms in the top 10 list of the replication corpus of Ciclosi et al. [34]. These terms correspond to the terms belonging to the top 10 of the source corpus except *Correct*, *Deactivate*, and *Personally identifiable information (PII)*. In parallel, the technical terms *Targeted ads*, *Track/Tracking*, and *Keys* who belong to the top list of source corpus do not belong to this top 10 list of the replication corpus.

Rank	Source Name	Replication	#Occurr.	Cum. Freq.
1	**Personal information**	Dati personali	1524	26,09%
2	**of which PII**	Dati personali	1013	17,34%
3	**Cookies**	Cookies	650	11,13%
4	**Third parties**	Terzi (or terze parti)	621	10,63%
5	**Privacy policy**	Informativa sul trattamento dei dati personali	473	8,10%
6	**Correct (amend & revise)**	Rettificare	172	2,94%
7	**IP address**	Indirizzo IP	143	2,45%
8	**Deactivate**	Disattivare	134	2,29%
9	**Encryption**	Cifratura (or crittografia)	87	1,49%
10	**Account information**	Informazioni sull'account	60	1,03%

survey's questions. The 26 terms' original list composition is as follows: 22 were derived from a previous pilot study's results[2], while the other four were in the survey questions.

In analyzing meaningfulness, we considered the technical terms coded as an outcome of the "auto-coding" activity.

We found that more than half of the technical terms considered in Tang's survey are infrequent (and sometimes even nonexistent) in the replication corpus. Appendix T describes the count comparison of terms appearing between corpora.

Another interesting result is that the numerosity of occurrences with which these 26 terms appear in the source and replication corpora often differs widely. For example, the technical term *keys* appears 71 times (10 positions per frequency) in the source corpus but only 13 times (38 positions per frequency) in the replication corpus. Table Q.1 of Appendix Q shows a summary of the number of occurrences and position with which each of Tang's 26 survey terms appears in the three corpora (source, replication, and original).

Key outcome: *The technical terms used in privacy policies may have different meaningfulness, moving from one language to another.*

[2]The list of 22 terms is composed as follows: (i) two high-frequency terms that most users correctly defined; (ii) ten high-frequency terms that users commonly misdefined; (iii) ten additional terms for which users exhibited significant misunderstandings.

Table 4.6: Frequency position of the Tang's technical terms in source and replication corpora of Ciclosi et al. [34]

The columns "*EN Occ.*" and "*EN Pos.*" respectively refer to the number of occurrences of a term and its position among the most frequent terms in the privacy policies of the source corpus. The columns "*IT Occ.*" and "*IT Pos.*" have the same meaning but refer to the privacy policies that belong to the replication one.

Technical Term	**EN Pos.**	**EN Occ.**	**IT Pos.**	**IT Occ.**
Personal information	1	1219	1	1524
Cookies	2	686	3	650
Privacy policy	3	437	5	473
Account information	4	280	10	60
Third parties	5	278	4	621
IP address	6	131	7	143
Targeted ads	7	130	14	47
Encryption	8	104	9	87
Track/tracking	9	95	28	23
Keys	10	71	38	13
PII	13	64	2	1013
Correct	15	44	6	172
Deactivate	26	18	8	134

Table Q.2 of Appendix Q shows a possible list of these terms, while Table Q.3 lists some technical terms whose use can be misleading in an Italian study.

Key outcome: *While most technical terms remain in the top lists, some terms have very different frequencies of occurrence, moving from one language to another.*

4.7 Methodology Italian Replication (Result RQ2.4)

We started with the quantitative survey questions described in [159, App. A.2] and built a new Italian version reflecting the Italian-speaking user's peculiarity. To build a replication of Tang's [159] survey in Italian (see Subsection 4.7.1), we aim to choose terms usable in this example replication language. Most terms used in Tang's study are significant for an Italian-speaking user who reads an Italian version of a privacy policy; hence we can use them in an Italian survey. For others terms there might be a big gap in occurrences. For example, for *account information*, *aggregate*, *anonymize*, *de-identified data*, *keys*, *permanent cookies*, *sensitive personal information*, *session cookies*, and *web beacons*, there is a difference between the occurrences found in English and Italian privacy policies.

To limit the length of the survey, the time required for its completion, and to min-

imize the fatigue experienced by the participant [172, 40], we re-engineered it by constructing three separate shorter surveys. According to this choice, we conducted three studies in which participants answered 29, 27, and 26 questions, respectively, rather than a single one where they had to respond to more than 80 questions simultaneously. Appendices U, V, and W contain the full text of the questions from each survey (in Italian and English).

4.7.1 Structure survey

The surveys aimed to determine the staff's understanding, misunderstanding, and misconception of technical terms in privacy policies. The first survey was designed to assess the respondents' understanding of each technical term individually. It consisted of 29 questions, each focusing on a specific technical term. Participants were presented with information related to various technical terms in privacy policies and were asked to respond to these questions. The second survey, comprising 27 questions, aimed to gauge the respondents' comfort level with technical aspects. Specifically, it sought to compare the respondents' comfort with data use practices described in technical terms versus the same practices described in non-technical, descriptive language. The third survey consists of 26 questions about how each respondent is comfortable with data protection. Similar to the second survey, the questions compare how comfortable users feel with such data protection practices.

The original survey proposed in [159] consisted of two parts, the first containing questions on the understanding of technical terms and the second on the comfort users have with these technical terms. The first survey we proposed covers the first part of the original one, while the others we proposed cover the second part.

Analyzing Tang et al.'s original survey [159], we noticed that some technical terms were used only in one of the two parts. Therefore, we modified these surveys further to enrich them so that for each technical term, there was at least a question related to its understanding and another related to users' comfort with it. The technical terms used only in the first part of Tang et al. were: *keys* and *privacy policy*. Those used only in the second part of Tang et al. were: *account information*, *device attributes*, *encryption*, *public information*, *sensitive personal information*, and *session cookies*. In the Italian version of the survey, we have balanced the presence of technical terms by adding the missing questions for the previous eight terms. As a result, each technical term has a question on its understanding and another on users' comfort with it.

Finally, we substituted six technical terms not very significant for Italian-speaking

users (*track/tracking*, *private browsing*, *personally identifiable information*, *fingerprinting*, *do not track*, and *cryptographically hashed*) with six other more significant according to these findings and belonging to the Tang'list (*usage data*, *third parties*, *affiliates*, *location-related information*, *targeted ads*, and *unique identifiers*). We rephrased questions containing the replaced technical terms using the same privacy policies belonging to the replication corpus of Ciclosi et al. [34].

4.7.2 Participant selection

Three educational institutions in different regions of Italy participated in the research and were chosen using purposive sampling [109]. These institutions encompass one kindergarten, ten elementary schools, two middle schools, and two high schools serving 1918 students. They employ 355 teachers and over 57 technical and administrative staff members (excluding janitors of Institutions 2 and 3). The researchers contacted the principals of such institutions to present the research goals, and after obtaining the institute's formal endorsement of the project, the researchers began the recruitment of participants. We recruited participants from school staff, namely teaching, administrative, technical, or auxiliary personnel, and service directors. All school staff received an email containing a QR code and an anonymous link to access the survey. Before the survey started, the participants visualized the research information sheet and the privacy policy. If the participant gave consent, the questions were administered. We received no requests from individuals to withdraw their consent after the study began.

4.7.3 Ethics

The research was submitted to and approved by the University of Trento's Ethics Institutional Review Board (IRB) as a recommended practice in security research [98]. Procedures for recruiting participants and for data protection and storage were defined. Participants were adults working at the selected educational institutions. They were asked to give written informed consent describing the research goals and stating that participation is voluntary. Withdrawing from the research was possible at any time. Informed consent was given digitally by checking the appropriate check box on the survey.

4.7.4 Data analysis

We analyzed the responses to the surveys in an aggregate form.

Table 4.7: The Institutions' staff participation in the three surveys

Institution ID	**1st survey**	**2nd survey**	**3rd survey**
Institution 1	88	66	72
Institution 2	14	10	5
Institution 3	10	5	5

4.8 Results Replication

4.8.1 Data collected

We collected 128, 92, and 111 responses to the first, second, and third surveys. Some of them were incomplete, so the final number of responses is 112, 81, and 83. Because participation in the survey was optional and anonymous, not all staff participated. Table 4.7 shows the participation in the three surveys, divided by Institution.

4.8.2 Regression with accuracy & comfortableness

The high turnout at Institution 1 is due to its principal organizing a training course in which survey administration was the first step. However, since the survey was conducted before the training, the participants' responses could not have been influenced by the training outcomes.

4.9 Discussion: Interpretation and recommendations (Result RQ2.5)

What is critical for an English speaker to understand a policy may be negligible for an Italian reading a policy of a corresponding or even the *same* company. It has significant implications for the replication and generalizability of usability studies and even for defining specific users' rights.

A possible explanation for this result is that the choice of the policies in a privacy policies corpora is not neutral. Hence, the distribution of the occurrences of the technical terms used in the source corpus is only partly overlapping with the distribution of the occurrences of the terms in the policies of the replication corpus. This is somewhat to be expected as a company might not even be present in the country of the replication language. Still, a partial overlap is present, which would allow a potential partial replication of Tang's study [159] in Italian.

These differences in occurrences between the English and Italian policies may derive from the difficulty of using an automated approach to intercept the possible translations of these technical terms in Italian, keeping the number of false positives low, or might be due to a different relevance of the activities for the intended audience of the privacy policy.

As takeaways, we provide the following recommendations to researchers who want to replicate a study on privacy policies in a language other than English.

- equip themselves with comparable cross-language corpora;
- construct their ground truth on the considered corpora through manual coding of a subset of policies and then analyze the results to determine the regular expressions to be used in a subsequent "auto-coding" phase on the entire corpora;
- check the validity of the results and ground truth by checking the precision and recall obtainable from analyzing the differences in the obtained results between manual coding and "auto-coding";
- ensure that the frequency and significance of key terms for the reference domain are similar between the two languages;
- make the necessary adjustments to the original study before replicating it according to the previous outcomes.

While Ciclosi et al. [34] provided a methodology for building interlinguistic corpora of privacy policies, we provide a method codified as manual algorithms to code and analyze these corpora and to determine the key domain terms' occurrence frequency or meaningfulness. We also provide some suggestions for replacing less interesting terms for users who speak the language of the replication study.

4.10 Conclusions and limitations

This study examined the cross-lingual mapping of technical terms in privacy policies, specifically from English to Italian.

By performing a mix of manual and automated analysis (carried out using a regular expressions search with the software Atlas.ti), this chapter reports the result of the mapping of the technical terms of Tang. et al. [159] in the cross-language corpora of Ciclosi et al. [34]. We found that the terms' frequency position in the replication corpus of privacy policies in Italian is *mostly* acceptable.

Yet, the significant difference in the occurrences of some technical terms in the privacy policies belonging to the source (in English) and replication (in Italian) corpora requires additional investigation before concluding that the analyzed list of terms is representative.

An example of such difference in sensibilities is the term *fingerprinting*. In the original study by Tang et al. [159], this term was among the selected ones. This analysis shows that the occurences of the term were negligible in the multi-lingual corpus from Ciclosi et al. [34]. This finding does not mean that Italian users do not worry about the problem. It simply means that using that *particular term* in a comprehension study in Italian would be wrong, and it should be replied by another term: compared to U.S. users, Italians do not worry about being 'fingerprinted'; they worry about being 'identified', and their privacy policies reflect this nuance.

In general, these findings suggest that for a meaningful replication of research on understanding privacy policies in other languages (e.g. the EU and its citizens), *the set of considered technical terms must be definitely adapted*.

Further, some technical terms must be better detailed to elicit a more accurate user reaction. For instance, the term *location information* refers to the location of something or someone, but this location data often includes *different* information. The data may refer to a generic location (e.g., a city- or country-level location) or a precise location (e.g., the location of a personal device obtained using a GPS service). Users might respond differently to a different query depending on the precision of the location. Hence, the privacy policy's full context must be considered.

The generalization of this study to other languages must account for some of its limitations. The first obvious limitation is about the choice of Italian as the language of the replication corpus. The findings related to Italian policies might transfer more easily to Spanish or French than to Dutch, German, or non-phonetic languages.

The second limitation concerns the small number of privacy policies investigated. Although adopting a larger number of policies can help to understand how the proposed approach generalizes, the presence of many false positives and negatives that we have identified even with such a small sample makes such task likely daunting.

A third limitation is the use of Alexa Top 10 to identify comparable policies. While this is the standard approach in the literature, this approach is not immediately applicable when considering cross language policies and must be adapted country to country as the top 10 worldwide might be far from the top (or even not exist at all) in the country target of investigation. Addressing this issue requires some manual choices on the representativeness of the selected organizations that might bias the analysis. Using the

multi-language collections from the proceeding of the European Parliament (*Europarl Corpus* [84]) might be an alternative solution of matched cross-lanaguage texts. Unfortunately, it is easy to argue that such collection is hardly representative of the online privacy policies that users might encounter. The corpus of the decision of European GDPR authorities provide another interesting source. Yet, its original source fails the requirement of having matched cross-language text that are not the result of manual comparisons.

The open issues above provide interesting avenues for future work on the replication of studies on understanding privacy policies that span different languages and cultures.

Dataset Availability

The dataset containing the technical terms in privacy policies in Italian and English is archived on Zenodo at `https://doi.org/10.5281/zenodo.8297075` [36], while Appendix I displays a description of the dataset characteristics and composition.

Chapter 5

GDPR in the Small: a field study of privacy and security challenges in schools

Abstract: The GDPR was enacted to reign in the mighty corporations of the internet. Then, it was unleashed on all organizations, large and small alike. This chapter reports the results of a multi-site field study on schools, and the challenges they face to implement the GDPR while running activities full of sensitive issues without an army of legal and compliance officers. The study did not find evidence of the privacy paradox (spotless on paper but careless on the field). In contrast, school staff mostly crumble when by-the-book procedures cannot be implemented with the resources that they actually have. The chapter discusses how privacy violations happen on the field, from critical with potential impact on pupils security and safety, to 'formal' violations for which life is too short to bother and how a risk-based approach would be more effective to address them.

Small organizations often struggle to comply with data protection rules. The ideal situation is, of course, bulding a corporate culture of data protection [94], but this is hard to do when tight budgets make it impossible to pay privacy trainings for regular staff or to allocate dedicated staff to privacy compliance[151, 43]. Even the Data Protection Officer (DPO) is often just an individual who end up playing at once advisor and guardian, executor and auditor [33]. The limited knowledge of the data processing, as actually implemented on the field, might also mislead staff into thinking that privacy management is sometimes unnecessary [72].

Among the different types of small organizations, schools are an ideal case study [161, 154, 27, 150, 80, 1]. To make ends meet, they focus on formal compliance: information sheet on processing, manuals, or register of processing activities. This phenomenon that also happens for small operators in charge of securing critical infrastructures [108]. On paper, they tick all the boxes. In practice, they might fall into violations as we also show in this chapter, GDPRxiv [155], an open-source archive collecting all official GDPR rulings from 2018 to 2023, reports that 123 rulings in EU member states were doled to educational institutions. Compare them to 34 rulings to the four Internet giants, 62 to telecom operators, and 138 to banks. A huge fine is often immaterial for banks, telcos or the big four. For a school, whose budget is mostly incompressible (staff salaries), a small fine might wipe out most of the discretionary budget of the institution.

Security technologies could help. Yet, when we deploy them we must think to field use. Consider a simple and apparently obvious question: *Does any device in the school need a password for individual staff members tailoring access to individual class school registers?* It is RBAC 101 and ensures GDPR security measures (art 32). Consider a Principal's challenge: the solution has to work when a teacher calls sick at 7:30am and at 8:00am a replacement needs to step in a class of 10 years old. At 7:55am you have finally found the person. Now, depending on the technological solution (tablet with teachers shared password in locked drawer or a top-notch RBAC system), you need an IT administrator to change access control in 5 minutes to a sensitive list of minors without making mistakes. Pity the only IT admin is fixing PCs for the 8am Physics class. Second challenge: a divorced parent is not allowed by a judge to see the kid. Make sure the replacement teacher (as all teachers of the class) knows they can't hand the kid to the father and nobody else knows it.

It is pointless to preach small organizations to a standard that they cannot achieve, as they lack the resources to do so. This is a recipe for privacy violations to occur.

The candidate's goal is to support schools to make sure we identify and mitigate structural and high risk violations and balance the resources available in the true spirit of the law[1]. The contribution of this chapter are manyfold:

1. We carried out a field observational study to explore the role of human factors in the socio-technical Italian school system by analyzing the practice on the field of educational institutions including different levels of schooling (i.e., kindergarten, primary, middle, and high school);

[1]Recital 4 GDPR: [...] The right to the protection of personal data is not an absolute right; it must be considered in relation to its function in society and be balanced against other fundamental rights, in accordance with the principle of proportionality.

2. Based on the observation, we propose a coding taxonomy of factual observations and *sequences of actions leading to violations*;

3. We propose a classification of violations in two types, *structural* violations related to a poorly designed organizational procedures and *occasional* violations, due to an incorrect and episodic behavior of staff

4. Finally, building on the analysis of the requirements and fines given to schools we were able to classify violations ranging from *high risk* violations, that is violations likely result in serious damages to the rights and freedom of people and *low* risk violations.

In summary, the candidate's findings are that for small organizations there is no privacy paradox (an inconsistency between people's privacy attitude and their behavior), but just a conflict between competing needs of staff who is 'called elsewhere' because there is no one else to answer the call and in doing so they may unfortunately generate data leaks. Thus, the solution to these privacy problems is the design of organizational protocols robust to interruptions and amenable to mitigations which balance costs with risk.

An important disclaimer applies here. The opinion whether a privacy violation belongs to the categories 'high risk' or 'life is too short' is entirely due to the authors and it should be in no way attributed to the people working at the institutes under investigation. As we mentioned in the abstract, unfortunately the GDPR applies identically to the large and to the small, to the tragedy and to the mockery.

5.0.1 Artifact Availability Statement

We make available through Zenodo (`https:\doi.org/10.5281/zenodo.15133671`) the following material: the templates of the information and consent forms, survey questionnaires and survey results, coding guidelines with selected examples, codes' occurrences and correlations. The field observations will be not publicly released to preserve the privacy of people participating in the study. Full text of the field annotations cannot be released as it could allow teachers and staff of the surveyed schools to re-identify most individuals responsible for the privacy violations. Researchers willing to analyze the raw material can contact the authors to view it on our premises.

The Italian School System in a nutshell. A public *Istituto Comprensivo* is an institution composed by N primary schools (usually at different sites) whose pupils continue in M middle schools, then in K high level school types (e.g. science and classics vs accounting), all under the same principal. Each level comprises few schools of each type. A principal is in charge of the entire organization and some deputies supervise the different sites. In Italy, there is not the USA level of inequality, so the public-private difference is not relevant.

Figure 5.1: The Italian Istituto Comprensivo

5.1 The Italian School System and its GDPR Requirements

5.1.1 The Italian School System

Since school systems are different from country to country, we provide here a brief introduction to the Italian school system. By the term 'school' one typically denotes a single type of school, such as an elementary school, with a principal and the teachers. For example an elementary school typically includes pupils of the age bracket 5-10.

In Italy, the schools are grouped into larger *Educational Institutions (Istituto Comprensivo)* (EIs for short) grouping early childhood, primary, junior secondary and secondary schools, or some secondary educational institutions of different type (for example scientific vs. vocational education). Figure5.1 summarizes this structure. As should be clear from the above explanation, the Italian Istituto Comprensivo is *not* a "*comprehensive school*", term used in some countries to designate post-primary schools that accommodate pupils between the ages of 11 and 18.

Each EI is a part of the national school system while maintaining its own administrative, educational, and organizational autonomy. EIs can be spread over several sites. The main site might include the whole educational pipeline from primary up to high-school, while a satellite site in a remote part of the county might only offer a primary school.

A *Principal* heads the EI and is responsible for both administrative and educational management. The school principal of each EI is selected from among tenured teaching staff in an open competition (so not necessarily among the teachers of the very schools of the EI). Moreover, each EI has a dedicated administrative office (secretary) for administrative purposes and to maintain relations with third parties. The EI principal decides all activities after consulting with the *School Council*, an internal body of the EI

including as its members the parents' representatives. If the EI is multi-site, a deputy is appointed for each site.

5.1.2 The GDPR Requirements

The handbook of the Italian Data Protection authority reports all privacy requirements peculiar to schools [63]. Schools can only process strictly necessary data, regardless of the nature of processing (electronic or paper). They are allowed to process personal data only for institutional purposes (e.g., education and training) defined by law. For other purposes, some types of data (e.g., those required from theater courses not included in the school curriculum) may be processed after obtaining the consent of the data owners. Special categories of personal data relating to pupils (such as ethnic or racial origins, religious beliefs, health status, and so on) may be processed only for specific purposes established by law. Schools must also make sure that data shared on online platforms is kept secure and protected. Documents with sensitive data should be stored and transported in closed folders for protection.

The *Data Controller* is the school's legal representative, i.e., the principal, who defines the activities to be undertaken and how to implement them. The school's administrative and teaching staff are authorized to process personal data in their respective areas of competence. The data controller must (i) prepare a processing register to have an up-to-date picture essential for the assessment and the analysis of the risk associated to such processing; and (ii) appoint a *Data Protection Officer* (DPO). In addition, a set of technical and organizational *minimal mandatory security measures* must be taken to prevent both circulation of data among unauthorized colleagues and its unwarranted disclosure to third parties (e.g., other students and families). Staff have a duty of confidentiality and professional secrecy regarding the data they process [63]

Table X.1 in the Section X.1 of appendix X lists the main issues of data processing the Italian legislation on the the protection of personal data in the school setting.

5.1.3 Data Protection Authority Measures

Private operators are typically fined because they took some person's data for some purpose and then started using for untold, own commercial purposes. In contrast, schools are typically fined because some staff did something in good faith that disclosed data, sometimes even of a single pupil whose parents complained.

The Italian Data Protection Authority (Garante della Privacy) uses three instruments to support the implementation of the GDPR.

- *Rules/manuals* apply to a category of organizations in terms of recommended practices and minimal security measures. There is a specific manual (Vademecum) on schools [63]).

- *Enforcements* typically includes fines and obligations of the affected parties to stop (or start) using some particular process or technology. They can be found in GDPRxiv.

- *Warnings* just notify the offending party that they have to fulfil a certain condition before is it does not want to incur stronger enforcement measures.

Since the full implementation of the GDPR, over the past seven years, 29 schools received either sanctions or warnings and 58 out of 2561 published issues in the database of the Italian Data Protection Authority refer specifically to educational institutions.

On GDPRXiv out of 388 enforcement actions concerning GDPR in Italy, 27 involved schools receiving sanctions for posting excessive personal information online (16 cases), in restricted areas of the electronic register (3 cases), or on bulletin boards (1 case).

Mere "warnings" are not in GDPRxiv, but they exist (14 out of 58 measures): the authority recognized that the violations were determined by small scales, lack of resources, or mere material error and only issued a warning. Analyzing the case history of rulings provides only a partial view of real-world occurrences. The scarcity of ruling case histories can be attributed to various factors, such as the absence of filed complaints with the Data Protection, arbitration of specific cases (e.g., when data subjects and data controllers reach their own agreements), the lack of publication of some measure on the Authority's website (it is a secondary sanction), and the Authority's inspection activities focused on other organizations.

Field observations are therefore crucial for understanding the extent and prevalence of privacy violations.

5.1.4 Non-Goal

After talking with the principals, a clear non-goal emerged: the management of digital data by third-parties. This was surprising as in 2019 there was a significant controversy in Germany on the use of Microsoft O365 online services in school, following a report of the German Data Protection Conference (DSK) – which consists of the German Federal Data Protection Authority and 16 state regulators. Office365 was then banned across

all schools[2].

Obviously, a part of the data processing happens when schools use third-party applications to administer, for example, emails, assignments, parent-teacher communication, and privacy violation can occur in such system. For example, some EIs made agreements with Google to offer emails to staff and students. Also financial data is often processed by software providers responding to procurement tenders at county level across multiple EIs.

While the responsibility for privacy violations always stay with the data controller according to the GDPR (in this case the EI's principal), if the contract with the third party is well drafted (and most are standard), then the EI has several legal protections. Loosely speaking, if the data of pupils is lost by the software in charge of processing grades, before the buck arrives at the door of the EI's principal, it would have made several other stops. The first target of the privacy authority would be the ministry or the county council (which procured the contract and therefore the EI could not refuse), then the third party itself for having failed the duty of care.

The major concern of the EI is the management of the school's own IT and the mishap of their own making. They are overwhelmingly fined for something their staff did, and not for something their suppliers did. For example, Zoom platform recordings were used for disciplinary purposes, video surveillance systems were used during school activities, and pupil health data was even communicated to other pupil families.

Appendix Y lists the main references to Italian legislation regarding the protection of personal data in the school setting.

5.2 Related Work

We reviewed literature on privacy management in small users' communities and small organizations taking into account the last five years of papers published in the proceedings of the following international conferences: *ACM CCS*, *ACM CHI*, *IEEE SSP*, *PoPETs*, *USENIX Security Symposium*, and *USENIX SOUPS*.

Some studies investigate how specific communities perceive and manage privacy and security issues. For example, some of them analyze how elderly people communities collaborate to cope with these issues (e.g. [89]), others investigate the privacy perceptions of bystanders (e.g. [175, 37]) or users [179] of smart homes or smart speakes [52, 93]. Finally, some studies report how software developers shape their responses to the security

[2] `https://datenschutzkonferenz-online.de/media/dskb/2022_24_11_festlegung_MS365_zusammenfassung.pdf`

and privacy requirements they receive in the development process (e.g. [103]).

Qualitative studies on organizational issues are uncommon. Kokulu et al. [87] analyze interviews with security workers on organizational issues in Security Operation Centers (SOCs). As another example, Hielscher et al. [76] studied how the Chief Information Security Officers (CISOs) perceive Human-Centered Security (HCS). A further example is related to investigating the role of regular security meetings of the software development team for achieving continuous software security [167].

More recently, several ethnographic studies exploring privacy issues were carried out. Chattopadhyay et al. [28] examine at which extent and how cognitive biases affect software development, and the practices and tools that developers adopt to cope with such biases. In their research, Palombo et al. [117] focus on secure software development processes in a software company, whereas Dalela et al. [44] conducted an ethnographic study of security and privacy practices in Danish companies, highlighting the outcomes differences between SMEs and large companies.

To our knowledge, however, there is only a few works targeting privacy in educational settings. Most authors address, indeed, the privacy implications of specific educational technologies (EdTech) at schools, such as WiFi service delivery (e.g. [81, 13]), VPN networks [174], proctoring systems [20], remote educational technologies [38], online conferencing platforms [156], and clouds [66]. Chanenson et al. [27] survey school staff to investigate which risks EdTech can produce for students' privacy and security. On the same note, some technical contributions were conceived to alleviate students' privacy risks in EdTech. Hasan and Fritz [75] propose an approach exploiting feature selection and adversarial censoring techniques to build privacy-preserving versions of learning analytics datasets. Hasan [74] implements a pipeline to automatically detect posts related to EdTech's privacy and security issues.

Other studies focus on how teaching staff approach to privacy and security issues. Mayer et al. [110] qualitatively study the use of password managers. Tu et al. [164] performed a series of telephone phishing experiments on university staff and faculty to explore why scam works and how to defend against it. Kumar et al. [91] investigate how educators take into account privacy and security analyzing their curricular and classroom management goals. In parallel, other works target the final stakeholders of education institutions, that is students and their parents. Zhong et al. [180] study the parents' awareness of privacy and security risks at schools, whereas Balash et al. [12] investigate students' concerns about sharing their data with third parties (proctoring platforms). McDonald et al. [111] focus on how librarians interpret or define their own privacy rules to protect users.

This multi-site study aims at filling the current lack of research on the assessment of the implementation of organizational procedures on the field.

5.3 Study Design and Data Collection

The methodology of this field study takes inspiration by the previous work described in [176, 146] and [148], and grounds on the simultaneous usage of field observations and self-assessments of privacy attitude of educational staffs in their daily activities. This dual approach was chosen to tackle the *privacy paradox* stating that people's attitude and behavior concerning privacy can be inconsistent [141]. Field observations were analyzed using a thematic analysis [70] and a grounded theory methodology [42] following recommendations suggested in the Appendix A of [64] to generate concepts potentially transferable to other scenarios.

The key steps of workflow are inspired by the work of Burrows et al. [21]. We briefly described here the steps. More details are reported in the next sections of the chapter.

The research plan and the related documents were reviewed and approved by the Ethics Institutional Review Board (IRB) of the University of Trento.

5.3.1 Ethical Approval

The research goals and the procedures for recruiting participants and data protection and storage were submitted to and approved by the Ethics Institutional Review Board (IRB) of the University of Trento

The informed consents signed by participants were retained by the main researcher and secured in a safe. Data collected through the preliminary survey were anonymous, whereas data collected through observations were pseudonymized to avoid potential identification. For research integrity purposes, the tables with the original paper observations that could enable re-identification were securely stored by the principal investigator of the study into a physical safe (hence our statement that if you want to see the raw data you have to travel to our site).

5.3.2 Setting and sample

The schools involved in the research were chosen using a purposive sampling [109], leveraging researchers' direct connections with principals. The target was three Italian EIs including kindergarten, primary, junior secondary and secondary schools in two

regions. The researchers contacted the EIs' principals to illustrate the research goals. After obtaining the institute's formal endorsement of the project, research began with the recruitment of participants.

We recruited participants among the EIs' staff, that is teaching / administrative / technical / auxiliary personnel as well as service directors, data protection officers and physicians. Students attending the schools and their parents were not included in the study. The researcher responsible for field observation met the candidate participants in a preliminary meeting organized at schools by the principals. During such a meeting the research goals were clearly explained. Sometimes the meeting followed a data protection course taught the school's staff by the researcher.

To protect the privacy of the participants, principals did not attend the meeting and did not have access to information about who did or did not join the research. At the conclusion of the meeting, participants who wished to contribute to the study were given the opportunity to provide written consent. They were notified that they could withdraw from the research at any time. A few weeks later, we began the field observations for people who provided consent. Section 5.4 shows the case study's details.

5.3.3 Preliminary Survey

Some days after the preliminary meeting, participants received an email from the school office containing a QR code and a link to access an anonymous online survey aimed at determining their understanding of the institution's practices on personal data protection. To fill up the survey, participants must read the research information sheet, the privacy policies, and express their consent by ticking the appropriate check box in the welcome page of the survey. The survey consisted of 22 multiple choice questions covering key data protection issues, such as: the role and function of data protection officer (DPO), the privacy notice, the data protection policy, the processing register, the data protection impact analysis (DPIA), and so on.

The questions were elaborated during several meetings with researchers expert on privacy that already carried out similar studies. All questions were written in Italian, the native language of participants, and formulated as much as possible with a neutral valence to avoid to elicit negative feelings. The survey also included some demographic questions to determine the size and type of schools the staff works in, their role, and their years of experience. The complete set of questions is available in Section X.2 of Appendix X.

Responses to the preliminary survey were analyzed in an aggregate form with the

aim of: (i) verifying whether the declared school staff's knowledge on the privacy organizational structure reflects the actual situation (e.g., if people of the staff state that the school has appointed a DPO, we verified whether really there is a DPO at the school); and (ii) making sure that at large, staff members were aware of privacy policies and the need of compliance even if they where not aware of where the lasted form was. The data on the survey is available on Zenodo (See Section5.0.1).

5.3.4 Field Observations

Procedure. A researcher with more than 10 years of experience in data protection carried out field observations and reported events at each EI in two distinct periods. As a design choice, half the time was spent in classrooms/laboratories, and another half in administrative offices. Behaviors were observed at different timings to cover, on a sample basis, the total work activity carried out by institutes' staffs throughout their daily and weekly duty hours. The researcher decided on the field the number of observations to make according to what happened in the different areas. More specifically, the researcher followed the personal data paths discovered in the field by choosing where to observe and what to observe in real-time. In their observation activities, the researcher moved independently between the areas. According to the grounding theory, there was no a priori decision before entering the field about what to look for (e.g., vulnerabilities and how to classify them). The researcher repeated the field observation procedure until data saturation was reached.

- In the first period, the observations covered data processing in administrative offices, such as the head office, the teaching secretary's office, and the administrative secretary's office.

- In the second period, observations concerned data processing carried out by teachers in the context of activities in classrooms and laboratories. If the school principal allows the observation of a small sample of teaching staff, the second observation period may be shorter, as fewer hours may be needed to reach data saturation.

Annotation schema. An annotation schema was defined following suggestions in Corbin and Strauss [42], Hoegl et al. [77], and Anderson et al. [8] that provide guidelines on assessing the quality of group work, and define a system of behavioral markers usable in the reference domain.

To describe how personal data travels and transforms all along the activities, the annotation schema reported the location where the observation took place (e.g. classrooms), the people present at that location (e.g. staff, collaborators, third parties), the ongoing activities and their pace, the relevant incidents, the interactions between people, and all information that the researcher in charge of the procedure considered relevant at that time. Annotations were performed as much as possible in real-time in a written form without recording or videotaping tools. A fine tuning of annotations was carried out right after each field observations to add contextual data useful for the research goals and further details [176]. Information acquired on the field study was kept confidential, and the observed dynamics did not impact staff members' performance evaluation. As reported on the experimental protocol, no data of student, their parents or members of the school's staff who did not provide their consent was collected.

Section X.3 of Appendix X shows the annotation schema we used to collect the observation notes.

5.4 Case Study

Three EIs located in different regions of Italy participated in the research. The sample study consisted of one kindergarten, ten primary schools, two junior secondary schools, and two secondary schools. They are divided into 17 sites, corresponding to a total of 355 teachers, more than 57 people of technical and administrative staff, 1918 studentes over 112 classes (excluding the number of kindergarten's pupils at EI1 in Table 5.1). Table 5.1 summarizes the figures of each EI (we did not include janitors of EI2 and EI3 in Table 5.1 as well as the three principals).

Table 5.1: Figures of the EIs involved in the study

(*) We did not include in the count the janitors of EI2 and EI3.

Number of	**EI1**	**EI2**	**EI3**	**Total**
Students	763	501	654	1918
Classes	41	30	41	112
Teachers	159	73	123	355
Other staff members	39	7(*)	11(*)	57

All the teaching and administrative staff's members participating in the study were requested to fill in the preliminary survey. We collected answers from 112 members. Only 98 surveys were complete and retained for the analysis.

Field observations lasted in total 81 hours (30h at EI1, 29h at EI2, and 22h at EI3). The annotation took 29 days (12 days at EI1, 10 days at EI2, and 7 days at EI3. Overall, 125 staff members were observed (68 at EI1, 38 at EI2, and 19 at EI3). Observations were carried out in administrative offices for 54.3% of the time and classrooms/laboratories for the remaining 45.7%. Table 5.2 shows the percentages of annotations made on the premises of each EIs. In two out of three cases, the percentage of observation coverage is 100%. In the third case, this is 50% overall but rises to 76.19% when considering the primary site of the EI, where most of the annotations were made. In addition, when analyzing the detail of the type of rooms observed in the primary site, the percentage of coverage is almost total for the main types (91.67% for classrooms, 100% for computer labs, and 100% for the other significant rooms). It is reduced to 20% for the other labs with computer equipment for teachers.

Table 5.2: Observations (%) made at each EI

Percentage values refer to the number of visited rooms and they are rounded to the nearest whole number. The *faculty rooms* wording includes the faculty room, server room, interview room, and lecture hall. The *offices* are the administrative ones.

EI	**Classrooms**	**Computer labs**	**Other labs**	**Faculty rooms**	**Offices**	**Total**
EI1	100%	100%	100%	100%	100%	100%
EI2	100%	100%	100%	100%	100%	100%
EI3 (all sites)	55%	80%	10%	100%	100%	80%
EI3 (main site)	92%	100%	20%	100%	100%	100%

The researcher chose the number of observations to make based on the dynamics in the field and the need to observe different circumstances (e.g. during class or recess). If a teacher did not give the consent to participate in the research, observations were made only at the end of lectures. Multiple observations were made in all administrative offices because multiple staff's members were in each room. In rare cases where a person did not consent to participate, the researcher did not observe them, focusing on the other staff members.

Thus, the number of observations' sessions made in classrooms / laboratories ranged from 1 to 3; that one of observations made in the faculty and server room varied from 2 to 7 and from 4 to 5, respectively. The number of observations in the administrative offices was higher because as the most significant flow of data was there, the researcher following the data path spent more time there.

Table X.8 in Section X.7 of Appendix X shows the codes detected day by day at the EIs: on the days in which observations were carried out in the administrative offices a higher number of codes was used. This disparity in the number of codes was to be

expected since administration is the beating heart of the data flow.

5.5 Data Pre-processing and Analysis

5.5.1 Pseudonymization

The following elements of annotations were pseudonymized: (i) the name of the EI; (ii) the date of observation; (iii) the place where the observation took place; and (iv) the first and last names of the EI's staff involved in the study. To further reduce the re-identifiability of the EI's staff members, gender references were removed. The correlation tables with the pseudonyms were stored in a paper format in a safe and they will be permanently destroyed at the end of the storage period specified in the experimental protocol. Analysis was carried out on pseudonymized individual or aggregate data to minimize the risks of linkage attacks on data privacy [140].

5.5.2 Unitizing and filtering

Unitizing is the process to segment data stream in meaningful data units, here identified by *applying a semantic criterion* of following the personal data paths observed in the field. A boundary of a unit is set when:

- the interactions between the EI staff's members processing data end;
- the link between the data and the EI staff members involved in data processing is broken;
- data is safely stored (or disclosed);
- the action(s) on data are concluded.

Thus, the units can have different temporal length. We manually inspected the segmented units to filter off those not relevant for this research.

5.5.3 Coding

Codebook. Starting from the annotations, a set of 21 unique codes was defined after several discussion among the researchers to summarize the paths of personal data observed in the field (see Section X.5 of Appendix X for the codebook template, and Section X.6 of the same Appendix for codes' details). The researchers discussed the

annotations using the Atlas.ti software and arbitrate their conflicts as recommended by Guest et al. [70]. In the rest of the section a *Quote Identifier* is denoted by QID.

The codes are centered on the notion of data processing, they describe the *types of data processed; people accessing and manipulating it; how people interact around it; different data processing; and the types of failure in data processing.* The rationale behind the definition of such codes is to concatenate them into sequences that end into a failure.

Table 5.4 provides an excerpt of the codebook with examples of when or not to use a specific code, while Tables X.5, X.3, X.4, X.6, and X.7 in Section X.6 of Appendix X provide its full version.

People and Data. We identified three *types of data*: physical data (PH), digital data (DI), and audio data (AU). Understanding what is physical or digital data is trivial. Audio data, however, is a transient data type, meaning that it is available to people only for a certain amount of time and in a certain location. Audio data can be the content heard by a person during a discussion or walk on a corridor; a WhatsApp voice message, instead, belongs to the family of digital data. The codebook defines four different *types of people accessing and manipulating data*: the teachers (T), the administrative staff (A), the head staff (H), and third parties (e.g. students, parents, other people alien to the educational institutions) to whom data is exposed in some way (TP).

Data Processing. Creating a new data is marked with the code C, the pairs of codes *C-DI C-PH* indicating the creation of a digital document (e.g. a file) and a physical document (e.g. a paper form), respectively. Consulting a physical or digital document is identified by the code Q. During processing data can undergo digital ($2D$) as well as physical ($2P$) transformations. Summarizing, we can have: i) physical-to-digital transformation (PH-$2D$-DI), ii) duplication of digital data (DI-$2D$), iii) digital-to-physical transformation (DI-$2P$-PH), and iv) duplication of physical data (PH-$2P$). Storing physical data is marked as SP. We did not use any specific code for the storing of digital data due to the difficulty to determine precisely, just through observations, when it is definitely stored. Transferring data between people is another typical processing. DTR is used when there is one or more senders and one or more receivers, and at the end of transfer the sender(s) cannot longer act on data; DTT is adopted when the senders(s) can continue to process data. Finally, M is employed to report when people stops to process data (e.g. going away). If nobody is present in the observation area but something relevant happens, the unit is coded from the perspective of actions done or not done by the person (e.g., not locking a computer or leaving a document on the

Table 5.3: Codebook: Types of failure in processing

Code	Long code	Description	When to use it	When not to use it
F	Failure	Something within the processing procedure does not work as expected and generates an error or failure (lack of compliance).	A printer does not print a document because of a paper jam. A person cannot access an online service because of an application error. A person leaves a room without locking the computer's operating system. A person exits a room, leaving open documents on the desk. A person cannot consult a document because of an error generated by the document management system. Two+ people discuss processing in a public place like a hallway. A person leaves a document unattended in the printer.	A person exits a room and leaves open documents on the desk because another person is processing them. An operator leaves a document on their desk in an access-controlled room where only another operator authorized to process the data in the document is present.
R	Resources failure	A procedure or process fails due to a lack of resources. Used also when two or more procedures have conflicting requirements that fail. **Jointly use:** This code must be used with the *F* code, as it specifies a particular failure circumstance.	In a room, an unauthorized person can listen to talks about data because of the physical layout of the premises and compliance with the organization's procedures. A single janitor has to be at the reception desk and simultaneously to perform activities elsewhere. An unauthorized person can hear the conversation in another room because of the unsoundproofed walls.	Because of non-compliance with the organization's procedures in a room, there are several people inside, not all authorized, who can listen to talks about data.

table). For instance, the unit *"There is an unlocked computer in the room"* will be interpreted and coded as *"The staff' member left data unattended by not locking the computer."*.

Interactions around data. Mid-way the coding process, a consistent pattern emerged. Failures tended to happen in presence of interactions (or failed interactions) among people and data. We identified thus four types of interactions: individual actions (*IA*); interaction (*I*); indirect interactions (*II*), and missing interactions (*MI*). The individual actions consist of actions performed on data by a single person; interactions are deliberate interactions among two or more people around data; indirect interactions occur when data processing is done not voluntarily in presence of third parties; missing interactions are interactions during which some of the people involved in data processing do not behave as they should (e.g. going away and so on).

Failures The last two codes of the codebook are about lack of compliance (*F*) or lack of resources (*F-R*), respectively. A compliance failure occurs when something within the processing procedure does not work as expected, so an error or failure happens. Such kind of failure is about actions or omissions of people involved in personal data management acting in a non-compliant way with respect to the law or the organization's rules. Conversely, a resource failure occurs if a procedure or process fails due to a lack of resources. In this case, the failure is under the organization's responsibility, because it is precisely by complying with the procedures that people commit failures. Table 5.3 illustrates the types of failures. An example of compliance failure is a teacher lecturing by connecting their personal laptop to the school network; or a staff's member leaving

the server room unmanned without locking the door. An example of resource failure is a janitor asked to perform simultaneously two contradictory tasks such as operating at the reception desk, and going in a classroom to drop off a document. Another example is not having rooms for private communications with parents, so that teachers share the available rooms with each other.

Sequences of Actions. Eventually, observations of sequences of actions are coded that ends into a failure. They make it then possible to understand what caused, or at least preceded a failure. Table X.10 in Section X.9 of Appendix X contains the text of the quotes.

Similar units can be coded in a different way according to the contect and the examples. Table X.2 in Section X.4 of Appendix X reports some annotations and the corresponding strings of codes about staff's members leaving a room without locking a computer. Although all the units share some codes (e.g., A (administrative staff), DI (digital data), and F (compliance failure)), some specific codes were used to specify the context in which observations about data were performed. Thus, for example, the TP (third parties) code was used if (i) third parties were present in the staffed room (Table X.2, QIDs: 129 and 982), or (ii) the room was unmanned (Table X.2, QIDs: 402 and 1557), with the possibility of third parties entering undisturbed, as the investigator did.

Table 5.4: A codebook's excerpt

ID	**Long code**	**Code Description**	**When to use it (examples)**	**When not to use it (examples)**
T	Teaching staff	a teaching staff member is involved in data processing.	A teaching staff member is on field.	An administrative or managerial staff member is on field.
I	Interaction	a deliberate interaction between two or more people processing data is occuring.	Two or more people are speaking on data.	One person knocks on the door of a room where there is another person and then, without interacting, leaves a document on the table. Two or more people greet each other.
C	Creating data	a new data (physical or digital) is created	A person is filling out a paper form. A person is writing an email.	A person is photocopying a document. A person is copying a file.

5.5.4 Violation Categories

A careful inspection of the units allowed us to define three privacy risk categories resulting by the combination of their impact on people and their likelihood to be detected. In particular, we focused on the technical means and the chances of attackers being

discovered along the methodology integrating safety and security by Eurocontrol. The risk of a violation is determined according not only to the type of information disclosed, but also the location, and who can access information.

High Risk violation happens if data can be directly used to impact the well being of a pupil / teacher without particular technical means and with low chances of being discovered while accessing the information; *Medium Risk* when data can be used for discrimination against a pupil / teacher or can be used as an escalating vector to acquire additional information by parties without major skills that would have otherwise high chance of being detected while accessing the information; Finally *Low Risk* data disclosure happens when data can only be used by parties who already know the information or can easily acquire it. Significant technical skills would be required to further escalate it, and the chances of being detected while accessing it would be extremely high. Each unit reporting a violation was marked with the code *High*, *Medium*, or *Low*.

The causes of violations are analyzed using a 2x2 matrix, which correlates the cause with the type of lack in data processing. Table 5.5 illustrates the causes of violations using a 2x2 matrix. It distinguishes the causes into structural, if a change in organizational processes can directly mitigate them, and occasional if they are both due to individual staff behavior and can be mitigated by increased behavior compliance with organizational process requirements. Processing failures are compliance failures in the absence of a procedure, in the presence of a poorly designed procedure that does not comply with regulations, or even in the presence of a well-designed procedure not executed by staff. In contrast, resource failures are characterized by the existence of a procedure (or an established practice) that is poorly designed and fails due to insufficient resources for its completion.

5.6 Results

5.6.1 Staff's knowledge about data protection

A DPO was appointed at each EI and only some member of EI1 and EI2 were either unaware of its existence at school (23.5%). In EI3, which performed a specific training, most participants stated that the schools have a DPO and that they know how to contact them (83.3%). During the observation periods, DPOs visited the schools and met staff's members QID775: "*In the room of Manager A, HED2 and MT02 meet with the DPO to take stock of the situation, discuss issues and acquire opinions*"). All EIs have a corporate data protection policy, and only 19.3% of staff members do not know

Table 5.5: Matrix of the causes of violations in data protection

Cause	Structural cause	Occasional cause
Lack of compliance	Either there is no business procedure or a procedure (or an established practice) that, being poorly designed, fails to comply with current regulations. The organizational process has regulatory nonconformities that cause the processing to fail precisely because the staff complies with the procedure's requirements.	The staff could perform a well-designed procedure, but it is not. Individual staff behavior may be determined by i) particular circumstances and ii) autonomous and personal adoption of practices not authorized by the educational institution.
Lack of resources	A procedure (or established practice) that is poorly designed fails because there are insufficient resources to complete it. The organizational process has conflicting requirements that cause the processing to fail precisely because the staff meets the procedure requirements.	Although well designed, there is a procedure that fails because of insufficient resources to complete it. Individual staff behavior tries to compensate for the lack of resources by deviations: i) with practices not authorized by the educational institution; ii) with specific actions determined by particular circumstances.

its content.

The major knowledge gap for rank and file members of staff arise in the *specifics of the GDPR*: all EIs have a register of processing activities but 54.1% of participants do not know what this document is and 16.3% knows what it is but think the school does not have it. Similarly, staff's member ignore the meaning of data protection impact assessment (66.3%), and formal procedure for data subjects' rights management (60.2%). We believe this is not only understandable but even legitimate. As a rank-and-file staff you are supposed to know what a policy is and follow it, but not necessarily how it is stored. Section X.10 of Appendix X contains some plots reporting the results mentioned above.

5.6.2 Physical secure IT equipment in edu spaces

The researcher made 150 observations in classrooms and laboratories in which at least one there was a computer: 84 of 150 observations (56%) were unique. In 11 out of 150 cases (7.3%), it was not possible to detect whether the operating system was locked or not, while in 37 cases (24.6%), the computer was turned off. In the remaining 102 cases (68%), the computer was turned on. More specifically, in 38 out of 102 cases (37.2%), the computer was locked or in use by a staff's member, while in 64 out of 102 cases (62.8%), it was unlocked and not in use. Analyzing in deep these last 64 cases, we found that the room was empty in 22 occurrences (34.4%), whereas a member of the staff was in the room in the other ones.

We found laptops in 57 out of the 150 annotations concerning rooms with electronic equipment inside: 35 (61.4%) were without an anti-theft cable, while 22 (38.6%) had an anti-theft cable. All the laptops equipped with anti-theft cables were at the same EI. Checking for a double violation is helpful to assess the real risk level for data breach, so that we analyze for example how many times staff's members left the room unlocked with a laptop inside without an anti-theft cable. We found that the room was empty in 19 cases (54.3%), while the room remained handled by a staff member in 16 cases (45.7%). When laptops were without the anti-theft cable, we found that most occurrences were recorded in classrooms or laboratories (17 out of 19 of the empty room cases) and at the specific times of the day detected for the case of the unlocked computer. Hence, the mobility of the staff's member across classrooms (or labs) and the possibility to access shared computers in such locations is a critical factor for data violations.

Table X.10 in Section X.9 of Appendix X contains the text of the quotes. Violations are almost all related to doors left open (QIDs 794, 1589, and 1596), unmanned classrooms (QIDs 234 and 734), or the presence of unlocked computers (QIDs 236, 880, and 1597). Moreover, we discovered further violations related to the failure to close doors in the server room (QID 794), in classrooms during parent-teacher meeting (QID 1589), and in rooms where there are video surveillance system monitors (QID 1596). We detected examples of unmanned classrooms during lunch break (QID 234), or when teachers did not use the classroom for classes (QID 734). Finally, we detected situations in which a computer was turned on, but not in use and without a lock screen both in the faculty room (QID 880) and classrooms (QID 1597 for portable systems or QID 236 for the fixed ones). The absence of an anti-theft cable connected to the notebooks is a separate issue since the EI's specific organizational choice determines it. If two or more of these violations co-occur, a multiplicative effect significantly increases the effects of individual violations. If, for example, one considers an unmanned room and a computer unlocked, each violation amplifies the consequences of the other.

5.6.3 Physical secure IT equipment in admin spaces

In assessing the physical security of computer equipment used by administrative staffs, we did not observe any cases of laptops not being connected to anti-theft cables. At each EI, we observed a number of staff members varying from two to six, making the case of computers being left unlocked in unsupervised rooms marginal. We detected, however, a few cases of such a violation in each EI especially when people were on duty

Table 5.6: Unlocked computers in the administrative offices

Institution	**No. People**	**No. Cases**	**% Cases**	**% People**
EI1	7	17	22,9%	28%
EI2	6	9	12,2%	24%
EI3	12	48	64,9%	48%
Total	25	74	100%	100%

according to a roster (e.g., on Saturdays or specific time slots). Overall, we detected 74 occurrences of computers left unattended and unlocked. Table 5.6 displays the count of individuals engaged in data processing and the instances of unlocked computers detected at each EI. An increase in the number of people involved in the processing often results in an increased number of occurrences detected.

Since the number of people observed is limited, the effects of individual behaviors are significant. For example, at EI3, the person who left the computer unlocked the most times determined 33.3% of the occurrences. Similarly, at the same institution, the number of occurrences determined by the first four person who do not lock their computers is 70.8% of all occurrences. By structuring the annotations on a sample basis to cover the entire work week and hourly schedule, we can limit distorting effects related to specific staff activities on a particular day.

5.6.4 Folklore and staff habits

During the field study, some cases deserving a specific discussion occurred. Although these cases are part of the folklore and staff habits, they are not interesting from the perspective of engineering good organizational privacy practices. In the qualitative analysis of the results, we report only those more significant violations that can be ranked as high risk violations (see 5.5.3).

There are obviously cases that result from the habits of a specific staff member. For example (QID 1156) , a person left an ID card unattended on the desk while the room was empty. In another example (QIDs 68 and 189) , a person used to leave post-it notes containing personal data stuck on their desk or monitor. Further, some staff's member used EI's IT resources for personal purposes, for example, participating in a video conference using a school's computer outside working hours (QID: 522). Table X.10 in Section X.9 of Appendix X reports the corresponding quotes.

Qualitative results highlight the most critical implication for small organizations: the ubiquitous lack of resources. In this respect, the most critical instances to assess a possible violation and corresponding structural mitigations are those in which *the staff*

Table 5.7: Examples of high risk/low risk violations observed in the field.
Context matters. The information of 962 and 970b are both disclosures of the very same personal information. Yet, the former is a 'life too short' violation as it is already known to all people who can likely read it. The latter can be used by a third party (e.g. a parent of child not belonging to the same class) who come to talk to a teacher to potentially molest a child going home alone.

ID	Room	Where	Type of data	Visible by Third Parties	Risk
962	Classroom	Inside the classroom	The children's names and means of transport for leaving school (e.g., alone on foot, by bus).	No	Low
970a	Faculty room	A paper behind the door	Children's personal data (e.g. first and last names, whether they take religion hours or not).	Yes	Medium
970b	Faculty room	A paper behind the door	Children's first and last names, the class they attend, whether they eat in the canteen, and means of transport for leaving school	Yes	High
946	Classroom	A paper on walls	Children's names.	No	Low
973	Faculty room	On the bulletin board	Documents reporting teachers' first and last names, such as the substitution sheet, hearing sheet, and committee attendance sheet.	Yes	Medium
908	Laboratory	A paper on the walls	Computers' accounts and passwords.	Yes	investigate (?)

member moves on, leaving the room unattended, fails to lock it, and leaves an unlocked computer inside or some physical private data.

5.6.5 Not all bullettin boards are equal

Let's consider a common practice at schools: posting information on bulletin boards, walls or classroom's door. The kind and the nature of such posted information broadly change and, sometimes, personal information are disclosed causing high risk privacy violations. Table 5.7 shows some examples of field observations that display privacy violations.

Consider row (QID 962) and (QID 970b). The information is essentially identical. It is a sheet reporting how children of that class go home (e.g. alone, by bus and so on). Knowing how children go home is definitely a sensitive information exposing children to several risks, such as abduction by a malicious party. In (QID 962) the information disclosure, however, happens in a location accessible only to pupils and school staff of the class who *already* know the information and in the case of the teacher have a *need to know* to make sure that a pupils should be accompanied to the bus. For a replacement teacher this is information that can be quickly accessed while escorting the pupils out of the class. For that reason, the violation is marked as low risk. An entirely different set-up is the posting in (QID 970). The particular faculty room is open to

third parties, in particular parents of pupils wishing to talk to the teacher(s). However, the information extend to all classes, so a parent might get information about children of classes different from the classes of ones' own. Knowing whether a child attend (or do not) religious classes can be used for discrimination but this is at a difference level than knowing that the child is authorized to walk home alone.

Another minor violation is that one labeled by the QID 946. On the walls there is a sheet with the list with the first and last names of students and their annual duties (e.g. cleaning the blackboard, emptying the rubbish). Although, the sheet disclose personal data, the potential breach of personal data due to the reading of this list of names (without photos) by a third party is low risk.

Knowing a teacher's first and last name and hearing or committee attendance plan is not a major privacy breach because it concerns their teaching role (see QID 973). One may argue that such information should be public.

Finally, another violation consists in posting on the lab bulletin board personal login credentials to lab computers (QID 908). The traditional IT approach would scream for severe security violations. We instead marked it as covering the entire spectrum. The key issue is what can be done on these computers: if the room is normally locked, there is no unregulated access to the Internet, and students can only access it to run education programs, the biggest risk is that they would copy each other's assignment. While this might be an academic violation, it is *not* a privacy violation. If the computers allow access to the personal emails of the students as provided the school this might classify as a medium violation as some official communication between the school and a student might be captured by other students logging as him or her. Finally, this might be classified as high risk if the computers allow unfettered access to the Internet as one student could impersonate another student with potentially severe consequences. Table X.10 in Section X.9 of Appendix X contains quotes' text.

5.6.6 Staff incorrect use of tools

Some violations involved the security requirements induced by the GDPR concerns how people use tools such as computers possibly containing sensitive information.

Table X.9 in Section X.8 of Appendix X reports some instances of them.

For example, having an unlocked computer (or not connected to an anti-theft cable) in an unattended room may become a more critical issue during lunch breaks or in the interval between two lectures. It's extremely unlikely that an attacker would enter a room with a staff member present, sit at a computer that isn't theirs, and start accessing

data without being questioned about their were doing.

5.6.7 I'm called elsewhere effect

Sometimes a staff's member is solicited with requests for activities to be carried out simultaneously. This situations force the staff's member to interrupt a task in progress that then remains unfinished, subsequently leading to a failure. For example, a staff's member who was filing folders in a locked cabinet was forced to rush out of the room, resulting in forgetting the open folders on their desk (QID 1092). Due to an inadequate service coverage, the same staff's member often faced with conflicting tasks simultaneously. Another example is reported in QID 884: the janitor's office is empty because JAN6 is busy on another floor where no janitor is available. In five cases, we observed the faculty room with unlocked computers. Such observations were randomly made during the morning while teachers lectured in classrooms. In further 17 cases, we observed this dual circumstance in classrooms and laboratories, but at specific times: during school recess, lunch break, time and teacher change, or when the classroom/laboratory was not used for classes. We observed that people in secretarial offices have the tendency to leave their computers unlocked when they have to be absent from their workstations. The most common reasons attributable to this behavior are (i) going to get print out from the network printer (which is almost always outside the room); (ii) going to make photocopies; (iii) going to deliver physical documents to someone; (iv) attending in-person meetings. Some people were observed turning off the monitor instead of locking their computer, leaving it unlocked. Interestingly, this behavior was detected among operators from the same office but only at EI1.

5.6.8 Resources don't live up to expectations

We found mistakes in the organizational procedures adopted by the analyzed EIs due to lack of physical resources. For example, dialogues with parents were held in places or premises unsuitable to protect the content of confidential conversations between teachers and parents. An unsuitable place is the hallway, where *"it is possible to hear and understand what some teachers (who do the dialogues down the hallway) say to parents"* (QID 1590). Another example is that one of a room with poorly insulated walls where *"it is possible to hear clearly and loudly everything that is being said in the room next door [...] where a teacher is talking to parents about educational evaluations given to some students"* (QID 467).

Lack of economic resources also encourages faulty design of procedures which are amplified by the move-on effect. For example, we observed the malfunction of an automatic locking mechanism of a main entrance door, which could not be repaired due to lack of resources. To remedy the problem, the institution designed a procedure requiring a janitor to manually close the door after visitors entered (QID: 928). Unfortunately, we observed one case in which the janitor could not provide the door closure because he was busy elsewhere, resulting in the burden of closure being entrusted to third parties visiting the institution (QID: 929). Table X.11 in Section X.9 of Appendix X reports the text of the above mentioned quotes.

5.6.9 Who watches the watchers?

Regarding video surveillance monitoring systems, we found them in EI1 and EI2, and we identified four types of problems: (i) unauthorized third parties who may watch video surveillance system monitors; (ii) erroneous definition of access criteria for monitoring video surveillance systems; (iii) failure to guard the room where there is a video surveillance monitoring system; (iv) failure to lock the rack of the video surveillance system.

A problem we detected concerns a staff's member leaving the room where the monitoring system is located without locking the door (coded using the string *A-IA-M*). This behavior could produce a possible indirect interaction with video surveillance data (coded as *TP-II-DI*). Such kind of failure is marked as a compliance failure (code *F*).

Another example is about the placement of the monitoring system in a staffed location (e.g., an administrative office) accessible to third parties. As in the previous circumstances, a possible indirect interaction could occur. In this case, however, the failure relates to a need for more resources (lack of suitable rooms) that prevented the proper design of the business procedure (codes *F-R*). In another case, unauthorized individuals (e.g. janitors) monitored the video surveillance system (coded as A-IA-Q-DTT-DI). Thus, a transient digital data transfer occurred due to a lack of human resources caused by an incorrect business procedure design (codes F-R).

We encountered several violations in which occur missing interactions with people in charge of manning the box where the video surveillance monitoring system is placed. For example, staff's members had to leave because they were engaged elsewhere in other activities (we coded it as A-IA-M to indicate the staff's members leaving and as *A-MI-TP* to indicate the failure to interact with third parties). In such a case, third parties could interact indirectly with data (*TP-II-DI*). Here, the problem lies in the lack of

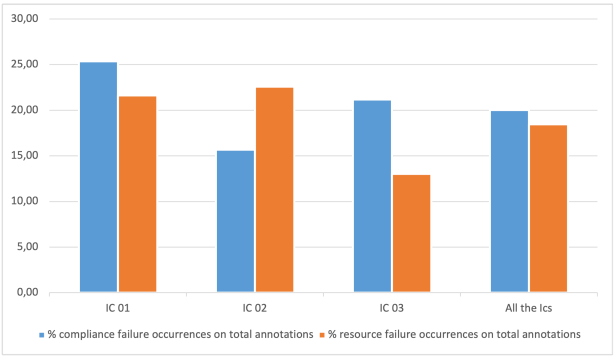

Figure 5.2: Number of compliance or resource failures on all annotations.

human resources. The procedure, indeed, requires that a staff's member performs two contradictory tasks simultaneously; hence, a resource failure happens (codes *F-R*).

5.6.10 The shadow IT

In classrooms and laboratories, we observed many IT solutions introduced by the teaching staff: the shadow IT. Haag et al. [71] define it as an IT entity (hardware, software, or services) that a worker builds, introduces, or uses for the job without prior informing the organization or acquiring its approval. Some authors [171] show how shadow IT is ingrained in the IT environments of higher education institutions and reveal that they can create new attack vectors. Specifically, we found that some teachers used their laptops for lectures by connecting them to the institution's network, acting like 'unmanaged PCs.' For example, a teacher used a personal laptop in the classroom (QIDs: 780 and 1574) or in the teaching room (QID: 777). According to van Acken et al. [171], self-acquired devices are the more frequently shadowed IT category. Table X.11 in Section X.9 of Appendix X reports the text of the above mentioned quotes.

5.7 Summary of failures

We detected overall 613 failures: 294 (48%) resource failures, and 319 (52%) compliance failures. This result denotes that, in almost half of the annotated cases, a staff's member fail to protect data while acting according to company rules or practices. Figure 5.2 shows (overall and by EIs) the percentage of units coded as compliance failure and resource failure, respectively. From the pattern, it is evident that about two-fifths of the annotated units of analysis result in a failure (compliance 20%, resource 18.5%). In addition, while resource failures prevail in EI2, compliance failures are the majority in

EI1 and EI3. Interestingly, failure of compliance is less prevalent in EI2, where 83.3% of respondents say the institute has a corporate data protection policy. In EI2, a 40% of respondents say they know the content of the policy (in line with data from other institutes), and another 40% say they do not know the policy because the corporate data protection rules are based on practice (which is a peculiarity of this institute). Thus, EI2 is characterized by a very practice-focused organizational structure, evident in interactions with the administrative staff. The scarcity of economic resources also results in high staff turnover. e.g. "*HED1 told that it is difficult to share procedures in the working group because this changes often*" (QID: 419).

We used the codes of Section 5.5 to understand how to classify the results. The sequences of codes (i.e., events) are the keys to findings, the threads linking all the anecdotes of Section 5.6. Figures 5.3 and 5.4 show the codes that preceded a failure: event X took place before event Y, or event Z was absent. For instance, we are concerned about unlocked PCs because the teacher doesn't lock the room when leaving (T+IA+DI+M+F sequence).

The annotations show a predominance of compliance failure in units in which digital data are involved (52.7%), followed by those involving physical data (42.6 %) and transient audio data (21.3%). The first data is probably higher considering the difficulty of determining precisely the actions performed by staff's members on digital data by observing their activities (see Subsection 5.5.3). In the case of resource failures, physical data involvement is slightly predominant (46.6%), followed by digital data (40.1%) and transient audio data (15.6%). However, the latter case is explained by the many annotations determined by inadequate logistical equipment or practices established at the corporate level in storing data in physical form (see Section 5.4). Similarly, Marjanov et al. [106] pointed out that most security of processing violations (ex Art. 32) of the GDPR involve data in digital format.

Table X.12 in Section X.9 of Appendix X contains the text of the quotes referred to all analyzed cases.

5.7.1 Missing Interactions as a Root Cause

Figure 5.3 shows the number of failures occurred with missing interactions (MI) and/or moving on (M). Figure 5.4 shows the number of failures due to the co-occurrence of a MI or a M with respect to the total number of failures.

The analysis of the data shown in the figures clearly shows that most violations were *preceded by* interactions in which people 'left the ball drop midway' (MI) or they were

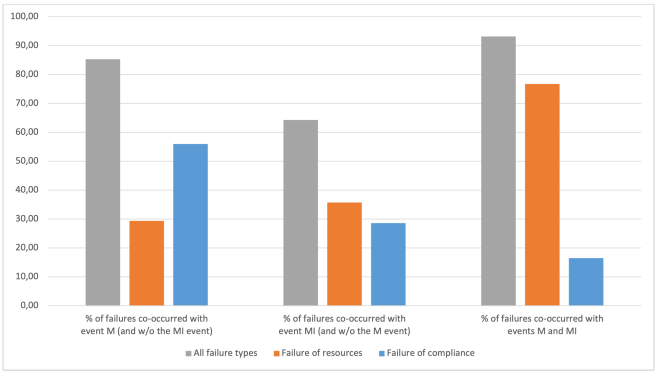

Figure 5.3: Number of failures related to missing interaction (MI) and/or move on (M)

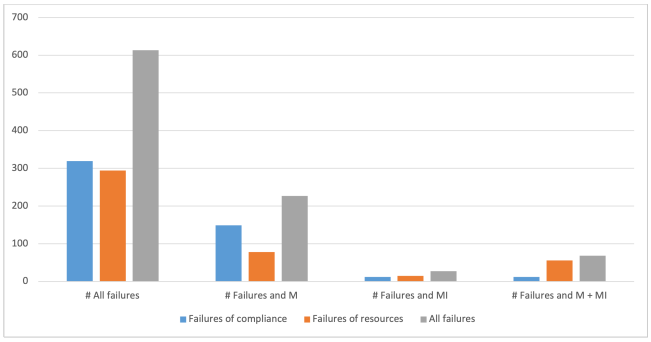

Figure 5.4: Number of failure co-occurred with MI or M with respect to the total number of failures

interrupted during the processing (M). The highest probability of failure was indeed observed when such events co-occurred (93.1%): compliance failures occurred in 16.4% of cases. whereas resources failures in the remaining 76.7%. When MI was not followed by M, we had 28.6% failures of compliance and 35.7% failures of resources. Conversely, when M was not followed by MI, we detected 56% failures of compliance and 29.3% failures of resources.

The co-occurrences of M or MI and a resource failure happened with unattended rooms (classrooms, laboratories, faculty room, administrative and janitors office) with open doors due to an established business practice. It may be related to a classroom (QIDs: 234, 626, 879, and 1417), a laboratory (QIDs: 309, 680, and 1444), the faculty lounge (QIDs: 484, 339, and 628), the janitorial office (QIDs: 884), or an administrative office (QIDs: 27 and 567). There are other cases in which there was a lack of manning at the EI's site entrance (QIDs: 623 and 1414) or in the hallways (QIDs: 834 - 923 - 988) because janitors were called to perform other duties. Two remarkable cases were observed. In the first case, due to turnover, the members of a workgroup changed unexpectedly near a regulatory deadline. This change led to mishandling, it

was unclear, indeed, who was in charge of what (QIDs: (568 - 577). In the second case, we observed that the corporate server and removable hard drives containing data backup were located in an unattended office having its door open (QIDs: 85). A design error in the processing procedures is evident in both these examples.

The co-occurrences of M or MI and a compliance failure are associated with unattended rooms or open doors due to a personal mistake by staff's members who chose not to comply with school procedures. Examples include the temporary absence of staff from administrative offices (QIDs: 203 - 498 - 513) or classrooms (QID: 889) to get a paper document from the network printer or to deliver it to someone. More critical cases occurred. For example, a staff's member left by mistake unattended and with an open door the room with surveillance system monitors. Mistakes are the sources of such failures.

5.7.2 Structural Violations

Incorrect design choices of actual security procedures are an incentive factor for the occurrence of security violations. Often, the seriousness and erroneousness of some of these design choices are trivial, as in the case (QID: 971) of the WiFi access password written on the whiteboard or the list of student access credentials hanging on the lab bulletin board (QID: 908). Sometimes the fallacy of a design choice is unclear, such as using a shared account and operating system autologin mechanism (IDs 300 and 332) with a subsequent application login phase for users. This configuration is particularly critical in laboratories where we found that it is possible to read the e-mails of students who had not logged out of the application.

5.8 Risk-based observations

To formulate the risk-based observations, we analyzed the following data sources related to Italy cases:

1. rulings of the Italian Supervisory Authority (the *Garante per la protezione dei dati personali*) found on GDPRxiv [3];

2. rulings of the Italian Supervisory Authority related to Security Violations applied to schools.

We investigate the rulings in Italian for homogeneity toward the schools considered that are located in the territory of the Italian Republic for the following two reasons:

1. The European reference legislation (the GDPR) is the same;
2. The Italian supervisory authority coordinates with its European SA counterparts in assessing the interpretation of the GDPR and defining homogeneous criteria for determining the type and amount of sanctions to be issued to non-compliant parties.

The GDPRxiv service database [4] collects a selection of GDPR injunctions, judgments, opinions, reports, and guidelines from all official GDPR sources from 25/05/2018 to 29/06/2023. The portal contains 4263 records related to measures provisioned by the supervisory authorities of 25 EU member states and the EDPB. Of these records, 388 referred to Italy (measures of the Italian Supervisory Authority), of which only 27 related to educational institutions. I conducted the analysis considering homogeneous data related to the Italian Supervisory Authority (SA) both because it is strongly represented (only the Spanish one has more records) and because the field study was conducted on Italian educational institutions and, therefore, under the direct jurisdiction of that SA. The Italian SA's measures database contains the complete list of GDPR injunctions, rulings, opinions, reports, and guidelines issued by the authority, for which online publication is required. A total of 12439 published measures were published, of which 170 were referable to educational institutions (1,37% of the total). I analyzed them by considering two subsets of these measures, namely those issued after the full entry into force of the GDPR (25/05/2018) and, respectively, the date of the last measure included in my dataset (16/11/2024) or the date of the last measure included in GDPRxiv dataset (26/06/2023). In the period from 25/05/2018 to 16/11/2024, the Italian SA published 2561 measures online, of which 58 related to educational institutions (2,07% of the total), while in a more limited period from 25/05/2018 to 26/06/2023, the SA published 1666 measures, of which 45 related to educational institutions (2,70% of the total). It shows that only 60 percent of the total measures issued by the Italian Supervisory Authority for educational institutions belong to the GDPRxiv dataset.

We point out that investigating the extent and prevalence of privacy violations through the investigation of sanction casuistry returns a partial picture of what happens in the real world. The possible absence (or scarcity) of a sanctioning case history may result from multiple factors, such as:

- the absence of complaints filed with the supervisory authority (e.g., because the data controller, when approached by the data subject, has resolved the alleged violation, making it unnecessary to file a complaint with the supervisory authority);

- the difficulty of ascertaining the case (e.g., in a complaint submitted to the *Garante* on the erroneous methods of administration of the surveys for organizational well-being organized by an education institution, there was no response because the education institution asserted to have decommissioned the database);
- the failure to publish the measure on the supervisory authority's website (because, the publication of the measure is an accessory sanction).

In addition, many measures arise due to the inspection activity carried out by the Italian supervisory authority. On its initiative, this inspection activity is also carried out employing the *Guardia di Finanza* Corps [125] and is determined by the "*Garante per la protezione dei dati personali*" on a semi-annual basis. In recent years (January 2018-December 2023, [135], [134], [130], [131], [126], [127], [133], [128], [123], [120], [121], and [124]), such activity has not been directed at schools, except the January-July 2024 [136] and July-December 2024 [138] semesters, where the own-initiative inspection activity taken care of by the Office of the "*Garante per la protezione dei dati personali*", including employing the "*Guardia di Finanza*", has been directed, among other things, to the investigations regarding the processing of personal data carried out at educational institutions through the "*electronic register platforms and digital suites.*" Unfortunately, the sanction proceedings are still in progress. Therefore, it was impossible to detect their outcomes in the consultation of the dataset of rulings issued in the time range considered (from 25/05/2018 to 16/11/2024).

Moreover, in databases of judgments (such as GDPRxiv [4], [3], [155]), there are only the sanctioning measures, which represent a subset of the measures issued by the Italian SA, which often, especially in the case of small educational institutions (such as schools), recognizing that the non-compliant processing was not on a large scale (as it was limited to students or teachers), as well as the absence of the profit-related purposes (often the violations were determined by lack of resources or mere material error), preferred to limit itself to a simple warning. A simple warning was also issued in cases deemed particularly significant by the Italian SA, to the point that related cases were included in the vademecum for schools published by the same SA [63].

Risk-based analysis of observations' findings

Overall, the field observations revealed 14 findings, which I analyzed using a risk-based approach.

Findings are reported according to the following scheme. First, I searched the GDPRxiv dataset and the Italian SA's measures dataset for evidence that the finding

is a repeated problem that has demonstrably led to quantifiable consequences or that there exists a major privacy violation that hits the news. If yes, I marked the risk as high. If not, I checked to see if any SA's measure says the finding represents a problem that could be exploited as a vector for further data access. If I had evidence of cases where some school was convicted because the vector was exploited, I marked the risk as medium-low (or high) risk, depending on how much evidence I had or how high profile they are on the news. Conversely, I classified this as low risk if I had no record where such a breach would be impactful.

Finding 01: There are diffuse violations of basic security hygiene, such as several computers accessible without passwords. Over 27 pronunciations in the past 388 enforcements (GDPRxiv), 58 measures in the past 2561 Italian SA measures (Italian SA database), and in the last 7 years in Italian Supervisory Authority measures, no school has ever been fined for not enforcing password lock-in on screens. Hence, we consider this a low risk for general computer use. However, there are two pronunciations [132] [137] where the Italian SA observes that organizations should ensure that only authorized personnel can access personal data. Secure computer access should be considered more important for computers with secure and direct access to personal data (e.g., administrative staff). However, we could not find any record in GDPRxiv and the Italian Supervisory Authority measures where such a breach would be impactful, so we classify this as low risk.

Finding 02: There are diffuse violations of basic security hygiene in the physical protection of laptops, such as several laptops without an anti-theft cable. Over 27 pronunciations in the past 388 enforcements (GDPRxiv), 58 measures in the past 2561 Italian SA measures (Italian SA database), and in the last 7 years in Italian Supervisory Authority measures, no school has ever been fined for not physically protecting laptops anchoring them to an anti-theft cable. Hence, we consider this a low risk for general computer use. However, the GDPR (Article 32) requires that the data controller and processor implement appropriate technical and organizational measures to ensure a level of security appropriate to the risk, including encryption of personal data and the ability to ensure on an ongoing basis the confidentiality, integrity, and availability of systems. According to this provision, many measures of the Italian SA highlight this aspect. Secure computer physical access is more important for laptops that process personal data, which could be stolen, leading to unlawful processing. However, we could not find any record in GDPRxiv and the Italian Supervisory Authority measures where such a breach would be impactful, so we classify this as low risk.

Finding 03: Staff is diffusely violating the organizational procedure for

access to the institution's computers, such as leaving several unused and unlocked computers or monitors off without locking the unused computer. Over 27 pronunciations in the past 388 enforcements (GDPRxiv), 58 measures in the past 2561 Italian SA measures (Italian SA database), and in the last 7 years in Italian Supervisory Authority measures, no school has ever been fined because staff members violated the institution's regulations by leaving the computer unmanned and unlocked. However, there is an interesting measure in which the Italian SA sanctioned a hospital company for unauthorized third-party access to some patients' personal data by using the authentication credentials of administrative staff who had left their workstations open or without bothering to guard their passwords [139]. So computer access should be protected, mainly for those located on unattended premises. If computers allow direct access to personal data processing systems (such as for administrative staff or faculty stations in the faculty lounge), logical access protection becomes a priority. However, we could not find any record in GDPRxiv and the Italian Supervisory Authority measures where such a breach would be impactful, so we classify this as low risk.

Finding 04: There are diffuse violations in the physical control of access to the premises, such as unmanned and unlocked classrooms or rooms. Over 27 pronunciations in the past 388 enforcements (GDPRxiv), 58 measures in the past 2561 Italian SA measures (Italian SA database), and in the last 7 years in Italian Supervisory Authority measures, no school has ever been fined because of a lack of physical control of access to premises. I got a similar finding by consulting the Supervisor's other 2561 measures not directly related to the schools. However, it is established in the literature that physical controls are a critical aspect of cybersecurity to protect against physical threats to systems and data, such as damage, tampering, and theft. So, the sensitive areas should be under access control to reduce the risk of unauthorized access, and institutions should adopt additional physical security measures to protect data and devices. However, because I could not find any record in GDPRxiv and the Italian Supervisory Authority measures where such a breach would be impactful, so I classify this as low risk.

Finding 05: There are diffuse violations in the physical control of access to technology rooms, servers, and network equipment, such as unmanned and unlocked rooms. Over 27 pronunciations in the past 388 enforcements (GDPRxiv), 58 measures in the past 2561 Italian SA measures (Italian SA database), and in the last 7 years in Italian Supervisory Authority measures, no school has ever been fined because of the absence in physical access control to sensitive areas, such as server rooms or technology rooms. I got a similar finding by consulting the Supervisor's other 2561

measures not directly related to the schools. However, the literature establishes as a critical point that all sensitive areas are under an access control procedure and that the data center is protected, aiming to protect the systems and data contained in it. Moreover, the data center (or, better yet, the server room in the school's case) should be considered a physical security perimeter according to Annex A 7.1 of the ISO 27001:2022 standard. However, I classify this as low risk because I could not find any record in GDPRxiv and the Italian SA measures where such a breach would be impactful.

Finding 06: Diffuse violations exist in the design and management of the video surveillance system, such as monitors or equipment located in unmanned premises, incorrect positioning of monitors whose contents are visible to unauthorized third parties, or an erroneous definition of access criteria for monitoring it. Over 27 pronunciations in the past 388 enforcements (GDPRxiv), 58 measures in the past 2561 Italian SA measures (Italian SA database), and in the last 7 years in Italian Supervisory Authority measures, no school has ever been fined for not properly implement and manage video surveillance systems. However, the Italian SA's general measure on video surveillance [119] requires that persons authorized to enter the premises where the monitoring stations are and authorized to view the images or perform further operations be allowed to a limited number of people who can only access them after authentication and that are previously punctually identified and named. It follows that the presence of monitors displaying video surveillance camera images in time on-premises accessible to the public (such as the student secretary's offices) or visible from the hallways (such as the janitor's desk) constitutes a wrong definition of access criteria and a violation of legal requirements by allowing unauthorized parties to view the video surveillance system images. However, I classify this as low risk because I could not find any record in GDPRxiv and the Italian Supervisory Authority measures where such a breach would be impactful.

Finding 07: There are diffuse violations of institutional procedures regarding publishing or posting documents containing unnecessary personal data (like passwords or data revealing religious convictions) on physical bulletin boards or in other non-restricted areas of the school. Over 27 pronunciations in the past 388 enforcements (GDPRxiv), in 20 cases, the schools received an administrative sanction from the Italian SA for posting excessive and unnecessary staff or student personal data online (on the website or the bulletin board) (16 cases), posting such data in restricted areas of the electronic register (3 cases), or posting such data on the physical bulletin board (1 case). In addition, over 58 measures in the past 2561 Italian SA measures (Italian SA database), and in the last 7 years in Italian Su-

pervisory Authority measures, the schools received a measure from the SA in 29 cases for the same reasons. The measures consist of an administrative sanction or a simple warning. In detail, in 23 cases, the schools were sanctioned for posting excessive and unnecessary staff or student personal data online on the website or the bulletin board (19 received an administrative sanction, while 4 received a simple warning). In addition, in the other 6 cases, the schools received an administrative sanction for posting staff or students' personal data in restricted areas of the electronic register (4 cases, plus a single case where the school received a simple warning) or posting such data on the physical bulletin board (1 case). Although the Italian SA measures referred to personal data published online or in restricted areas of the electronic register, the observations carried out at the involved education institutions reveal an issue in the physical publication of documents containing personal data. Because the schools received a sanction online, a physical risk can be considered a medium risk according to the number of penalties received from schools. Table 5.8 lists the document IDs of the cited measures and pronunciations on the Italian SA's website.

Finding 08: There are diffuse violations in the access control to the school's buildings procedure, such as unmanned access to the building or malfunction of an automatic door locking mechanism. Over 27 pronunciations in the past 388 enforcements (GDPRxiv), 58 measures in the past 2561 Italian SA measures (Italian SA database), and in the last 7 years in Italian Supervisory Authority measures, no school has ever been fined for not correctly implementing physical security measures at boundaries. Hence, I consider this a low risk for a data breach. However, according to Annex A 7.1, control of the standard ISO 27001:2022 [60], it is crucial to safeguard the institutions' resources and premises by implementing physical security measures. According to this vision, it is critical to identify the physical boundaries of buildings or areas and to control who has access to them, when, where, and how. Hence, a building without someone controlling people's access or a faulty door-locking mechanism does not protect the physical perimeter from unauthorized access. However, I classify this as low risk because I can not find any record in GDPRxiv and the Italian Supervisory Authority measures where such a breach would be impactful.

Finding 09: There are diffuse violations of the confidentiality of conversations due to inadequate premises, such as dialogues with parents held in places or premises unsuitable for protecting the confidentiality or the presence of several people for different processing in the same room. Over 27 pronunciations in the past 388 enforcements (GDPRxiv), 58 measures in the past 2561 Italian SA measures (Italian SA database), and in the last 7 years in Italian Supervisory

Table 5.8: Document IDs of the Italian Supervisory Authority measures and pronunciations divided by the type of publication carried out by the schools.

The "*GDPRxiv (Yes/No)*" field indicates "*Yes*" if the document is both contained in the GDPRxiv dataset and "*No*" otherwise. In the "*Simple warning (Yes/No)*" field, "*Yes*" is indicated if the measure of the Italian SA was a simple warning or "*No*" if the measure was an administrative sanction. The documents can be found by conducting a search at the URI: www.garanteprivacy.it.

Document ID	**Publication type**	**GDPRxiv (Yes/No)**	**Simple warning (Yes/No)**
9574101	Online publication (web or bulletin board)	Yes	No
9283029	Online publication (web or bulletin board)	Yes	No
9685245	Online publication (web or bulletin board)	No	No
9572226	Online publication (web or bulletin board)	No	No
9283014	Online publication (web or bulletin board)	Yes	No
9718196	Online publication (web or bulletin board)	Yes	No
9365159	Online publication (web or bulletin board)	No	Yes
9451734	Online publication (web or bulletin board)	Yes	No
9576756	Online publication (web or bulletin board)	Yes	No
9445324	Publication on physical bulletin board	Yes	No
9565218	Communication to third parties	Yes	No
9896845	Online publication (web or bulletin board)	Yes	No
9789541	Online publication (web or bulletin board)	Yes	No
9718134	Online publication (web or bulletin board)	Yes	No
9075229	Online publication (web or bulletin board)	No	No
9828059	Online publication (web or bulletin board)	Yes	No
9688099	Online publication (web or bulletin board)	Yes	No
9987578	Online publication (web or bulletin board)	No	No
9960948	Communication to third parties	No	No
9902516	Online publication (web or bulletin board)	No	Yes
9920274	Online publication (web or bulletin board)	No	Yes
9365147	Publication in restricted area of electronic register	No	Yes
9852255	Publication in restricted area of electronic register	Yes	No
9811361	Publication in restricted area of electronic register	Yes	No
9777200	Publication in restricted area of electronic register	Yes	No
9711517	Publication in restricted area of electronic register	Yes	No
9777156	Communication to third parties	Yes	No
9776444	Communication to third parties	Yes	No
9897931	Communication to third parties	No	No
9742923	Online publication (web or bulletin board)	Yes	No
9463997	Online publication (web or bulletin board)	No	Yes
9857610	Online publication (web or bulletin board)	Yes	No
9697781	Online publication (web or bulletin board)	Yes	No
9583865	Online publication (web or bulletin board)	Yes	No

Authority measures, no school has ever been fined for not guaranteeing the confidentiality of conversations due to inadequate premises. Hence, I consider this a low risk for confidentiality. However, according to Annex A 7.1 (Physical Security Perimeters) of the ISO 27001:2022 standard, it is crucial to take under control who can have access to a room and when, as well as who can see or even hear into the office from inside or worst from outside. Analogously, it is crucial to pay attention to documents or information (like information on whiteboards or flipcharts) posted in the room that can become accessible to third parties when they are inside the room. However, I classify this as

low risk because I can not find any record in GDPRxiv and the Italian Supervisory Authority measures where such a breach would be impactful.

Finding 10: There are diffuse security violations in data processing due to the scarcity of human resources performing multiple processes simultaneously, such as activities carried out simultaneously, interruption of a task in progress, or frequent staff moving from one place to another. Over 27 pronunciations in the past 388 enforcements (GDPRxiv), 58 measures in the past 2561 Italian SA measures (Italian SA database), and in the last 7 years in Italian Supervisory Authority measures, no school has ever been fined because of workforce-related risks. Hence, I consider this a low risk for personal data processing at the workplace. However, managing human resources risks is a real problem and has repeatedly been investigated [14], such as when scarcity compelled them to perform multiple treatment activities simultaneously (and sometimes chaotically). Therefore, improving the organization of work activities by ensuring that staff members can process continuously and organize without being continually interrupted can improve data protection by reducing the risk of human error or individual adoption of inadequate operating practices. It would enable work to be carried out without violating institutional procedures and current regulations. However, I could not find any record in GDPRxiv and the Italian Supervisory Authority measures where such a breach would be impactful, so I classify this as low risk.

Finding 11: There are diffuse violations of the institutional rules on IT resource management, such as using institutional IT resources for personal purposes or connecting non-prior-authorized personal devices to the institutional network. Over 27 pronunciations in the past 388 enforcements (GDPRxiv), 58 measures in the past 2561 Italian SA measures (Italian SA database), and in the last 7 years in Italian Supervisory Authority measures, no school has ever been fined for not properly implement and manage a Bring Your Own Device (BYOD) procedures. However, from the literature, it is known that the use of personal devices connected to the institutional network or using institutional IT resources for personal purposes brings the protected and controlled environment of the institutional network into contact with the unprotected and uncontrolled environment of the personal sphere of staff members, exposing the corporate information system to the occurrence of unlawful processing of personal data. Thus, to implement appropriate technical and organizational measures to reduce the risk of a personal data breach, a Bring Your Own Device (BYOD) policy is recommended to establish and promote correct and conscious behaviors by staff. However, I classify this as low risk because I could not find any record in GDPRxiv and

the Italian SA measures where such a breach would be impactful.

Finding 12: Diffuse security violations occur due to individual staff behaviors that do not comply with institutional regulations, such as leaving ID cards or keys unattended on the desk or sticking Post-it notes with personal data on the desk or monitor. Over 27 pronunciations in the past 388 enforcements (GDPRxiv), 58 measures in the past 2561 Italian SA measures (Italian SA database), and in the last 7 years in Italian Supervisory Authority measures, no school has ever been fined for individual staff behavior that which could be prodromal to data exfiltration or identity theft. Hence, I consider this a low risk for personal data processing. However, there was an injunction order against a University Health Care Company where unauthorized access to a patient's personal data by staff members who were not authorized to process it happened. Despite I found no records in GDPRxiv or from the Italian SA indicating for a school a significant impact of this breach, in other realities the impact has been critical, thus I classify it as a medium risk.

Finding 13: There are diffuse security violations in document storing, such as leaving documents unattended on the desk or not storing files in the archive. Over 27 pronunciations in the past 388 enforcements (GDPRxiv), 58 measures in the past 2561 Italian SA measures (Italian SA database), and in the last 7 years in Italian Supervisory Authority measures, no school has ever been fined for not protecting the confidentiality, availability, and integrity of documents containing personal data due to mishandling of their storage. Hence, I considered this risk of low level for physical document management. However, the Italian SA has issued requirements regarding the processing of special categories of personal data [122], which, in the case of paper documents, provide for their transmission as a rule in a sealed envelope. These provisions are incompatible with the presence of documents left unattended on desks in publicly accessible rooms. However, I classify this as low risk because I could not find any record in GDPRxiv and the Italian Supervisory Authority measures where such a breach would be impactful.

Finding 14: There are diffuse security violations in document printing procedures, such as printing documents sent for printing without the user's prior authentication (with a pin or badge) or placing printers in publicly accessible areas. Over 27 pronunciations in the past 388 enforcements (GDPRxiv), 58 measures in the past 2561 Italian SA measures (Italian SA database), and in the last 7 years in Italian Supervisory Authority measures, no school has ever been fined for not protecting the confidentiality, availability, and integrity of documents containing personal data due to a wrong printing procedures or printers management. Hence, I

considered this risk a low level for physical document management. However, the Italian SA has issued requirements regarding the processing of special categories of personal data [122], which are incompatible with the presence of documents left unattended on network printers placed in publicly accessible premises, because the network's printer configuration prints them immediately when the printing job is run without verifying the physical presence of someone who can pick them up. However, I classify this as low risk because I could not find any record in GDPRxiv and the Italian Supervisory Authority measures where such a breach would be impactful.

Of course, all security and privacy holes should be plugged in (if you have Google or Meta resources). However, if you are firefighting, you may want to send your crew to first quench those fires that might burn the house to the ground in the next hour. The ones that flickered unnoticed for years might as well wait for tomorrow.

5.9 Online publications observations

The study checked the data content of EIs' online publications. While surveying the data published on the institutional websites, we experienced the impossibility of verifying and analyzing the data in the documents published in the restricted areas of the electronic register because we did not have access credentials. Therefore, the investigation focused on analyzing the public areas of the EIs' websites.

In analyzing the websites, we checked for the existence of the following elements (according to the list in Table X.1 of Appendix X):

1. to check whether pupil evaluations are posted online;
2. to check whether pupil evaluations are posted in restricted areas of the electronic register [this can be verified only indirectly by reading the content of circulars posted online];
3. to check whether there are pupils' personal data in circulars and communications directed to multiple people;
4. to check whether there are pupils' health data in circulars posted online;
5. to check whether data subjects have been informed that their personal data will be disclosed to third parties for specific activities, such as job placement and tutoring [this is very difficult to verify because it can only be done indirectly by reading the content of circulars posted online];

6. to verify whether consent has been sought from the interested parties for the online publication of photos or videos containing pupils [this can only be verified indirectly by checking the presence of the online consent forms or by analyzing the content of the school regulations];
7. to verify the presence of the disclosures for the use of platforms dedicated to distance learning;
8. to check the institute regulations regarding the possibility of recording lectures for personal use;
9. to check the institute regulations regarding the possibility of disseminating online or on social media the audio/video recording of previously recorded lectures for personal use;
10. to check the school regulations regarding the possibility of making recordings of lectures from which class dynamics emerge;
11. to check whether the class composition is published online, as well as what personal data is contained in these documents;
12. to verify whether the class composition is published in restricted areas of the electronic register, as well as what personal data is contained in these documents [this is very difficult to verify because we do not have access to the electronic register, so it can only be done indirectly by analyzing the content of circulars and school regulations published online];
13. to check whether ATA staff and document rankings are published online, as well as what personal data are contained in these documents;
14. to verify whether there is a regulation for the use of video surveillance systems and what is contained in this document [this is difficult to verify because this document is probably not published online].

Table 5.9 shows the results of the observations made on the websites of the schools involved in the study.

5.10 Discussion

Incorrect personal data processing procedures design exposes schools to GDPR non-compliance with consequent fines [106]. Insufficient organizational measures may be

Table 5.9: Observations on online publication of personal data carried out by schools

Type of data	**Processing description**	**EI1**	**EI2**	**EI3**
Student's grades	Online publication	No	No	No
Student's grades	Posting in the electronic register but not in a restricted area with limited access to the student or family	No	No	No
Personal data of pupils	Preparation of school communications not addressed to specific recipients / services related to the school activities (e.g. school bus, canteen)	Unverifiable	Unverifiable	Unverifiable
Health data	Dissemination of pupil health data (e.g., including the names of students with disabilities in a circular published online)	No	No	No
Data on educational, intermediate, and final outcomes	Communication to third parties to facilitate career guidance, training, and job placement	No	No	Yes
Audio /photos /videos recorded at school or during school-related activities	Dissemination on the Internet (website / social networks and so on)	Yes	Yes	Yes
Personal data	Use of distance education systems within institutional purposes	Yes	Yes	Yes
Audio/video shooting	Dissemination on the web and social networks of audio/video footage of a previously recorded lecture for personal purposes	No	No	No
Video Shooting	Video recording of the lesson in which class dynamics are manifested	No	No	No
Class composition data	Publication of data on the institutional website	No	No	No
Class composition data	Posting in the electronic register but not in a restricted area with limited access to the student or family	No	No	No
Rankings of teachers and administrative technical and auxiliary staff (ATA), other than first name, last name, score, and ranking position	Online publication of rankings for selection procedures	No	No	No
Biographical data and income data	Posting the names of individuals who use the canteen service free of charge for income reasons on the school website or bulletin board.	No	No	No
Video recordings	Whether video surveillance within the building is regulated and how access to recordings is managed.	Unverifiable	Unverifiable	Unverifiable
Personal data	Online publication of student names for school bus services in a form accessible by anyone.	No	No	No
Biographical and/or contact information	Collaboration contracts with external subjects.	No	No	No
Photograph and/or handwritten signature	Collaboration contracts with external subjects.	Yes	Yes	Yes
Biographical and/or contact information	Assignments to internal staff.	No	No	No
Photograph and/or handwritten signature	Assignments to internal staff.	No	No	No
Biographical and/or contact information	Curriculum vitae of collaborators.	Yes	Yes	Yes
Photograph and/or handwritten signature	Curriculum vitae of collaborators.	Yes	Yes	Yes

enough to be subject to a data protection authority (DPA) fine even without a personal data breach. Analyzing DPAs' fine in the EU Member State [59, 53, 5], Marjanov

et al. [106] show that organizational mistakes are the most costly for a non-compliant Data Controller.

Although results highlight a gap between the school staff's claims to care about privacy and their actual behaviors (i.e. the privacy paradox), already studied by Colnago et al. using an online sample of US participants [39], this is not always the case. We observed that a lack of resources pushes people to adopt organizational practices aimed at optimizing work that accidentally violates privacy requirements. For example, the governance of the EI is forced to choose a suboptimal logistical arrangement of staff due to a shortage of premises, consequently triggering inappropriate practices on the part of staff trying to remedy the problem. We observed the presence of multiple persons involved in processing activities in a single publicly accessible room in which they talk about data while third parties being there (QID: 1249). Such a promiscuity of authorized and unauthorized people can create confidentiality problems that staff's members tackle by implementing behaviors that can result in violations. For example, to talk with a colleague about an urgent and confidential matter, a staff member working in a room with third parties present "*takes a document with him and leaves the room leaving the computer unlocked*" (QID: 1494)." In other cases, people are overwhelmed by the number of tasks they have to perform, and to optimize time and resources, they adopt inappropriate practices. For example, "*while OP05 is making photocopies using the hallway copier, OP02 comes out of secretary room A and brings him signed documents. Then, because OP05 is engaged in another activity, he tells OP02 to place the documents on his desk. So, OP02 enters secretarial room B and does as he was told*" (QIDs: 1512 and 1513). In another case, multiple simultaneous tasks are assigned to the same person, so a colleague helps them by making their scans. "*Then OP05 tells OP01 to log in using his (OP05's) badge so that the scans are directly sent to him*" (QID 1519). In this sense, *there is no privacy paradox*, the one we observed is more of a concrete need to balance privacy with work leading to *accidental violations due to the limited resources available.*

The analysis reveals a dimensional problem and an organizational difference among the EIs. The resources available for data protection management are sometimes different and their number decrease in educational institutions' secondary sites especially when the secondary sites are in different buildings or located far from the primary sites. Thus, sub-optimal data protection practices occur especially in secondary sites .

Different violations were observed in educational institutions at the main and secondary sites. For example, violations about the video surveillance system or the institutional server system cannot occur in the secondary sites because such systems are

installed / managed at the primary sites only. Violations related to the missing identification of unknown people are instead more common in the primary sites. Such sites, indeed, are visited everyday by a large amount of people including third parties.

5.11 Threats to Validity

We took several methodological expedients to ensure the study's internal validity and data quality. However, the findings should be interpreted with the following limitations.

We conducted this study in Italy where GDPR is enforced. Hence, the results may only be partially replicable in contexts with a different legislation. Some of the findings might be due to particular beliefs and characteristics of the country [170] and might not generalize to other countries. To mitigate this risk, we developed a flexible methodology to try to capture general phenomena occurring in a variety of context [64]. A fine-tuning of the codebook indeed could be needed. We acknowledge that a finite and limited codebook does not allow for a fine-grained description of what occurred, but the goal was to develop a valuable and quickly extendable tool for performing a quick macro-analysis of personal data flow in small privacy scenarios such as educational settings.

Another limitation is that some aspects are not directly observable in the field. For example, it is tough to observe some digital data processing happening when EIs use third-party applications to manage emails, assignments, or parent-teacher communication. As we mentioned, EIs do not have a direct control over this type of software, which sometimes is chosen at a national level. However, it can be interesting to understand the perception of the EI's staff about this third party software. Future work could address this issue using semi-structured interviews with selected staff's members.

The observations were performed by a researcher without using any digital recording instruments. Although audio-video recordings could have provide more information about the flow of personal data, their use would have make more complex data collection and processing. Indeed, either obtaining ethical authorizations and informed consents by all parents whose children *could* have been captured, or editing videos displaying children whose parents have not given consent for filming, would have made the setup of the study particularly time-consuming and challenging. Moreover, the adoption of recording devices would have potentially affect the natural behavior of the staff and raise severe, and justified concerns from parents. The expertise in data protection matter of the researcher who carried out the field study should have limited the bias in the observations.

Audio/video would have also collected data that was privacy protected (the very case of the conversation in the hallway about pupils) without actually contributing to the research goal of identifying *structural* failures.

A limit in the observations is that the experimental protocol approved by the University's IRB asserts that researchers must not record data about parents, students or school staff who do not provide their consent. For example, the researchers will act as if the staff member who did not give consent is not present in the room. So it might well be that there were even more than two people in a room where a violation took place. Only when we write that a room was left empty it was really empty (of observable or unobservable people).

5.12 Conclusions

Field study's qualitative results offer valuable insights for governing schools and small organizations, helping them identify and correct mistakes in personal data processing procedures.

The key takeaway of this research is that privacy violations due to lack of resources when implementing the GDPR in the small, such as in schools or small companies, are ubiquitous and often insuperable. We argue that it is possible to better protect the personal data with less secure but practically feasible procedures than perfectly secure procedures that will be forcibly poorly implemented. Therefore, a more concrete approach should adopt a concrete and implementable paradigm oriented to risk mitigation. This approach may not guarantee a full compliance with all the tiny provisions and legal minutiae of the GDPR. However, in the presence of non-waivable processing, such as those of EIs, compliance with the principles of the law is still ensured in the sense of continuous improvement.

A possible solution to address this concern is staff training: one-third of the surveyed staff would have appreciated more training. However, rather than training them in the terminology of the law, making school staff more aware and attentive to data protection processing principles [76] would allow them to react better to procedures that fail and to avoid designing processes that are great on paper but doomed to fail on the field.

Acknowledgments

I would like to thank the IEs' principals for their trust and support in performing this study and all IEs' members for their help and support in this work.

I would like to thank the principals of the EIs for trusting and supporting us in performing this study and all members of the EIs for their help and support in this work.

Chapter 6

Conclusions and Future Works

The second chapter explored the dual role of the Data Protection Officer (DPO) under the GDPR, serving as both a person who controls data processing activities and an advisor to management. This duality often leads to tension, as DPOs need access to information while managers may hesitate to share it. The role becomes particularly challenging in organizations with a negative stance on data protection, where DPOs face conflicts when balancing legal compliance with organizational goals. Some studies have shown that DPOs often encounter opposition in such environments, revealing that staff members were pressured to follow organizational norms conflicting with the company's stated data protection policies over data protection laws. It creates friction for DPOs trying to uphold GDPR standards. The research was focused on Italian cases emphasizing the GDPR's uniform application across the EU and cross-border validation through extensive case reviews. While I did not provide technical solutions, helping to understand these challenges is a crucial step forward.

The third chapter highlights that replicating research in a different language is feasible but requires adaptations due to elements that do not transfer directly. Key contributions include identifying how components, like privacy policy corpora, can be adjusted to maintain comparability. The proposed case study demonstrates that, with careful adaptation, cross-language corpora can support research into how humans understand privacy policies. Future research avenues include evaluating automated analysis methods for identifying term occurrences in policies and comparing precision and recall between automated and manual coding of technical terms. Another focus is mapping technical terms across multilingual corpora, as initial findings reveal challenges in identifying and standardizing terms. Developing indicators to filter irrelevant, infrequent, or unreliable terms could enhance replication efforts, such as adapting understandability

assessments from prior studies for Italian privacy policies. This groundwork provides a foundation for improving cross-language research methodologies.

The fourth chapter illustrated a study investigating cross-lingual mapping of technical terms in privacy policies, focusing on English-to-Italian translation. Using a mix of manual and automated analyses, the research identified differences in term usage between the two languages, taking as an example the technical terms of Tang et al. [159] and the privacy policy corpora of Ciclosi et al. [34]. The study emphasized the need to adapt technical terms for multilingual research to ensure relevance and accuracy. Despite some limitations, including the small sample size, reliance on Italian policies, and challenges in selecting representative cross-language policies, other alternatives, like Europarl Corpus or GDPR decisions, were less applicable to online privacy contexts. Future research should address these challenges to enhance the understanding of privacy policies across diverse languages and cultures.

The fifth chapter presented a study highlighting small organizations and schools' pervasive challenges in adhering to GDPR due to resource constraints. Perfectly secure data protection procedures were often poorly implemented, leading to privacy risks. Instead, the study advocated for a risk mitigation approach with practical, feasible procedures, prioritizing continuous improvement over strict compliance with every GDPR detail. Staff training emerged as a critical solution; while one-third of staff desired more instruction, the focus should shift from legal jargon to practical data protection principles. This approach can empower staff to identify and address failing procedures, avoiding designs that succeed theoretically but fail in practice.

In future work to analyze the DPO's everyday activities and challenges, it is possible to consider another perspective, the real DPOs' point of view, through semi-structured interviews with them. To complete the vision of what DPOs do, it will consider the vision of those who work in small organizations like schools. To understand what the DPO does, it is fundamental to introduce another empirical analysis, which considers several points of view. The idea is to consider the view other stakeholders (the schools' staff, in that case) involved in data protection activities have of the DPO's role. It will be carried out by conducting semi-structured interviews with several schools' staff other than DPOs (e.g., school headmaster, ATA staff, faculty, and DGSA) on their vision of DPO's role. To this end, semi-structured expert interviews can be conducted with other data protection stakeholders (e.g., managers, employees, and members of supervisory authority). The results can be compared to those described above. As a future work to improve the empirical analysis of educational institutions' data protection organizational structures, it is possible to realize qualitative semi-structured interviews with

their staff focused on stakeholders' practices and staff's points of view about the role of DPO. These interviews will be coded and used as an empirical base for thematic analysis, allowing the researchers to derive theoretical constructs.

Bibliography

[1] School Mental Health Is Not Just for Students: Why Teacher and School Staff Wellness Matters. Report on emotional & behavioral disorders in youth, 17:6, 2017.

[2] GitHub - PrivApp/IT100-Corpus, 2020.

[3] Gdprxiv.org - enforcement database, 2023.

[4] Github - lawfulcomputing/gdprxiv, 2023.

[5] Privacy Affairs. Gdpr fines list: Find all gdpr fines & detailed statistics.

[6] Gonçalo Almeida Teixeira, Miguel Mira da Silva, and Ruben Pereira. The critical success factors of GDPR implementation: a systematic literature review. Digital Policy, Regulation and Governance, 21(4):402–418, 2019.

[7] Ryan Amos, Gunes Acar, Elena Lucherini, Mihir Kshirsagar, Arvind Narayanan, and Jonathan Mayer. Privacy Policies over Time: Curation and Analysis of a Million-Document Dataset. WWW '21: Proceedings of the Web Conference 2021, pages 2165–2176, 2021.

[8] Neil R Anderson and Michael A West. Measuring climate for work group innovation: development and validation of the team climate inventory. Journal of Organizational Behavior, 19:235–258, 5 1998.

[9] Siddhant Arora, Henry Hosseini, Christine Utz, Vinayshekhar Bannihatti Kumar, Tristan Dhellemmes, Abhilasha Ravichander, Peter Story, Jasmine Mangat, Rex Chen, Martin Degeling, Thomas Norton, Thomas Hupperich, Shomir Wilson, and Norman Sadeh. A tale of two regulatory regimes: Creation and analysis of a bilingual privacy policy corpus. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and

Stelios Piperidis, editors, Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 5460–5472, Marseille, France, June 2022. European Language Resources Association.

[10] (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, 2017.

[11] Article 29 Data Protection Working Party. Guidelines on Data Protection Officers ('DPOs') - WP 243 rev. 01, 2017.

[12] David G Balash, Dongkun Kim, Darika Shaibekova, Rahel A Fainchtein, Micah Sherr, and Adam J Aviv. Examining the examiners: Students privacy and security perceptions of online proctoring services. pages 633–652. USENIX Association, 8 2021.

[13] David G Balash, Elena Korkes, Miles Grant, Adam J Aviv, Rahel A Fainchtein, and Micah Sherr. Educators perspectives of using (or not using) online exam proctoring. pages 5091–5108. USENIX Association, 8 2023.

[14] Karen Becker and Michelle Smidt. A risk perspective on human resource management: A review and directions for future research. Human Resource Management Review, 26:149–165, 6 2016.

[15] Rizwan Beg, Qamar Abbas, and Alok Joshi. A method to deal with the type of lexical ambiguity in a software requirement specification document. Proceedings - 1st International Conference on Emerging Trends in Engineering and Technology, ICETET 2008, pages 1212–1215, 2008.

[16] Steven Bellman, Eric J. Johnson, Stephen J. Kobrin, and Gerald L. Lohse. International Differences in Information Privacy Concerns: A Global Survey of Consumers. http://dx.doi.org/10.1080/01972240490507956, 20(5):313–324, nov 2010.

[17] Daniel M Berry, Erik Kamsties, and Michael M Krieger. From contract drafting to software specification: Linguistic sources of ambiguity. Univ. of Waterloo, Tech. Rep, 2003.

[18] Jaspreet Bhatia and Travis D. Breaux. Semantic incompleteness in privacy policy goals. Proceedings - 2018 IEEE 26th International Requirements Engineering Conference, RE 2018, pages 159–169, oct 2018.

[19] Sandeep Bhatt, Pratyusa K. Manadhata, and Loai Zomlot. The operational role of security information and event management systems. IEEE Security and Privacy, 12(5):35–41, 2014.

[20] Ben Burgess, Avi Ginsberg, Edward W Felten, and Shaanan Cohney. Watching the watchers: bias and vulnerability in remote proctoring software. pages 571–588. USENIX Association, 8 2022.

[21] Alison Burrows, David Coyle, and Rachael Gooberman-Hill. Privacy, boundaries and smart homes for health: An ethnographic study. Health & Place, 50:112–118, 3 2018.

[22] Rochelle A Cadogan et al. An imbalance of power: the readability of internet privacy policies. Journal of Business & Economics Research (JBER), 2, 2004.

[23] Nadine Casutt and Nico Ebert. Data protection officers: Figureheads of privacy or merely decoration. In 16th European Conference on Management, Leadership and Governance, page 39. Academic Conferences International limited, 2020.

[24] Nadine Casutt and Nico Ebert. Data protection officers: Figureheads of privacy or merely decoration. In 16th European Conference on Management, Leadership and Governance, page 39. Academic Conferences International limited, 2020.

[25] Grazia Cecere, Fabrice Le Guel, and Nicolas Soulié. Perceived Internet privacy concerns on social networks in Europe. Technological Forecasting & Social Change, 96:277–287, 2015.

[26] (Westin Research Center). Data protection officer requirements by country, 2021.

[27] Jake Chanenson, Brandon Sloane, Navaneeth Rajan, Amy Morril, Jason Chee, Danny Yuxing Huang, and Marshini Chetty. Uncovering privacy and security challenges in k-12 schools. Association for Computing Machinery, 2023.

[28] Souti Chattopadhyay, Nicholas Nelson, Audrey Au, Natalia Morales, Christopher Sanchez, Rahul Pandita, and Anita Sarma. A tale from the trenches: cognitive biases and software development. pages 654–665. Association for Computing Machinery, 2020.

[29] Aristeidis Chatzipoulidis, Theodosios Tsiakis, and Theodoros Kargidis. A readiness assessment tool for GDPR compliance certification. Computer Fraud & Security, 2019(8):14–19, 2019.

[30] Aristeidis Chatzipoulidis, Theodosios Tsiakis, and Theodoros Kargidis. A readiness assessment tool for GDPR compliance certification. Computer Fraud & Security, 2019(8):14–19, 2019.

[31] Mohamed Chebel, Chiraz Latiri, and Eric Gaussier. Bilingual lexicon extraction from comparable corpora based on closed concepts mining. In Pacific-Asia Conference on Knowledge Discovery and Data Mining, pages 586–598. Springer, 2017.

[32] Francesco Ciclosi and Fabio Massacci. The data protection officer, an ubiquitous role nobody really knows. arXiv preprint arXiv:2212.07712, 2022.

[33] Francesco Ciclosi and Fabio Massacci. The data protection officer: A ubiquitous role that no one really knows. IEEE Security & Privacy Privacy, 21:66–77, 2023.

[34] Francesco Ciclosi, Silvia Vidor, and Fabio Massacci. Building cross-language corpora for human understanding of privacy policies. In Antonio Skarmeta, Daniele Canavese, Antonio Lioy, and Sara Matheu, editors, Digital Sovereignty in Cyber Security: New Challenges in Future Vision, pages 113–131, Cham, 2023. Springer Nature Switzerland.

[35] Francesco Ciclosi, Silvia Vidor, and Fabio Massacci. Cross-language corpora of privacy policies, 3 2023.

[36] Francesco Ciclosi, Silvia Vidor, and Fabio Massacci. Technical terms in privacy policies in italian and english dataset, 1 2025.

[37] Camille Cobb, Sruti Bhagavatula, Kalil Anderson Garrett, Alison Hoffman, Varun Rao, and Lujo Bauer. " i would have to evaluate their objections": Privacy tensions between smart home device owners and incidental users. Proc. Priv. Enhancing Technol., 2021:54–75, 2021.

[38] Shaanan Cohney, Ross Teixeira, Anne Kohlbrenner, Arvind Narayanan, Mihir Kshirsagar, Yan Shvartzshnaider, and Madelyn Sanfilippo. Virtual classrooms and real harms: Remote learning at u.s. universities. pages 653–674. USENIX Association, 8 2021.

[39] Jessica Colnago, Lorrie Faith Cranor, and Alessandro Acquisti. Is there a reverse privacy paradox? an exploratory analysis of gaps between privacy perspectives

and privacy-seeking behaviors. Proceedings on Privacy Enhancing Technologies, 2023:455–476, 1 2023.

[40] Thomas D Cook, Donald Thomas Campbell, and William Shadish. Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin Boston, MA, 2002.

[41] Juliet Corbin and Anselm Strauss. Basics of qualitative research: Techniques and procedures for developing grounded theory. Sage publications, 2014.

[42] Juliet Corbin and Anselm Strauss. Basics of Qualitative Research. Techniques and Procedures for Developing Grounded Theory. SAGE Publications, Inc., fourth edition edition, 3 2015.

[43] M da C Freitas and Miguel da Silva. GDPR Compliance in SMEs: There is much to be done. Journal of Information Systems Engineering & Management, 3:30, 2018.

[44] Asmita Dalela, Saverio Giallorenzo, Oksana Kulyk, Jacopo Mauro, and Elda Paja. A mixed-method study on security and privacy practices in danish companies, 2021.

[45] Fabiano Dalpiaz, Elda Paja, and Paolo Giorgini. Security requirements engineering via commitments. Proceedings - 2011 1st Workshop on Socio-Technical Aspects in Security and Trust, STAST 2011, pages 1–8, 2011.

[46] Fabiano Dalpiaz, Elda Paja, and Paolo Giorgini. Security requirements engineering: designing secure socio-technical systems. MIT Press, 2016.

[47] Fabiano Dalpiaz, Ivor van der Schalk, and Garm Lucassen. Pinpointing ambiguity and incompleteness in requirements engineering via information visualization and NLP. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10753 LNCS:119–135, 2018.

[48] Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. We value your privacy... now take some cookies: Measuring the gdpr's impact on web privacy. arXiv preprint arXiv:1808.05096, 2018.

[49] Jose M. Del Alamo, Danny S. Guaman, Boni García, and Ana Diez. A systematic mapping study on automated analysis of privacy policies. Computing, 2022.

[50] Vasiliki Diamantopoulou, Aggeliki Tsohou, and Maria Karyda. From ISO/IEC27001:2013 and ISO/IEC27002:2013 to GDPR compliance controls. Information and Computer Security, 28(4):645–662, 2020.

[51] Nicola Dragoni, Alberto Lluch Lafuente, Fabio Massacci, and Anders Schlichtkrull. Are we preparing students to build security in? A survey of european cybersecurity in higher education programs. IEEE Security and Privacy, 19(1):81–88, 1 2021.

[52] Jide S Edu, Jose M Such, and Guillermo Suarez-Tangil. Smart home personal assistants: A security and privacy review. ACM Comput. Surv., 53, 12 2020.

[53] CMS Legal Services EEIG. Gdpr enforcement tracker - list of gdpr fines.

[54] Tatiana Ermakova, Benjamin Fabian, and Eleonora Babina. Readability of Privacy Policies of Healthcare Websites. In Wirtschaftsinformatik Proceedings 2015 (WI 2015), 2015.

[55] Parliament EU and Council EU. Consolidated text: 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), 2016.

[56] (European Data Protection Supervisor). Position paper on the role of Data Protection Officers in ensuring effective compliance with Regulation (EC) 45/2001, 2005.

[57] Benjamin Fabian, Tatiana Ermakova, and Tino Lents. Large-Scale Readability Analysis of Privacy Policies. WI '17: Proceedings of the International Conference on Web Intelligence, pages 18–25, 2017.

[58] Charles J Fillmore et al. Frame semantics and the nature of language. In Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech, volume 280, pages 20–32. New York, 1976.

[59] NOYB European Center for Digital Rights. Gdprhub.

[60] (International Organization for Standardization). Iso/iec 27001:2022(en) information security, cybersecurity and privacy protection — information security management systems — requirements, 2022.

[61] Ciclosi Francesco, Vidor Silvia, and Massacci Fabio. Building cross-language corpora for human understanding of privacy policies. pages 113–131. Springer Nature Switzerland, 6 2023.

[62] Matthias Gallé, Athena Christofi, and Hady Elsahar. The case for a GDPR-specific annotated dataset of privacy policies. volume 4, page 2, 2019.

[63] Garante per la protezione dei dati personali. La scuola a prova di privacy - vademecum ed. 2023 (doc. web. 9886884), 6 2023.

[64] Dennis A Gioia, Kevin G Corley, and Aimee L Hamilton. Seeking qualitative rigor in inductive research: Notes on the gioia methodology. Organizational research methods, 16:15–31, 2013.

[65] Paolo Giorgini, Fabio Massacci, John Mylopoulos, and Nicola Zannone. Modeling security requirements through ownership, permission and delegation. In 13th IEEE International Conference on Requirements Engineering (RE'05), pages 167–176. IEEE, 2005.

[66] Moritz Gruber, Christian Höfig, Maximilian Golla, Tobias Urban, and Matteo Große-Kampmann. "we may share the number of diaper changes": A privacy and security analysis of mobile child care applications. Proceedings on Privacy Enhancing Technologies, 2022:394–414, 2022.

[67] Danny S. Guaman, Jose M. Del Alamo, and Julio C. Caiza. GDPR Compliance Assessment for Cross-Border Personal Data Transfers in Android Apps. IEEE Access, 9:15961–15982, 2021.

[68] Danny S Guamán. Gdpr compliance assessment for cross-border personal data transfers in android apps. 2, 2021.

[69] Greg Guest, Kathleen M MacQueen, and Emily E Namey. Applied thematic analysis. sage publications, 2011.

[70] Greg Guest, Kathleen M MacQueen, and Emily E Namey. Applied thematic analysis. sage publications, 2011.

[71] Steffi Haag and Andreas Eckhardt. Shadow it. Business and Information Systems Engineering, 59:469–473, 12 2017.

[72] Irit Hadar, Tomer Hasson, Oshrat Ayalon, Eran Toch, Michael Birnhack, Sofia Sherman, and Arod Balissa. Privacy by designers: software developers' privacy mindset. Empirical Software Engineering, 23(1):259–289, 2018.

[73] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G Shin, and Karl Aberer. Polisis: Automated analysis and presentation of privacy policies using deep learning. pages 531–548, 2018.

[74] Rakibul Hasan. Understanding edtech's privacy and security issues: Understanding the perception and awareness of education technologies' privacy and security issues. Proceedings on Privacy Enhancing Technologies, 2023:269–286, 10 2023.

[75] Rakibul Hasan and Mario Fritz. Understanding utility and privacy of demographic data in education technology by causal analysis and adversarial-censoring. Proceedings on Privacy Enhancing Technologies, 2022:245–262, 4 2022.

[76] Jonas Hielscher, Uta Menges, Simon Parkin, Annette Kluge, and M Angela Sasse. Employees who dont accept the time security takes are not aware enough: The ciso view of human-centred security. pages 2311–2328. USENIX Association, 8 2023.

[77] Martin Hoegl and Hans Georg Gemuenden. Teamwork quality and the success of innovative projects: A theoretical concept and empirical evidence. Organization Science, 12:435–449, 8 2001.

[78] Henry Hosseini, Martin Degeling, Christine Utz, and Thomas Hupperich. Unifying privacy policy detection. Proc. Priv. Enhancing Technol., 2021:480–499, 2021.

[79] Mitra Bokaei Hosseini, Travis D. Breaux, Rocky Slavin, Jianwei Niu, and Xiaoyin Wang. Analyzing privacy policies through syntax-driven semantic analysis of information types. Information and Software Technology, 138, 2021.

[80] Cristina K Hudson and Winny Shen. Understaffing: An under-researched phenomenon. Organizational Psychology Review, 5:244–263, 2015.

[81] Man Hong Hue, Joyanta Debnath, Kin Man Leung, Li Li, Mohsen Minaei, M Hammad Mazhar, Kailiang Xian, Endadul Hoque, Omar Chowdhury, and

Sze Yiu Chau. All your credentials are belong to us: On insecure wpa2-enterprise configurations. pages 1100–1117. Association for Computing Machinery, 2021.

[82] (Il Manifesto). Marche, il buco dello screening: dati accessibili a chiunque, 2021.

[83] Max Van Kleek, Ilaria Liccardi, Reuben Binns, Jun Zhao, Daniel J Weitzner, and Nigel Shadbolt. Better the devil you know: Exposing the data sharing practices of smartphone apps. 2017.

[84] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. pages 79–86, 9 2005.

[85] Ron Kohavi. Mining e-commerce data: The good, the bad, and the ugly. pages 8–13. Association for Computing Machinery, 2001.

[86] Reinhard Köhler. Statistical Comparability: Methodological Caveats. In Building and Using Comparable Corpora, pages 77–91. Springer, 2013.

[87] Faris Bugra Kokulu, Ananta Soneji, Tiffany Bao, Yan Shoshitaishvili, Ziming Zhao, Adam Doupé, and Gail-Joon Ahn. Matched and mismatched socs: A qualitative study on security operations center issues. pages 1955–1970. Association for Computing Machinery, 2019. Studio qualitativo (interviste) con i security worker per i problemi organizzativi dei SOC.

[88] Douwe Korff and Marie Georges. The DPO Handbook Guidance for data protection officers in the public and quasi-public sectors on how to ensure compliance with the European Union General Data Protection Regulation, 2019.

[89] Jess Kropczynski, Zaina Aljallad, Nathan Jeffrey Elrod, Heather Lipford, and Pamela J Wisniewski. Towards building community collective efficacy for managing digital privacy and security within older adult communities. Proc. ACM Hum.-Comput. Interact., 4, 1 2021.

[90] Barbara Krumay and Jennifer Klar. Readability of Privacy Policies. In A. Singhal and J. Vaidya, editors, DBSec 2020. Lecture Notes in Computer Science, pages 388–399. Springer, Cham, 2020.

[91] Priya C. Kumar, Marshini Chetty, Tamara L. Clegg, and Jessica Vitak. Privacy and security considerations for digital technology use in elementary schools. pages 1–13. ACM, 5 2019.

[92] Sophia Kununka, Nikolay Mehandjiev, and Pedro Sampaio. A comparative study of android and iOS mobile applications' data handling practices versus compliance to privacy policy. IFIP Advances in Information and Communication Technology, 526:301–313, 2018.

[93] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. Alexa, are you listening? privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. Proc. ACM Hum.-Comput. Interact., 2, 11 2018.

[94] Roslyn Layton and Silvia Elaluf-Calderwood. A social economic analysis of the impact of GDPR on security and privacy practices. In 2019 12th CMI Conference on Cybersecurity and Privacy (CMI), pages 1–6, 2019.

[95] Jens Leicht and Maritta Heisel. A Survey on Privacy Policy Languages: Expressiveness Concerning Data Protection Regulations. 2019 12th CMI Conference on Cybersecurity and Privacy (CMI), pages 1–6, 2019.

[96] Bo Li and Eric Gaussier. Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 644–652, 2010.

[97] Bo Li and Eric Gaussier. Exploiting comparable corpora for lexicon extraction: Measuring and improving corpus quality. In Building and using comparable corpora, pages 131–149. Springer, 2013.

[98] Kevin Li, Zhaohui Wang, Ye Wang, Bo Luo, and Fengjun Li. Poster: Ethics of computer security and privacy research - trends and standards from a data perspective. pages 3558–3560. Association for Computing Machinery, 2023.

[99] Yao Li, Alfred Kobsa, Bart P Knijnenburg, M-H Carolyn Nguyen, et al. Cross-cultural privacy prediction. Proc. Priv. Enhancing Technol., 2017:113–132, 2017.

[100] Timothy Libert. webxray domain owner list, 2017.

[101] Timothy Libert. An automated approach to auditing disclosure of third-party data collection in website privacy policies. pages 207–216. International World Wide Web Conferences Steering Committee, 2018.

[102] Thomas Linden, Rishabh Khandelwal, Hamza Harkous, and Kassem Fawaz. The privacy policy landscape after the gdpr. Proceedings on Privacy Enhancing Technologies, 2020(1):47–64, 2020.

[103] Tamara Lopez, Helen Sharp, Arosha Bandara, Thein Tun, Mark Levine, and Bashar Nuseibeh. Security responses in software development. ACM Trans. Softw. Eng. Methodol., 32, 4 2023.

[104] Masike Malatji, Annlizé Marnewick, and Suné von Solms. Validation of a socio-technical management process for optimising cybersecurity practices. Computers & Security, 95:101846, 2020.

[105] Angelini Marco, Ciccotelli Claudio, Franchina Luisa, Marchetti Spaccamela Alberto, and Querzoni Leonardo. Framework Nazionale per la Cybersecurity e la Data Protection. CINI - Cybersecruity National Lab, 2 2019.

[106] Tina Marjanov, Maria Konstantinou, Magdalena Jóźwiak, and Dayana Spagnuelo. Data security on the ground: Investigating technical and legal requirements under the gdpr. Proceedings on Privacy Enhancing Technologies, 2023:405–417, 7 2023.

[107] Yod-Samuel Martin and Antonio Kung. Methods and tools for GDPR compliance through privacy and data protection engineering. In 2018 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), pages 108–111, 2018.

[108] Fabio Massacci, Raminder Ruprai, Matthew Collinson, and Julian Williams. Economic impacts of rules-versus risk-based cybersecurity regulations for critical infrastructure providers. IEEE Security & Privacy, 14(3):52–60, 2016.

[109] Joseph A Maxwell et al. Designing a qualitative study, volume 2. The SAGE handbook of applied social research methods, 2008.

[110] Peter Mayer, Collins W Munyendo, Michelle L Mazurek, and Adam J Aviv. Why users (dont) use password managers at a large educational institution. pages 1849–1866. USENIX Association, 8 2022.

[111] Nora McDonald, Rachel Greenstadt, and Andrea Forte. Intersectional thinking about pets: A study of library privacy. Proceedings on Privacy Enhancing Technologies, 2023:480–495, 4 2023.

[112] Jan Mendling, Mark Strembeck, and Jan Recker. Factors of process model comprehension—findings from a series of experiments. Decision Support Systems, 53:195–206, 4 2012.

[113] Abraham Mhaidli, Selin Fidan, An Doan, Gina Herakovic, Mukund Srinath, Lee Matheson, Shomir Wilson, and Florian Schaub. Researchers' experiences in analyzing privacy policies: Challenges and opportunities. Proceedings on Privacy Enhancing Technologies, 4:287–305, 2023.

[114] National People's Congress of the People's Republic of China. Personal Information Protection Law of the People's Republic of China, 2021.

[115] Elda Paja, Fabiano Dalpiaz, and Paolo Giorgini. Modelling and reasoning about security requirements in socio-technical systems. Data & Knowledge Engineering, 98:123–143, 7 2015.

[116] Elda Paja, Fabiano Dalpiaz, Mauro Poggianella, Pierluigi Roberti, and Paolo Giorgini. Modelling security requirements in socio-technical systems with sts-tool. CEUR Workshop Proceedings, 855:155–162, 2012.

[117] Hernan Palombo, Armin Ziaie Tabari, Daniel Lende, Jay Ligatti, and Xinming Ou. An ethnographic understanding of software (in)security and a co-creation model to improve secure software development. pages 205–220. USENIX Association, 8 2020.

[118] Monica Lestari Paramita, David Guthrie, Evangelos Kanoulas, Rob Gaizauskas, Paul Clough, and Mark Sanderson. Methods for collection and evaluation of comparable documents. In Building and using comparable corpora, pages 93–112. Springer, 2013.

[119] (Garante per la protezione dei dati personali). Provvedimento in materia di videosorveglianza - 8 aprile 2010 [1712680], 4 2010.

[120] (Garante per la protezione dei dati personali). Deliberazione del 1° febbraio 2018 - attività ispettiva di iniziativa curata dall'ufficio del garante, anche per mezzo della guardia di finanza, limitatamente al periodo gennaio-giugno 2018 [7810451], 2 2018.

[121] (Garante per la protezione dei dati personali). Deliberazione del 26 luglio 2018 - attività ispettiva di iniziativa curata dall´ufficio del garante, anche per mezzo della guardia di finanza, limitatamente al periodo luglio-dicembre 2018 [9025338], 7 2018.

[122] (Garante per la protezione dei dati personali). Provvedimento che individua le prescrizioni contenute nelle autorizzazioni generali nn. 1/2016, 3/2016, 6/2016, 8/2016 e 9/2016 che risultano compatibili con il regolamento e con il d.lgs. n. 101/2018 di adeguamento del codice - 13 dicembre 2018 [9068972], 12 2018.

[123] (Garante per la protezione dei dati personali). Deliberazione del 12 settembre 2019 - attività ispettiva di iniziativa curata dall'ufficio del garante, anche per mezzo della guardia di finanza, limitatamente al periodo luglio-dicembre 2019 [9147297], 9 2019.

[124] (Garante per la protezione dei dati personali). Deliberazione del 14 febbraio 2019 - attività ispettiva di iniziativa curata dall´ufficio del garante, anche per mezzo della guardia di finanza, limitatamente al periodo gennaio-giugno 2019 [9096661], 2 2019.

[125] (Garante per la protezione dei dati personali). Deliberazione del 4 aprile 2019 - regolamento n. 1/2019 concernente le procedure interne aventi rilevanza esterna, finalizzate allo svolgimento dei compiti e all'esercizio dei poteri demandati al garante per la protezione dei dati personali [9107633], 4 2019.

[126] (Garante per la protezione dei dati personali). Deliberazione del 10 dicembre 2020 - attività ispettiva di iniziativa curata dall'ufficio del garante, anche per mezzo della guardia di finanza, limitatamente al periodo gennaio-giugno 2021 [9521843], 12 2020.

[127] (Garante per la protezione dei dati personali). Deliberazione del 1° ottobre 2020 - attività ispettiva di iniziativa curata dall'ufficio del garante, anche per mezzo della guardia di finanza, limitatamente al periodo luglio-dicembre 2020 [9468750], 10 2020.

[128] (Garante per la protezione dei dati personali). Deliberazione del 6 febbraio 2020. attività ispettiva di iniziativa curata dall'ufficio del garante, anche per mezzo della guardia di finanza, limitatamente al periodo gennaio-giugno 2020 [9269607], 2 2020.

[129] Garante per la protezione dei dati personali. Provvedimento correttivo e sanzionatorio nei confronti di TIM S.p.A. - 15 gennaio 2020 [9256486], 2020.

[130] (Garante per la protezione dei dati personali). Deliberazione del 22 dicembre 2021 - attività ispettiva di iniziativa curata dall'ufficio del garante, anche per mezzo

della guardia di finanza, limitatamente al periodo gennaio-giugno 2022 [9737049], 12 2021.

[131] (Garante per la protezione dei dati personali). Deliberazione del 22 luglio 2021 - attività ispettiva di iniziativa curata dall'ufficio del garante, anche per mezzo della guardia di finanza, limitatamente al periodo luglio-dicembre 2021 [9689657], 7 2021.

[132] (Garante per la protezione dei dati personali). Ordinanza ingiunzione nei confronti di aicomply s.r.l. - 10 giugno 2021 [9685947], 6 2021.

[133] (Garante per la protezione dei dati personali). Deliberazione del 21 luglio 2022 - attività ispettiva di iniziativa curata dall'ufficio del garante, anche per mezzo della guardia di finanza, limitatamente al periodo luglio-dicembre 2022 - 21 luglio 2022 [9809072], 7 2022.

[134] (Garante per la protezione dei dati personali). Deliberazione del 26 gennaio 2023 - attività ispettiva di iniziativa curata dall'ufficio del garante, anche per mezzo della guardia di finanza, limitatamente al periodo gennaio-giugno 2023 [9862660], 1 2023.

[135] (Garante per la protezione dei dati personali). Deliberazione del 3 agosto 2023 - attività ispettiva di iniziativa curata dall'ufficio del garante, anche per mezzo della guardia di finanza, limitatamente al periodo luglio-dicembre 2023 [9920683], 8 2023.

[136] (Garante per la protezione dei dati personali). Provvedimento del 29 dicembre 2023 - attività ispettiva di iniziativa curata dall'ufficio del garante, anche per mezzo della guardia di finanza, limitatamente al periodo gennaio-luglio 2024 [9981356], 12 2023.

[137] (Garante per la protezione dei dati personali). Provvedimento del 20 giugno 2024 [10028498], 6 2024.

[138] (Garante per la protezione dei dati personali). Provvedimento del 4 luglio 2024 - attività ispettiva di iniziativa curata dall'ufficio del garante, anche per mezzo della guardia di finanza, limitatamente al periodo luglio/dicembre 2024 [10056323], 7 2024.

[139] (Garante per la protezione dei dati personali). Provvedimento del 9 maggio 2024 [10027595] - garante privacy, 5 2024.

[140] Jovan Powar and Alastair R Beresford AlastairBeresford. Sok: Managing risks of linkage attacks on data privacy. Proceedings on Privacy Enhancing Technologies, 2023:97–116, 4 2023.

[141] Sören Preibusch. Guide to measuring privacy concern: Review of survey and observational instruments. International Journal of Human-Computer Studies, 71:1133–1143, 12 2013.

[142] Joel R. Reidenberg, Travis Breaux, Lorrie Faith Cranor, Brian French, Amanda Grannis, James T. Graves, Fei Lie, Aleecia McDonald, Thomas B. Norton, Rohan Ramanath, N. Cameron Russell, Norman Sadeh, and Florian Schaub. Disagreeable privacy policies: Mismatches between meaning and users' understanding. Berkeley Technology Law Journal, 30(1):39–68, 2015.

[143] Joel R Reidenberg, Travis Breaux, Lorrie Faith Cranor, Brian French, Amanda Grannis, James T Graves, Fei Liu, Aleecia McDonald, Thomas B Norton, and Rohan Ramanath. Disagreeable privacy policies: Mismatches between meaning and users' understanding. Berkeley Technology Law Journal, 30:1–88, 2015.

[144] Julie M. Robillard, Tanya L. Feng, Arlo B. Sporn, Jen-Ai Lai, Cody Lo, Monica Ta, and Roland Nadler. Availability, readability, and content of privacy policies and terms of agreements of mental health apps. Internet Interventions, 17, 2019.

[145] Marco Robol, Mattia Salnitri, and Paolo Giorgini. Toward GDPR-compliant socio-technical systems: modeling language and reasoning framework. In IFIP Working Conference on The Practice of Enterprise Modeling, pages 236–250. Springer, 2017.

[146] Per Runeson and Martin Höst. Guidelines for conducting and reporting case study research in software engineering. Empirical Software Engineering, 14:131–164, 2009.

[147] Paul Ryan, Martin Crane, and Rob Brennan. Design challenges for GDPR RegTech. In Proceedings of the 22nd International Conference on Enterprise Information Systems. SCITEPRESS - Science and Technology Publications, 2020.

[148] Nyyti Saarimäki. Methodological issues in observational studies. SIGSOFT Softw. Eng. Notes, 44:24, 10 2020.

[149] David Sarne, Jonathan Chler, Alon Singer, Ayelet Sela, and Ittai Bar Siman Tov. Unsupervised Topic Extraction from Privacy Policies. WWW '19: Companion Proceedings of The 2019 World Wide Web Conference, pages 563–568, 2019.

[150] Kim Schildkamp, Cindy L Poortman, Johanna Ebbeler, and Jules M Pieters. How school leaders can build effective data teams: Five building blocks for a new wave of data-informed decision making. Journal of Educational Change, 20:283–325, 2019.

[151] Sean Sirur, Jason R C Nurse, and Helena Webb. Are We There Yet? Understanding the Challenges Faced in Complying with the General Data Protection Regulation (GDPR). In Proceedings of the 2nd International Workshop on Multimedia Privacy and Security, pages 88–95. Association for Computing Machinery, 2018.

[152] Inguna Skadiņa, Andrejs Vasiļjevs, Raivis Skadiņš, Robert Gaizauskas, Dan Tufiş, and Tatiana Gornostay. Analysis and evaluation of comparable corpora for under resourced areas of machine translation. In The 5th Workshop on Building and Using Comparable Corpora, page 17. Citeseer, 2012.

[153] Nili Steinfeld. "i agree to the terms and conditions": (how) do users read privacy policies online? an eye-tracking experiment. Computers in Human Behavior, 55:992–1000, 2 2016.

[154] Jennifer L Styron. Critical issues facing school principals. Journal of College Teaching & Learning (TLC), 8:1–10, 2011.

[155] Chen Sun, Evan Jacobs, Daniel Lehmann, Andrew Crouse, and Supreeth Shastri. Gdprxiv: Establishing the state of the art in gdpr enforcement. Proceedings on Privacy Enhancing Technologies, 2023:484–499, 10 2023.

[156] Yuanyi Sun, Sencun Zhu, and Yu Chen. Zoomp3: Privacy-preserving publishing of online video conference recordings. Proceedings on Privacy Enhancing Technologies, 2022:630–649, 7 2022.

[157] Ali Sunyaev, Tobias Dehling, Patrick L Taylor, and Kenneth D Mandl. Availability and quality of mobile health app privacy policies. Journal of the American Medical Informatics Association, 22:e28–e33, 4 2015.

[158] Tuomas Talvensaari, Jorma Laurikkala, Kalervo Järvelin, Martti Juhola, and Heikki Keskustalo. Creating and exploiting a comparable corpus in cross-language information retrieval. ACM Transactions on Information Systems (TOIS), 25(1):4–es, 2007.

[159] Jenny Tang, Hannah Shoemaker, Ada Lerner, and Eleanor Birrell. Defining privacy: How users interpret technical terms in privacy policies. Proc. Priv. Enhancing Technol., 2021(3):70–94, 2021.

[160] Jenny Tang, Hannah Shoemaker, Ada Lerner, and Eleanor Birrell. Defining Privacy: How Users Interpret Technical Terms in Privacy Policies. Proceedings on Privacy Enhancing Technologies, 3:70–94, 2021.

[161] Edith S Tatet. Teaching in under-resourced schools: The teach for america example. Theory into Practice, 38(1):37–45, 1999.

[162] Welderufael B Tesfay, Peter Hofmann, Toru Nakamura, Shinsaku Kiyomoto, and Jetzabel Serna. I Read but Don't Agree: Privacy Policy Benchmarking Using Machine Learning and the EU GDPR. pages 163–166. International World Wide Web Conferences Steering Committee, 2018.

[163] Welderufael B Tesfay, Peter Hofmann, Toru Nakamura, Shinsaku Kiyomoto, and Jetzabel Serna. Privacyguide: Towards an implementation of the eu gdpr on internet privacy policy evaluation. pages 15–21. Association for Computing Machinery, 2018.

[164] Huahong Tu, Adam Doupé, Ziming Zhao, and Gail-Joon Ahn. Users really do answer telephone scams. pages 1327–1340. USENIX Association, 8 2019.

[165] Joseph Turow, Lauren Feldman, and Kimberly Meltzer. Open to exploitation: America's shoppers online and offline. Departmental Papers (ASC), page 35, 2005.

[166] Joseph Turow, Michael Hennessy, and Nora Draper. Persistent Misperceptions: Americans' Misplaced Confidence in Privacy Policies, 2003–2015. Journal of Broadcasting & Electronic Media, 62(3):461–478, 2018.

[167] Inger Anne Tøndel and Daniela Soares Cruzes. Continuous software security through security prioritisation meetings. Journal of Systems and Software, 194:111477, 12 2022.

[168] Blase Ur and Yang Wang. A cross-cultural framework for protecting user privacy in online social media. pages 755–762. Association for Computing Machinery, 2013.

[169] Matthew W. Vail, Julia B. Earp, and Annie I. Antón. An Empirical Study of Consumer Perceptions and Comprehension of Web Site Privacy Policies. IEEE Transactions on Engineering Management, 55(3):442–454, 2008.

[170] Matthew W. Vail, Julia B. Earp, and Annie I. Antón. An empirical study of consumer perceptions and comprehension of web site privacy policies. IEEE Transactions on Engineering Management, 55(3):442–454, 2008.

[171] Jan-Philip van Acken, Joost F Gadellaa, Slinger Jansen, and Katsiaryna Labunets. Poster: The unknown unknown: Cybersecurity threats of shadow it in higher education. pages 3633–3635. Association for Computing Machinery, 2023.

[172] Lisa Wallander. 25 years of factorial surveys in sociology: A review. Social Science Research, 38(3):505–520, sep 2009.

[173] Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N. Cameron Russell, Thomas B. Norton, Eduard Hovy, Joel Reidenberg, and Norman Sadeh. The Creation and Analysis of a Website Privacy Policy Corpus. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 1:1330–1340, 2016.

[174] Ka Lok Wu, Man Hong Hue, Ngai Man Poon, Kin Man Leung, Wai Yin Po, Kin Ting Wong, Sze Ho Hui, and Sze Yiu Chau. Back to school: On the (in)security of academic vpns. pages 5737–5754. USENIX Association, 8 2023.

[175] Yaxing Yao, Justin Reed Basdeo, Oriana Rosata Mcdonough, and Yang Wang. Privacy perceptions and designs of bystanders in smart homes. Proc. ACM Hum.-Comput. Interact., 3, 11 2019.

[176] R.K. Yin. Case Study Research and Applications: Design and Methods Sixth Edition. SAGE Publications, 9 2018.

[177] Razieh Nokhbeh Zaeem, Safa Anya, Alex Issa, Jake Nimergood, Isabelle Rogers, Vinay Shah, Ayush Srivastava, and K. Suzanne Barber. PrivacyCheck v2: A

Tool that Recaps Privacy Policies for You. CIKM '20: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pages 3441–3444, 2020.

[178] Sherali Zeadally and Stephanie Winkler. Privacy Policy Analysis of Popular Web Platforms. IEEE Technology and Society Magazine, 35(2):75–85, 2016.

[179] Eric Zeng, Shrirang Mare, and Franziska Roesner. End user security and privacy concerns with smart homes. pages 65–80. USENIX Association, 7 2017.

[180] Victoria Zhong, Susan McGregor, and Rachel Greenstadt. "im going to trust this until it burns me" parents privacy concerns and delegation of trust in k-8 educational technology. pages 5073–5090. USENIX Association, 8 2023.

[181] Sebastian Zimmeck and Steven M. Bellovin. Privee: An Architecture for Automatically Analyzing Web Privacy Policies. In 23rd USENIX Security Symposium (USENIX Security 14), pages 1–16, San Diego, CA, August 2014. USENIX Association.

[182] Sebastian Zimmeck, Rafael Goldstein, and David Baraka. PrivacyFlash Pro: Automating Privacy Policy Generation for Mobile Apps. In Network and Distributed Systems Security (NDSS) Symposium, 2021.

[183] Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel R. Reidenberg, N. Cameron Russel, and Norman Sadeh. MAPS: Scaling Privacy Compliance Analysis to a Million Apps. Proceedings on Privacy Enhancing Technologies, 2019(3):66–86, 2019.

[184] Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel R Reidenberg, N Cameron Russell, and Norman Sadeh. Maps: Scaling privacy compliance analysis to a million apps. Proc. Priv. Enhancing Tech., 2019:66–86, 2019.

Appendix A

Description of Supplementary Tables in DPO's article

In distilling our reference scenarios, we followed the methodology suggested by Yin in [176]. According to this vision, multiple (at least six) sources of evidence could exist for a case study building. Because better case studies rely on various sources usually used to build triangulation, we employed multiple sources of evidence (for example, court decisions, supervisory authorities' decisions, job advertisements, and newspaper articles). The distilled scenarios mask the original persons or organization involved in the original case study to make it more general and less vulnerable to retraction based on the right to be forgotten.

Table A.1 shows how each scenario from Table 2.2 covers a DPO function as described in Table 2.1. Table A.2 links reference scenarios described in Table 2.2 with an exemplification of possible impacts (consequences of the vulnerabilities) that are useful in better describing these scenarios.

Tables A.4 and A.5 lists the sources of case studies used to distill the scenario presented in Table 2.2 (in Sections 2.4 and C.1, respectively.). Table A.6 lists the sources of the other case study examples cited in the article that do not belong to the twelve scenarios.

The source of the example cited in Section 2.7 as a case of a court that reverted an attempt to mandate qualifications is the first one listed in Table A.6. For the years of professional experience in data protection that a DPO should have, one can refer to the second row of Table A.6. This document specifies that the head of the investigation expects the DPO to have at least three years of professional experience in data protection. Starting from the third row of the same reference in Table A.6 are

Table A.1: Scenarios' correlations to DPO functions

This table shows how each scenario from Table 2.2 covers a DPO function as described in Table 2.1.

Short name	**Primary Function(s)**	**Ancillary Function(s)**
Wrong-Advice	Advisory function	Organizational function
Ignored-DPO-Advice	Organizational function, Advisory function	(lack of) Enforcement
DPO-Advice-Not-Sought	Monitoring of compliance	Advisory function, Organizational function, Cooperative function, (lack of) Enforcement
Website-Forces-Choices	Monitoring of compliance	Enforcement, Advisory function
Admin-Asks-For-Everything	Handle queries or complaints (or) Monitoring of compliance	Monitoring of compliance (or) Handle queries or complaints, Enforcement, Information and raising awareness function
Neglected-Subject-Right	Cooperative function, Handle queries or complaints	(lack of) Monitoring of compliance, (lack of) Organizational function
Subject-Right-Request	Handle queries or complaints	Monitoring of compliance, Organizational function
No-Data-Protection-Principles	Organizational function	Advisory function
Unchecked-Remote-Monitoring	Organizational function	Advisory function
Wrong-Public-Procurement	Advisory function	Information and raising awareness function
Software-End-of-Life	Advisory function	Organizational function
Subcontractor-Violates-Privacy	Investigation, Monitoring of compliance	Advisory function, Organizational function, Enforcement, Cooperative function

listed the related documents (such as the job descriptions and advertisements for hiring DPOs, and internal regulations describing the DPO's requirements) that we looked for in Section 2.7.

In Table A.7 are listed the sources of evidence used to build the case related to an integrated video surveillance system by using an OCR camera (P1, Table 2.4).

Finally, we analyzed the GDPR's fines case types in the sites listed in [53]. We mapped the number of cases for each of them. Moreover, we related these types to our twelve scenarios (See Table A.9 for this classification).

Table A.2: Scenarios' possibles consequences

This table links reference scenarios described in Table 2.2 with an exemplification of possible impacts (consequences of the vulnerabilities) that are useful in better describing these scenarios.

Short name	**Examples of consequences**
Wrong-Advice	The controller evaluation of the risk of processing on rights and freedom of individuals could be wrong.
Ignored-DPO-Advice	The processing could be risky for the rights and freedoms of natural persons.
DPO-Advice-Not-Sought	Anyone could pretend to be another data subject and access his or her personal data.
Website-Forces-Choices	It is lost the legal basis of the processing.
Admin-Asks-For-Everything	Excess and irrelevant personal data are disseminated over the Internet by the controller, making them accessible to an indefinite number of unauthorized subjects.
Neglected-Subject-Right	The infringement of the data subjects' rights provisions expose the controller to administrative fines up to 20000000 EUR, or up to 4% of the total worldwide annual turnover.
Subject-Right-Request	The controller may not be able to respond to the data subject's request, exposing it to complaints and administrative fines.
No-Data-Protection-Principles	A personal data breach may happen because of unauthorized disclosure of personal data.
Unchecked-Remote-Monitoring	The technicians can connect to a workstation independently without notifying their access request user and acquiring its preventive consent. Moreover, they can connect to a workstation, stealthy monitoring all activities the user logged into the computer is doing.
Wrong-Public-Procurement	A contractor that offers products or services, not GDPR-compliant can win a bid, and the call for tender may be subject to litigation.
Software-End-of-Life	It is impossible to guarantee adequate security and functionality of the equipment.
Subcontractor-Violates-Privacy	The tender may be subject to litigation.

Table A.3: Document' sources of DPO's functions description section

Document source	**Document description**
European Data Protection Supervisor	"Position paper on the role of Data Protection Officers in ensuring effective compliance with Regulation (EC) 45/2001," Brussels, pp. 1–11, 2005.
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," pp. 1–22, 2017.
Il Manifesto (newspaper)	"Marche, il buco dello screening: dati accessibili a chiunque," 2021. [Online]. Available: `https://ilmanifesto.it/marche-il-buco-dello-screening-dati-accessibili-a-chiunque/` [English translation: Marche, the screening hole: data accessible to anyone].
Garante per la protezione dei dati personali	"Ordinanza ingiunzione nei confronti di Azienda sanitaria unica regionale Marche - 13 gennaio 2022 [9744496]," Roma, 2022. [Online]. Available: `https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9744496` [English translation: Injunction order against single regional health company Marche - Jan. 13, 2022].
European Commission	"EU Digital COVID Certificate." [Online]. Available: `https://ec.europa.eu/info/live-work-travel-eu/coronavirus-response/safe-covid-19-vaccines-europeans/eu-digital-covid-certificate_en`
European Parliament and of the Council	"Regulation (EU) 2021/953 of the European Parliament and of the Council of 14 June 2021 on a framework for the issuance, verification and acceptance of interoperable COVID-19 vaccination, test and recovery certificates (EU Digital COVID Certificate) to facilitate free movement during the COVID-19 pandemic," pp. 3–25, jun 2021.
Repubblica Italiana	"DECRETO-LEGGE 22 aprile 2021, n. 52. Misure urgenti per la graduale ripresa delle attivit'a economiche e sociali nel rispetto delle esigenze di contenimento della diffusione dell'epidemia da COVID-19." pp. 3–12, apr 2021. [English translation: DECREE-LAW No. 52 of April 22, 2021. Urgent measures for the gradual resumption of economic and social activities in compliance with the needs to contain the spread of the COVID-19 epidemic.].
Azienda sanitaria unica regionale delle Marche)	"Regolamento organizzativo aziendale privacy dell'Azienda sanitaria unica regionale delle Marche (Determina DG n. 349 del 30/05/2018," Ancona, 2018. [English translation: Corporate organizational privacy regulations of the Single Regional Health Authority of the Marche Region (DG Determination No. 349 of 30/05/2018)].
Banca d'Italia	"Relazione del Responsabile della Protezione dei Dati - anno 2020," Roma, Tech. Rep., 2020. [English translation: Report of the Data Protection Officer].

Table A.4: Document' sources of case studies described in Table 2.2, Section 2.4

Case study	**Document' sources**
Wrong-Advice	Provvedimento del Garante per la protezione dei dati personali n. 118 del 2 luglio 2020 [9440025]. Available: https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9440025 Provvedimento del Garante per la protezione dei dati personali n. 317 del 16 settembre 2021 [9703988]. Available: https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9703988
Ignored-DPO-Advice	Provvedimento del Garante per la protezione dei dati personali n. 207 del 25 maggio 2021 [9590466]. Available: https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9590466 Risposta a un quesito sull'identificazione degli intestatari del Green Pass del Garante per la protezione dei dati personali [9688875]. Available: https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9688875 Parere del Garante per la protezione dei dati personali sul DPCM di attuazione della piattaforma nazionale DGC per l'emissione, il rilascio e la verifica del Green Pass - 9 giugno 2021 [9668064]. Available: https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9668064 Provvedimento del Garante per la protezione dei dati personali n. 170 del 7 aprile 2022 [9773687]. Available: https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9773687 Provvedimento del Garante per la protezione dei dati personali n. 273 del 22 luglio 2021 [9683814]. Available: https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9683814 Provvedimento del Garante per la protezione dei dati personali n. 430 del 13 dicembre 2021 [9727220]. Available: https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9727220 Green Pass: le ultime indicazioni del Garante Privacy, ma rimangono aspetti da chiarire https://www.cybersecurity360.it/legal/privacy-dati-personali/green-pass-le-ultime-indicazioni-del-garante-privacy-ma-rimangono-aspetti-da-chiarire/ Green pass, controlli durante l'orario di servizio o al termine delle lezioni. Lo Snals dice stop: "Manca il rispetto della privacy" https://www.orizzontescuola.it/green-pass-controlli-durante-lorario-di-servizio-o-al-termine-delle-lezioni-lo-snals-dice-stop-manca-il-rispetto-della-privacy/ Segnalazione del Garante per la protezione dei dati personali al Parlamento e al Governo sul Disegno di legge di conversione del decreto-legge n. 127 del 2021 (AS 2394), in relazione alla possibilità di consegna, da parte dei lavoratori dei settori pubblico e privato, di copia della certificazione verde, al datore di lavoro, con la conseguente esenzione, dai controlli, per tutta la durata della validità del certificato [9717878]. Available: https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9717878
DPO-Advice-Not-Sought	Scambio sacche per trapianto midollo a San Martino di Genova [English translation: "Marrow transplant bag exchange at San Martino in Genoa, Italy"]. Available online: https://genova.repubblica.it/cronaca/2018/10/17/news/scambio_sacche_per_trapianto_midollo_a_san_martino_di_genova-209207054/ Vimercate, donna muore in ospedale dopo trasfusione: sacche di sangue scambiate per omonimia [English translation: "Vimercate, woman dies in hospital after transfusion: blood bags mistaken for homonymy"]. Available online:https://www.ilmessaggero.it/italia/monza_ospedale_sacche_sangue_trasfusione_errore_donna_morta-4737967.html?refresh_ce
Website-Forces-Choices	Provvedimento del Garante per la protezione dei dati personali n. 119 del 27 gennaio 2021 [9711630]. Available: https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9711630 Provvedimento del Garante per la protezione dei dati personali n. 317 del 16 settembre 2021 [9703988]. Available: https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9703988 Provvedimento del Garante per la protezione dei dati personali n. 7 del 15 gennaio 2020 [9256486]. Available: https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9256486 Provvedimento del Garante per la protezione dei dati personali n. 195 del 26 maggio 2022 [9788986]. Available: https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9788986
Admin-Asks-For-Everything	Provvedimento del Garante per la protezione dei dati personali n. 173 del 1 ottobre 2020 [9483375]," Roma, 2020. [Online]. Available: https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9483375 Provvedimento del Garante per la protezione dei dati personali n. 272 del 17 dicembre 2020 [9557593]," Roma, 2020. [Online]. Available: https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9557593 Provvedimento del Garante per la protezione dei dati personali n. 204 del 26 maggio 2022 [9780409]. Available: https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9780409 Dati personali pubblicati online: interviene il Garante. Come difenderci dagli abusi. [English translation: Personal data published online: the Guarantor intervenes. How to defend against abuse]. Available: https://www.repubblica.it/economia/diritti-e-consumi/diritti-consumatori/2022/06/19/news/dati_personali_pubblicati_online_interviene_il_garante_come_difenderci_dagli_abusi-354299516/
Neglected-Subject-Right	Provvedimento del Garante per la protezione dei dati personali n. 40 del 26 febbraio 2020 [9365135]," Roma, 2020. [Online]. Available: https://garanteprivacy.it/web/guest/home/docweb/-/docwebdisplay/docweb/9365135 Provvedimento del Garante per la protezione dei dati personali n. 174 del 12 maggio 2022

Table A.5: Document' sources of case studies described in Table 2.2, Section C.1

Case study	**Document' sources**
No-Data-Protection-Principles	Provvedimento del Garante per la protezione dei dati personali n. 285 del 22 luglio 2021 [9685994]. Available: `https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9685994`
	Provvedimento del Garante per la protezione dei dati personali n. 9 del 13 gennaio 2022 [9744496]. Available: `https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9744496`
	(Il Manifesto), "Marche, il buco dello screening: dati accessibili a chiunque", 2021. [English translation: Marche, the screening hole: data accessible to anyone]. Available: `https://ilmanifesto.it/marche-il-buco-dello-screening-dati-accessibili-a-chiunque`
	Garante privacy, multa da un milione ad Atac e Comune di Roma: "Dati degli automobilisti non tutelati". [English translation: Privacy Guarantor, one million fine to Atac and Rome Municipality: 'Motorists' data not protected']. Available: `https://roma.repubblica.it/cronaca/2021/09/10/news/garante_privacy_multa_atac_e_comune-317239933/`
Unchecked-Remote-Monitoring	Provvedimento del Garante per la protezione dei dati personali n. 190 del 13 maggio 2021 [9669974]. Available: `https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9669974`
	Provvedimento del Garante per la protezione dei dati personali n. 137 del 15 aprile 2021 [9670738]. Available: `https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9670738`
	CEDU, Judgment 5 September 2017, CASE OF BĂRBULESCU v. ROMANIA (Application no. 61496/08). Available: `https://hudoc.echr.coe.int/eng#{%22itemid%22:[%22001-177082%22]}`
Wrong-Public-Procurement	Provvedimento del Garante per la protezione dei dati personali n. 16 del 30 gennaio 2020 [9283857]. Available: `https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9283857`
Software-End-of-Life	Provvedimento del Garante per la protezione dei dati personali n. 83 del 4 aprile 2019 [9101974]. Available: `https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9101974`
	Provvedimento del Garante per la protezione dei dati personali n. 548 del 21 dicembre 2017 [7400401]. Available: `https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/7400401`
	Software dispositivi medici è obsoleto. [English translation: Medical device software is obsolete]. Available: `https://www.ansa.it/sito/notizie/tecnologia/software_app/2015/06/22/software-dispositivi-medici-e-obsoleto_98a87cdc-accd-4f8d-b500-5b1ef7cce47d.html`
	Cybersecurity,89% strutture sanità Italia ha software datati. [English translation: Cybersecurity,89% healthcare facilities Italy has outdated software]. Available: `https://www.ansa.it/sito/notizie/tecnologia/hitech/2021/12/07/cybersecurity89-strutture-sanita-italia-ha-software-datati_b5eda4c6-73a6-47f4-a007-b8930aba084b.html`
Subcontractor-Violates-Privacy	Provvedimento del Garante per la protezione dei dati personali n. 16 del 30 gennaio 2020 [9283857]. Available: `https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9283857`
	Provvedimento del Garante per la protezione dei dati personali n. 43 del 10 febbraio 2022 [9751498]. Available: `https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9751498`
	Provvedimento del Garante per la protezione dei dati personali n. 44 del 10 febbraio 2022 [9754332]. Available: `https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9754332`

Table A.6: Document' sources of case studies mentioned in the chapter extending the scenarios described in Table 2.2

References	Document' sources
Side Channels	(Youtvrs), "Tolentino, il video choc dell'incidente alla rotonda in viale Buozzi," 2019. [Online]. Available: https://www.youtvrs.it/tolentino-il-video-choc-dellincidente-alla-rotonda-in-viale-buozzi/ [English translation: "Tolentino, shocking video of the accident at the traffic circle on Buozzi Avenue]. Provvedimento del Garante per la protezione dei dati personali n. 163 del 28 aprile 2022 [9777996]. Available: https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9777996
Section 2.5, Table 2.3	Provvedimento del Garante per la protezione dei dati personali n. 419 del 2 dicembre 2021 [9733053]. Available: https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9733053 Provvedimento del Garante per la protezione dei dati personali n. 277 del 22 luglio 2021 [9693442]. Available: https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9693442 Provvedimento del Garante per la protezione dei dati personali n. 83 del 4 aprile 2019 [9101974]. Available: https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9101974 Attacco hacker: furto di dati personali ai clienti EasyCoop. [English translation: Hacker attack: theft of personal data from EasyCoop customers]. Available: https://www.repubblica.it/economia/diritti-e-consumi/diritti-consumatori/2022/06/09/news/attacco_hacker_furto_di_dati_personali_ai_clienti_easycoop-353134656/ Germania, attacco hacker a ospedale ha provocato morte donna. [English translation: Germany, hacker attack on hospital resulted in woman's death]. Available: https://www.ansa.it/sito/notizie/tecnologia/internet_social/2020/09/18/germania-attacco-hacker-a-ospedale-ha-provocato-morte-donna_663eee64-a732-4fb5-84b2-863335ed9b22.html Attacco hacker, dopo la Bicocca allarme per i servizi sanitari. [English translation: Hacker attack, after Bicocca alert for health services]. Available: https://milano.corriere.it/notizie/cronaca/17_maggio_13/attacco-hacker-la-bicocca-allarme-servizi-sanitari-349880d4-37d1-11e7-ad4d-8609abc1aa8e.shtml Attacco hacker Regione Lazio, ripartono solo le prenotazioni dei vaccini. Ecco l'elenco di tutti i servizi non ancora accessibili. [English translation: Lazio Region hacker attack, only vaccine reservations restart. Here's a list of all services not yet accessible]. Available: https://www.ilfattoquotidiano.it/2021/08/06/attacco-hacker-regione-lazio-ripartono-solo-le-prenotazioni-dei-vaccini-ecco-lelenco-di-tutti-i-servizi-non-ancora-accessibili/6284498/ Regione Lazio, a un mese dall'attacco hacker ancora disagi ai sistemi informatici. Fatture, tamponi e green pass: i servizi rallentati. [English translation: Lazio Region, one month after hacker attack still disrupting IT systems. Bills, buffers and green passes: services slowed down]. Available:https://www.ilfattoquotidiano.it/2021/09/03/regione-lazio-a-un-mese-dallattacco-hacker-ancora-disagi-ai-sistemi-informatici-fatture-tamponi-e-green-pass-i-servizi-rallentati/6308625/ Il Mise ha nascosto un furto di dati dai suoi sistemi per mesi. [English translation: Mise hid data theft from its systems for months]. Available: https://www.wired.it/internet/web/2021/03/16/furto-dati-mise-hackerato-ministero/ Uber, rubati i dati di 57 milioni di utenti e autisti. E l'azienda ha insabbiato tutto. [English translation: Uber, data of 57 million users and drivers stolen. And the company covered it up]. Available: https://www.wired.it/attualita/tech/2017/11/22/cyberattacco-uber-riscatto/ Dati di clienti e autisti non protetti: multe contro Uber per le falle nella privacy. [English translation: Unprotected customer and driver data: fines against Uber for privacy flaws]. Available: https://www.wired.it/internet/regole/2018/11/27/dati-uber-attacco-privacy-multa/ Il Garante della privacy ha multato Uber per 4 milioni di euro. [English translation: Privacy watchdog fined Uber 4 million euros]. Available: https://www.wired.it/article/il-garante-della-privacy-ha-multato-uber-per-4-milioni-di-euro/ Provvedimento del Garante per la protezione dei dati personali n. 101 del 24 marzo 2022 [9771142]. Available: https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9771142 Provvedimento del Garante per la protezione dei dati personali n. 127 del 7 aprile 2022 [9771545]. Available: https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9771545 Provvedimento del Garante per la protezione dei dati personali n. 49 del 10 aprile 2022 [9756869]. Available: https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9756869
Section 2.7	Sentenza del Tribunale Amministrativo del Friuli-Venezia Giulia, la n.287 del 13 settembre 2018, Trieste, 2018. [English translation: "Judgment of the Administrative Court of Friuli-Venezia Giulia, No. 287 of September 13, 2018"]. Available: https://www.giustizia-amministrativa.it/portale/pages/istituzionale/ucm?id=5LLMWH2MBE2JVPC536FUMJHNYU&q Decision of the National Commission sitting in restricted formation on the outcome of investigation no. [...] carried out at public establishment A Deliberation no. 38FR/2021 of October 15, 2021 [original text in French: Décision de la Commission nationale siégeant en formation restreinte sur l'issue de l'enquête n° [...] menée auprès de l'établissement public A Délibération n° 38FR/2021 du 15 octobre 2021]. Available: https://cnpd.public.lu/content/dam/cnpd/fr/decisions-fr/2021/Decision-38FR-2021-sous-forme-anonymisee.pdf Regolamento protezione dati personali della Provincia di Fermo https://www.provincia.fermo.it/public/2019/12/13/1_regolamento-per-il-trattamento-dei-dati-personali.pdf Regolamento comunale per l'attuazione del Regolamento UE 2016/679 relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali del Comune di Tolentino

Table A.7: Document' sources of case study related to an integrated video-surveillance system, referred in Table 2.4

Source type	**Document' sources**
Online newspaper	Patti per videosorveglianza, la firma di 21 Comuni. [English translation: Covenants for video surveillance, the signing of 21 municipalities.]. Available: `https://www.cronachemaceratesi.it/2021/12/09/patti-per-videosorveglianza-la-firma-di-21-comuni/1591504/`
Online newspaper	Videosorveglianza, Camerano aderisce al protocollo di intesa di Macerata. [English translation: Video surveillance, Camerano joins Macerata's memorandum of understanding]. Available: `https://www.cronacheancona.it/2019/12/06/videosorveglianza-camerano-aderisce-al-protocollo-di-intesa-di-macerata/204561/`
Online newspaper	Sistemi di videosorveglianza in rete: i Comuni riuniti a Macerata per rafforzare la sicurezza. [English translation: Networked video surveillance systems: municipalities gather in Macerata to strengthen security]. Available: https://www.cronacheancona.it/2019/06/04/sistemi-di-videosorveglianza-in-rete-i-comuni-riuniti-a-macerata-per-rafforzare-la-sicurezza/170249/
Online newspaper	Telecamere e lettura delle targhe, al vaglio l'integrazione dei sistemi di videosorveglianza. [English translation: Cameras and license plate reading under consideration for integration of video surveillance systems]. Available: https://www.cronacheancona.it/2019/05/30/telecamere-e-lettura-delle-targhe-al-vaglio-lintegrazione-dei-sistemi-di-videosorveglianza/169125/
Online newspaper	Verso un sistema integrato di videosorveglianza urbana: tavola rotonda a Macerata per mettere a fuoco obiettivi e strategie comuni. [English translation: Toward an integrated urban video surveillance system: roundtable in Macerata to focus on common goals and strategies]. Available: https://www.tmnotizie.com/verso-un-sistema-integrato-di-videosorveglianza-urbana-tavola-rotonda-a-macerata-per-mettere-a-fuoco-obiettivi-e-strategie-comuni/
Online newspaper	Videosorveglianza, Grottammare dice sì al Comune di Macerata per l'uso condiviso dei dati. [English translation: Video surveillance, Grottammare says yes to Macerata municipality for shared use of data]. Available: `http://www.ilgraffio.online/2020/02/06/videosorveglianza-grottammare-dice-si-al-comune-macerata-luso-condiviso-dei-dati/`
Online newspaper	Telecamere in rete per tanti Comuni di Ancona, Macerata e Fermo. Obiettivo sicurezza. [English translation: Networked cameras for many municipalities in Ancona, Macerata and Fermo. Security goal]. Available: `https://www.centropagina.it/osimo/telecamere-in-rete-per-tanti-comuni-di-ancona-macerata-e-fermo-obiettivo-sicurezza/`
Online newspaper	La videosorveglianza va ko: i topi masticano la fibra ottica. [English translation: Video surveillance goes down: rats chew through fiber optics]. Available: `https://www.cronachemaceratesi.it/2021/11/06/la-videosorveglianza-va-ko-i-topi-masticano-la-fibra-ottica/1580995/`
Memorandum of understanding	Protocollo di intesa per la gestione del sistema di videosorveglianza del Comune di Macerata. [English translation: Memorandum of understanding for the management of the video surveillance system of the Municipality of Macerata]. Available: http://www.prefettura.it/macerata/download.php and http://www.prefettura.it/macerata/download.php
Memorandum of understanding	Protocollo di intesa per la gestione del sistema integrato di videosorveglianza e di lettura targhe e dei comuni di Rovigo ed Occhiobello coordinato con i profili di interesse operativo delle forze di polizia territoriali. [English translation: Memorandum of understanding for the management of the integrated video surveillance and license plate reading system and the municipalities of Rovigo and Occhiobello coordinated with the operational interest profiles of the territorial police forces]. Available: `https://www.interno.gov.it/sites/default/files/2021-02/protocollo_videosorveglianza_rovigo_e_occhiobello_stralcio_funzionale_polesine_sicuro.pdf`
Online newspaper	Videosorveglianza: Monteprandone aderisce ad un protocollo sovracomunale. Sarà realizzato un sistema centralizzato integrato in collaborazione con diversi Comuni, capofila Macerata. [English translation: Video surveillance: Monteprandone joins a supra-municipal protocol. A centralized integrated system will be implemented in collaboration with several municipalities, led by Macerata]. Available: https://ascoli.cityrumors.it/prima-pagina/videosorveglianza-monteprandone-aderisce-ad-un-protocollo-sovracomunale-sara-realizzato-un-sistema-centralizzato-integrato-in-collaborazione-con-diversi-comuni-capofila-macerata.html
Online newspaper	Incontro a Macerata sull'istituzione di una rete di videosorveglianza sul territorio. [English translation: Meeting in Macerata on the establishment of a video surveillance network in the area]. Available: `https://www.maceratanotizie.it/40864/incontro-a-macerata-sullistituzione-di-una-rete-di-videosorveglianza-sul-territorio`
Online newspaper	Macerata, nuovi sistemi di videosorveglianza: incontro tra i Comuni del territorio. [English translation: Macerata, new video surveillance systems: meeting between area municipalities]. Available: `https://picchionews.it/attualita/macerata-nuovi-sistemi-di-videosorveglianza-incontro-tra-i-comuni-del-territorio`
Memorandum of understanding	Approvazione del "protocollo di intesa per la realizzazione di un sistema integrato di videosorveglianza (mediante l'utilizzo di telecamere OCR)". [English translation: Approval of the "memorandum of understanding for the implementation of an integrated video surveillance system (using OCR cameras)"]. Available: `http://www.comune.montaltodellemarche.ap.it/zf/index.php/atti-amministrativi/delibere/dettaglio/atto/G5WpRMETUWT0-A/provvedimenti/1` and `http://www.comune.montaltodellemarche.ap.it/c044032/de/attachment.php?serialDocumento=009NVR0202201`

Table A.8: Document' sources of DPO's appointment

Type	**Document description (for URLs it refers to Section 2.7 of Table A.6)**
City regulation	"Regolamento comunale per l'attuazione del Regolamento (UE) 2016/679 relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali", Tolentino, 2020. [English translation: Municipal Regulations for the Implementation of Regulation (EU) 2016/679 on the Protection of Individuals with regard to the Processing of Personal Data].
County regulation	"Regolamento protezione dati personali della Provincia di Fermo," Fermo, 2019. [English translation: Personal data protection regulations of the Province of Fermo].
University regulation	Decreto Rettorale n. 137 del 17 aprile 2019, dell'Università degli Studi di Macerata [English translation:Rectoral Decree No. 137, April 17, 2019]
Ministry regulation	"Direttiva Ministro dello sviluppo economico in data 28 gennaio 2020 di individuazione dei soggetti attraverso i quali il Ministero dello sviluppo economico esercita le funzioni di titolare del trattamento ai sensi del Regolamento UE 2016/679", Roma, 2020. [English translation: Directive Minister of Economic Development dated January 28, 2020 identifying the entities through which the Ministry of Economic Development exercises the functions of controller under EU Regulation 2016/679].
Health Care Authority regulation	"Regolamento organizzativo aziendale privacy dell'Azienda sanitaria unica regionale delle Marche (Determina DG n. 349 del 30/05/2018)", Ancona, 2018. [English translation: Corporate organizational privacy regulations of the Single Regional Health Authority of the Marche Region (DG Determination No. 349 of 30/05/2018)].
Court Decision	"Sentenza del Tribunale Amministrativo del Friuli-Venezia Giulia, la n.287 del 13 settembre 2018", Trieste, 2018. [English translation: Judgment of the Administrative Court of Friuli-Venezia Giulia, No. 287 of September 13, 2018].
Call for an internal DPO's appointment	"Bando di selezione interna per il conferimento dell'incarico di Responsabile della protezione dei dati (RPD) per l'Ateneo ai sensi dell'art. 37 del Regolamento UE 2016/679 (RGPD)", Camerino, 2018. [English translation: Notice of internal selection for the appointment of Data Protection Officer (DPO) for the University in accordance with Article 37 of the EU Regulation 2016/679 (RGPD)].
Call for an external DPO's appointment	"Procedura negoziata in modalità telematica finalizzata alla esternalizzazione del servizio di D.P.O. (data protection officer) occorrente alla Azienda sanitaria unica delle Marche n. CIG 7409654FE6", Ancona, 2018. [English translation: Negotiated procedure in telematic mode aimed at outsourcing the service of D.P.O. (data protection officer) needed by the Single Health Authority of the Marche Region].

Table A.9: Mapping Claims to Enforcement Tracker

GDPR's fines case types	**Occurences**	**Reference scenarios**
Insufficient fulfilment of information obligations	124	Website-Forces-Choices
Insufficient legal basis for data processing	489	Website-Forces-Choices, Wrong-Advice, Admin-Asks-For-Everything, Wrong-Public-Procurement
Insufficient fulfillment of data subjects rights	136	Neglected-Subject-Right, Subject-Right-Request
Insufficient technical and organizational measures to ensure information security	274	Wrong-Advice, DPO-Advice-Not-Sought, Ignored-DPO-Advice, Admin-Asks-For-Everything, No-Data-Protection-Principles, Unchecked-Remote-Monitoring, Wrong-Public-Procurement, Software-End-of-Life
Non-compliance with general data processing principles	332	Ignored-DPO-Advice, Website-Forces-Choices, Wrong-Advice, DPO-Advice-Not-Sought, Admin-Asks-For-Everything, Unchecked-Remote-Monitoring, Wrong-Public-Procurement, Subcontractor-Violates-Privacy
Insufficient data processing agreement	9	Website-Forces-Choices, Wrong-Public-Procurement
Insufficient fulfillment of data breach notification obligations	25	No-Data-Protection-Principles
Insufficient cooperation with supervisory authority	56	Neglected-Subject-Right, Subject-Right-Request
Insufficient involvement of data protection officer	14	DPO-Advice-Not-Sought
Unknown	16	*These are particular cases that, according to GDPR Enforcement Tracker's site, are not generalizable.*

Appendix B

Examples of DPO's functions

It is helpful to analyze from a practical point of view some scenarios in which the DPO carries out his or her tasks, to understand the seven functions of the DPO identified by EDPS.

Typical examples of organizational function are high-risk processing. They will typically require the execution of a data protection impact assessment (DPIA) [10], which is a procedure that allows the controller to assess and demonstrate compliance with personal data protection laws. For example, in a local healthcare management organization, a DPO will help a controller to do a DPIA before starting new processing relating to a Covid-19 screening data acquisition and management system (Tab. 2.4, P2). Suppose the controller does not carry out this DPIA. In that case, a data breach may occur by exploiting an existing system vulnerability [82]; consequently, the competent supervisory authority (SA) sanctions the controller. In another example, the DPO helps the controller assess risks posed by new processing built on an integrated territorial video surveillance system (implemented by OCR cameras). In that case, the DPO of the leading proponent municipality summons brings together and chairs a technical table with the mayors, the DPOs, and the technical delegates of all the Municipalities involved (this is a case of joint controllers). This committee defines all the project's technical and organizational aspects (Tab. 2.4, P1). According to the privacy by design principle, this assessment must happen before starting the processing activities. After that, if the analysis results confirm that the processing complies with the data protection laws, the DPO updates the records of processing activities. An example of monitoring activity is personal data breach management. (Tab. 2.4, R2, R7, and R10). An example of monitoring activity is personal data breach management. One more example of this function is the internal complaints handling the investigation.

For example, an employee reports to the DPO that the staff responsible for checking the possession of the EU Digital COVID Certificate for access to the workplace illegally also controls the specific category of the certification to understand which persons are not vaccinated against COVID-19. Upon this notification, the DPO initiates an investigative activity to verify the correctness of the control procedures of the EU Digital COVID Certificate held by the staff (Tab. 2.4, I3).

Another function is supporting the controller to comply with data protection by design principle. In Tab. 2.4, D1 and D2, we described further examples. In another typical example, after a personal data breach occurs, the DPO of a Health Care Authority summons brings together and chairs a technical table with the area managers. Then he or she reports directly to the controller, and if a breach results in a high risk to the rights and freedom of natural persons, communicate the breach to data subjects on behalf of the controller (Tab. 2.4, R10). A similar case is reported in Tab. 2.4, R7.

For the cooperation function, the DPO helps the SA facilitate access to the documents and information both for the performance of the SA's tasks and for the exercise of the latter's investigative, corrective, authorization, and advisory powers (Tab. 2.4, R8). By carrying out these tasks, the DPO **is not** an agent of the SA; he or she acts as an expert that is part of the controller's organization in which he or she works.

Relating to the handle queries or complaints function, the DPO must give a prompt response to data subjects who exercises their rights (Tab. 2.4, R11 and R12).

For the information and awareness-raising function in small and medium-sized enterprises (SME), first, DPOs monitor both the data protection laws changes and the indications of the bodies in charge, and then they prepare short information pills for the staff or create some quick information notes for the team. These informational materials could be made available also by publishing them on a specific internal website (Tab. 2.4, P4).

Despite the DPO having the competence to monitor compliance with the GDPR and handle complaints, his enforcement powers are minimal. For example, in the case of new processing that could present a high risk for the rights and freedoms of natural persons, the controller consults with the DPO on whether or not to carry out a DPIA. If a DPIA is necessary, this consults about the best methodology to follow when carrying out that DPIA and which safeguards to apply to mitigate any related risks. If the controller disagrees with the advice provided by the DPO, it may not take it into account; the controller is only obliged to document in writing why it chose to operate like this. Indeed, according to the accountability principle, the advice of a DPO is not binding for a controller (Tab. 2.4, R1, R3, and R5).

Appendix C

A Map of Risks

The challenges of DPOs are varied, as the organizations they work for. However, it is helpful to classify them into two categories of risks that are typical of a socio-technical system:

- **Technical Risks (TR):** resulting from technical or technological solutions implemented in processing that does not guarantee the protection of the subjects whose personal data is processed.
- **Socio-technical Risks (SR)**: resulting from not optimal execution or designing of organizational procedures, as well as from an incorrect interaction between human actors.

C.1 Technical Risks (TR)

For the readers of this magazine, Technical Risks are likely the most interesting. We can subdivide them into two other groups. The first type relates to design problems, such as choosing a configuration that complies with the data protection by default principle or choosing insecure technical solutions in the call for tenders. The second type involves misconfigurations or errors which appear during the implementation.

The DPO must support and advise the controller on many technical choices. For example, the choice of technical and organizational measures ensures that, by default, only personal data necessary for each specific purpose are processed. The incorrect technical configuration of a telecommunication software switchboard may result in a data breach (Tab. 2.2, No-Data-Protection-Principles) because everyone can view and download the call's metadata and registration. In another example (Unchecked-

Remote-Monitoring), an incorrect processing design may cause choosing tools that do not use encryption or allow unchecked remote desktop connections.

A further example is related to the selection procedures for technical solutions in calls for tenders (Wrong-Public-Procurement) if applicants are not required to "demonstrate" the GDPR compliance of their products or services. As a result, applicants that can demonstrate compliance with the GDPR do not receive a competitive advantage compared to the others. The tender commission may not be able to motivate the choice of the winner of the tender technically. Worse, this omission could make a compliant contractor lose a bid against a not-compliant contractor and kick-start litigation. In the scenario (Subcontractor-Violates-Privacy), a tender winner is violating the contract, for example, because the processor owns the data encryption key and could read all the personal data processed.

In Software-End-of-Life, there is an example of a technical choice for a company that uses obsolete software. In this case, a DPO noticed that many workstations used a version of the Microsoft Windows operating system, obsolete or with the vendor's extended support expired.

C.2 Socio-technical Risks (SR)

Socio-technical risks can appear daily while the DPO supports the controller in arranging processing activities.

We can specialize these risks into four additional categories, such as *auditing* (e.g., Tab. 2.2, Ignored-DPO-Advice and Subcontractor-Violates-Privacy), *communication* (e.g., No-Data-Protection-Principles and Unchecked-Remote-Monitoring), *processing designing* (e.g., Software-End-of-Life) and *relationship* (e.g., Neglected-Subject-Right).

The first is related to activities in which the DPO works on the processing inventory or processing auditing, including keeping a record of the processing activities and the risk analysis or the DPIA execution. The second involves all the issues in which we have an exchange of information between two distinct entities, for example, between the DPO and the staff involved in the processing activities or between the DPO and the controller's management. The third is related to activities of new processing design or an existing modification. The last refers to relationships between the controller and other different bodies (such as a processor, a third party, a data subject, or a SA).

Conflicting requirements are a particular instance of these risks. An example is the conflicting data flow resulting from the execution of procedures meeting conflicting

requirements (e.g., freedom of information acts vs. data protection acts) where personal data of the involved subjects may (accidentally or maliciously) surface.

DPOs often experience *side-channel attacks* from the most unexpected sources. A well-designed processing activity may become unlawful because a staff member operates differently from what is prescribed by the procedure and from the instructions received by the controller.

For example, a DPO discovers that a local online newspaper released a video of a traffic accident, which appears to derive from some closed-circuit video surveillance cameras used by the local police. In that case, the risk lies in something other than the failure of data flow management or the choice of technical solutions. Rather presumably, it lies in the unplanned wickedness of an operator who had chosen to film the screen of the closed-circuit control station with the smartphone and disseminate the video on social channels until it was re-broadcasted by the local newspaper.

In another example, a gym employee registers a new customer who signs up for a quarterly access subscription, asking her to sign her consent to processing personal data for marketing purposes. Unfortunately, violating the organization's procedures could expose the controller to administrative fines or penalties if the employee takes this consensus without first delivering a copy of the information to the data subject.

Appendix D

Methodology short description

D.1 Methodology short description

In this article we used a methodology to derive the scenario from some sources of information publicly available, without basing our consideration only on the subjective perception of one of the authors.

Hence, the insights described in this article are derived using the following listed methodology, and summarized in the Diagram D.1, which illustrates the steps followed to construct the reference scenarios analyzed in the article.

1. Analyzing the existing data protection laws as well as opinions and recommendations of relevant authorities (such as the EDPS and the EDPB).

2. Analyzing the seven DPO's functions that the European Data Protection Supervisor identified in its position paper on the role of DPO in compliance with Regulation (EC) 45/2001.

3. Identifying all the DPO's activities which are involved to the DPO's functions, and correlating them together. In making this position, we take into account how some EU SAs (namely the supervisory authorities of Bulgaria, Croatia, Italy, Poland and Spain, which are involved in the T4DATA EU project) used these functions to group all the DPO's tasks. This grouping is described by D. Korff and M. Georges in "The DPO Handbook Guidance for data protection officers in the public and quasi-public sectors on how to ensure compliance with the European Union General Data Protection Regulation".

4. Using of personal experience to parse the case studies related to the DPO's tasks, by selecting the more illustrative, as described in deep in the next step.

5. Identified some illustrative case studies of the DPO's tasks in which the activities (identified in the previous steps) are carried out.

6. Looking for some publicly available information and used them as reference to illustrate the case studies.

7. Masking the selected publicly available information (but without altering their essence) in order to become the actors who are involved not directly recognizable.

8. Building the article's scenarios. Using the case studies, we carried out analytical work on these to identify possible vulnerabilities and the possible consequences associated.

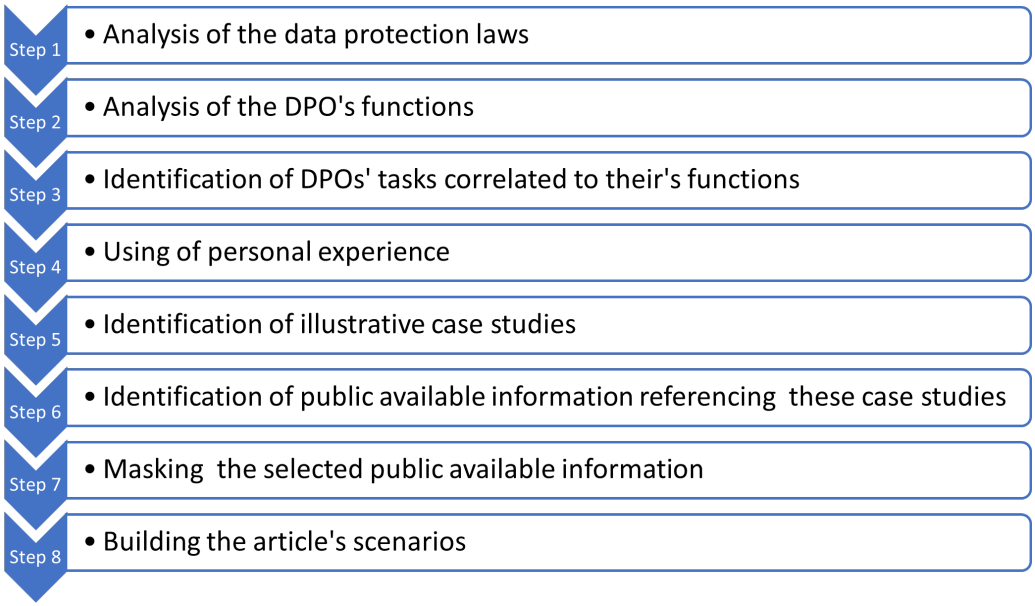

Figure D.1: Methodology's diagram

Appendix E

Scenarios' sources evidence

The case studies analyzed in our work are those of the final scenarios building in step 8 of Appendix D. In building our scenarios we followed the methodology suggested by Yin in [176]. According to this vision, despite one of the most important sources of case study evidence being the interview, this is not the only one. In particular, there could be multiple (at least six) sources of evidence for a case study building. Hence the better case studies rely on a variety of sources that usually are used together to build triangulation. In our work we used multiple sources of evidence that have been processed according to the methodology described in Diagram D.1. Direct observation, documentation and archival records are used in step 4 and 5 using the first author's personal experience to identify some intermediate case studies related to the DPO's tasks that are particularly illustrative of the activities carried out by this role. Instead, documentation and archival records (which are publicly available) are used in step 6 as a reference to illustrate with publicly available evidence the previous intermediate case studies. Prior to building the final scenarios (which are the final case studies analyzed in our work) these evidences are masking according to the step 7 content.

E.1 Where the scenarios come from?

The scenarios come from the personal experience of the first author, but to illustrate them we took into account a series of cases that are in the public domain (for example, judgments, revert in court, supervisory authorities' decisions, newspaper articles). The scenarios were chosen so we can have at least one analyzable practical example for each of the DPO's main tasks, as they are defined by the relevant authorities. We can make the list of the original source material as an additional material, as better detailed

replies to document named *Additional material number 3*.

The authors do not directly cited them for the following two reasons:

- there is a limitation in the number of citations usable in the article (it must be not more than 15), while the number of case studies is wide larger;
- they choose to focus on the scenarios characteristics, avoiding making them fingerprintable, in order to minimize the possibility that in the future the scenarios' actors can exercise their right to be forgotten.

For example, it is possible (because it has happened in the past) that in the future a judgment will force a newspaper to remove some personal data mentioned in a previously published article. Hence, because we masked they, the Magazine is safe from these kinds of problems.

Appendix F

DPO's functions vulnerability

To understand what the DPO role concretely means it is useful taking into account the seven functions identified by the EDPS [56] [88], as listed in Table 2.1.

Table F.1 summarises in a matrix notation how much each of the seven main functions of the DPO is affected by each of the vulnerability categories described in the article. In the Table F.1 the coding used is as follows: "-" if the function is not affected by the vulnerability, otherwise the number of stars ($\star$) to specify the grade with which the DPO function is involved from the vulnerabilities or rather the extent to which the presence of the vulnerability in question could make carrying out each function (in detail, $\star$ for a low value, $\star\star$ for a medium value, and $\star\star\star$ for an high value).

Table F.1: DPO Vulnerabilities

This table describes some different socio-technical vulnerability categories that could affect the DPO's functions

Vuln. category	Vulnerability type	Organisational function	Monitoring of compliance	Advisory function	Cooperative function	Handle queries or complaints	Information and raising awareness	Enforcement
Conflicting Data Flow	Auditing	$\star\star\star$	$\star\star\star$	$\star\star\star$	$\star$	$\star\star$	$\star$	$\star\star\star$
	Communication	$\star\star$	$\star\star\star$	$\star\star\star$	$\star\star$	$\star\star\star$	$\star\star\star$	$\star\star\star$
	Process designing	$\star\star\star$	$\star\star$	$\star\star\star$	$\star\star$	-	$\star\star$	$\star\star\star$
	Relationship	$\star\star$	$\star\star\star$	$\star\star$	$\star\star\star$	$\star\star\star$	$\star\star$	$\star\star$
Technical Issues	Wrong Design	$\star\star\star$	$\star\star$	$\star\star\star$	$\star\star$	-	$\star\star$	$\star\star\star$
	Misconfiguration	$\star\star$	$\star\star\star$	$\star\star$	$\star$	$\star$	$\star$	$\star$

Appendix G

The Data Protection Officer, a ubiquitous role nobody really knows: dataset

The documentary sources of case studies on the issues a data protection officer faces daily dataset is published on Zenodo at the URL: doi:10.5281/zenodo.7879103

The dataset contains the text of the documents that are sources of evidence used in [33] and [32] to distill our reference scenarios according to the methodology suggested by Yin in [176].

The dataset is composed of 95 unique document texts spanning the period 2005-2022. This dataset makes available a corpus of documentary sources useful for outlining case studies related to scenarios in which the DPO finds himself operating in the performance of his daily activities.

The language used in the corpus is mainly Italian, but some documents are in English and French. For the reader's benefit, we provide an English translation of the title of each document.

The documentary sources are of many types (for example, court decisions, supervisory authorities' decisions, job advertisements, and newspaper articles), provided by different bodies (such as supervisor authorities, data controllers, European Union institutions, private companies, courts, public authorities, research organizations, newspapers, and public administrations), and redacted from distinct professional roles (for example, data protection officers, general managers, university rectors, collegiate bodies, judges, and journalists).

The documentary sources were collected from 31 different bodies. Most of the

documents in the corpus (a total of 83 documents) have been transformed into Rich Text Format (RTF), while the other documents (a total of 12) are in PDF format. All the documents have been manually read and verified. The dataset is helpful as a starting point for a case studies analysis on the daily issues a data protection officer face. Details on the methodology can be found in the accompanying papers.

The available files are as follows:

- **documents-texts.zip** –> contain a directory of .rtf files (in some cases .pdf files) with the text of documents used as sources for the case studies. Each file has been renamed with its SHA1 hash so that it can be easily recognized.
- **documents-metadata.csv** –> Contains a CSV file with the metadata for each document used as a source for the case studies.

This dataset is the original one used in the publication [33] and the preprint containing the additional material [32].

Appendix H

Cross-language corpora of privacy policies: dataset

The cross-language corpora of privacy policies dataset is published on Zenodo at the URL: doi:10.5281/zenodo.7729546

The dataset consists of three different privacy policy corpora (in English and Italian) composed of 81 unique privacy policy texts spanning the period 2018-2021. This dataset makes available an example of three corpora of privacy policies. The first corpus is the English-language corpus, the original used in the study by Tang et al. [159]. The other two are cross-language corpora built (one, the source corpus, in English, and the other, the replication corpus, in Italian, which is the language of a potential replication study) from the first corpus.

The policies were collected from:

- the Alexa top 10 Italy and U.S. websites rank;
- the Play Store apps rank in the "most profitable games" category of the Play Store for Italy and the U.S.

We manually analyzed the Alexa top 10 Italy websites as of November 2021. Analogously, we analyzed selected apps that, in the same period, had ranked better in the "most profitable games" category of the Play Store for Italy.

All the privacy policies are ANSI-encoded text files and have been manually read and verified. The dataset is helpful as a starting point for building comparable cross-language privacy policies corpora. The availability of these comparable cross-language privacy policies corpora helps replicate studies in different languages. Details on the methodology can be found in the accompanying paper.

The available files are as follows:

- **policies-texts.zip** –> contains a directory of text files with the policy texts. File names are the SHA1 hashes of the policy text.
- **policy-metadata.csv** –> Contains a CSV file with the metadata for each privacy policy.

This dataset is the original dataset used in the publication [61]. The original English U.S. corpus is described in the publication [159].

Appendix I

My terms are (not always) your terms: technical terms in privacy policies in Italian and English: dataset

The technical terms in privacy policies in Italian and English datasets are published on Zenodo at the URL: doi:10.5281/zenodo.8297075.

Appendix J

Summary of Policies

Table J.1 describes the composition of privacy policies belonging to the corpora of Ciclosi et al. [34]. The table is extracted from Table 1 of the cited article.

Table J.1: Composition of privacy policy corpora from [34]

This table (extracted from Table 1 of the article [34]) describes the whole compositions of the new source and replication corpora, respectively, in English and Italian. The *Referred company* field shows the company (or its product) whose privacy policy is considered. The *Notes* field summarizes the rationale behind choices in determining the privacy policies to retrieve from selected apps and websites. The *original corpus* is the corpus of Tang et al. [159] that Ciclosi et al. [34] rebuilt. The *source corpus* is the new corpus in English built by Ciclosi et al. [34] from the *original corpus*. Meanwhile, the *replication corpus* is the new corpus in Italian of Ciclosi et al. [34].

Company	**Notes**
Amazon, Apple, Ebay, Facebook, Firefox (documentation), Google, King Games (app Candy Crush), LinkedIn, Twitter, Wikipedia, Yahoo	These policies had a privacy policy in both the source language and the replication languages.
Moon Active (app Coin Master), Zinga (app Empires & Puzzles)	These policies replace, in both the source and replication corpora, the policies of *Random Logic Games (app Infinite Word Search)* and *Athena FZE (app Woody Puzzle)* from the original corpus because the policies of the latter games did not exist in the replication language.
Telegram (FAQ doc.), WhatsApp	In both the source and replication corpora, these policies replace Signal's comparable privacy policy from the original corpus because the app's policy did not exist in the replication language.
Unicredit	This policy replaces the policy of *Bank of America* from the original corpus in both the source and replication corpora because the latter company did not exist in the replication language. The company used as a replacement has both an English and Italian policy.
Vodafone	In both the source and replication corpora, this policy replaces the policies of *Reddit* and *Verizon* from the original corpus because the company chosen for replacement is focused on the reality of the replication language's Country.
Zoom, Microsoft	These policies have been added to the source and replication corpora from the top 10 Alexa Italian and U.S. to include updated sites in the Alexa rankings concerning the original corpus.
Poste (only IT)	This policy has been added to the replication corpus because it is of a company from the top 10 Alexa in the replication language. The aim is to align the replication corpus with the original corpus design to have all the top 10 Alexa websites.
USPS (only US),	This policy has been added to the source corpus to have a company with features and products similar to the Italian *Poste*.
Teamsystem (only IT)	This policy has been added to the replication corpus to have a company with features and products similar to the U.S. *Force*.
Force (only US)	This policy has been added to the source corpus because it is of a company from the top 10 Alexa in the source language. The aim was to align the source corpus with the original corpus design to have all the top 10 Alexa websites.

Appendix K

Composition of privacy policy sets

The first set of policies consists of the privacy policies of the company Amazon Inc. We investigated three versions of these policies corpora: the Italian one (included in the replication corpus), the U.S. one, and, finally, the Dutch one (written in English). These latter two versions were included in the source corpus of this study. The Dutch version is particularly relevant for a policies comparison because it is drawn up under the EU GDPR [55] legislation as the Italian one but is written in English as the U.S. one. Therefore, the analysis of the Dutch policy is useful to understanding how different jurisdictions can affect the same language version of a privacy policy.

The second set of policies includes the privacy policies of different banks. They are the Bank of America policy (from the original corpus in English), and the Unicredit group policies belonging to the replication corpus in Italian and the source corpus in English. The banking example was chosen as the English policy targeted non-Italian-speaking people from European Union instead of people from U.S. Therefore, it played the same role as the English policy of the Dutch version of Amazon. These policies should only be minimally affected by legal system changes as they target the same legal system.

Appendix L

Correspondence mapping process

Table L.1 shows examples of correspondence mapping various technical terms in the source language onto their translations in the replication language. In this table, we compared a technical term used in Tang et al. with two terms: the related term most frequently used in replication language privacy policies (belonging to the replication corpus in Ciclosi et al.), and the equivalent legal term introduced by the GDPR.

Table L.1 describes how the terms *personal information* and *personally identifiable information* overlap in the EU legislation.

Regarding the concepts of orphans and widows, an orphan occurs when the term occurs in the replication policy in a paragraph or section with no equivalent in the source policy. In contrast, a widow occurs in the opposite case.

Let us suppose an orphan or a widow are detected somehow; in that case, the corresponding literal translation might be wrong. On the contrary, the two terms denote the same concept if there are neither orphans nor widows. In this case, we can use the same technical term to identify the different terms used in the replication and source language. For example, we mapped the English technical term *personal information* present in the source policies onto the Italian term "*dati personali*" (*personal data*) in the replication policies. Although "*dati personali*" is not the literary translation of the technical term, it denotes the same concept.

Table L.1: Mapping of correspondence onto terms translations

Mapping of correspondences onto the translations for the Tang's technical terms. We obtain these translations by consulting Ciclosi et al., who have the replication corpus of privacy policies available in the replication language.

Original technical term (U.S.)	**Italian technical term**	**Term from GDPR legislation**	**Comments**
personal information	informazioni personali	dati personali	The current EU legislation uses the expression "*dati personali*," but in many privacy policies belonging to the replication corpus, it is used the one "*informazioni personali.*"
personally identifiable information	informazioni personali	dati personali	According to the EU legislation do not exist the distinction (otherwise present in US legislation) between the concepts "*personal information*" and "*personally identifiable information.*"
sensitive personal information	informazioni personali sensibili	categorie particolari di dati personali (dati sensibili)	The current EU legislation uses the expression "*categorie particolari di dati personali*" of which the expression "*dati sensibili*" is a subset. Despite this, many privacy policies belonging to the replication corpus use the one "*informazioni personali sensibili.*"
track/tracking	controllo	tracciamento	According to the current EU legislation, it is better to use the word "*tracciamento*" than the word "*controllo.*" Yet, many privacy policies belonging to the replication corpus use the latter term. Unfortunately, the words *controllo* and *tracciamento* are not synonyms in Italian.

Appendix M

Coding process

Two algorithms were adopted to assign codes to mark the accuracy and quality with which a technical term appears in each privacy policy. Table M.1 reports the codes and their meaning.

Two groups of codes were conceived to measure the policy's precision and quality, respectively. Precision is about how well technical terms appear precisely in the policy of the source corpus, whereas quality is about how well the technical terms in the replication corpus match the source corpus.

The codes *exact*, *synonym*, and *locution* were used to measure accuracy in the source language policies analysis. Hence, if a source corpus policy has an entry in which one of the 58 Tang's technical terms occurs, then one of these three codes is alternatively used. We used the code *exact* whenever, in a sentence of a source corpus policy, there is an exact occurrence of a Tang's technical term. If this term is not present as an exact entry but in a related form (like word or expression having the same meaning), we have used the code *synonym*. Moreover, if the term indirectly appears in the sentence as a locution form, that is as a group of words in grammatical relationship to each other but lacking the formal and meaningful completeness of the sentence, we used the code *locution*. For instance, the expression "*information about your Internet service provider*" is a locution that refers to the technical term *device signal*.

In the source corpus policies' coding activity, we also adopted the codes: *ambiguous* and *missing*. The former was employed when an entry with an ambiguous classification is found in a sentence. The latter was used when a term was missing in the US study because it reflects the EU's particular data protection aspects.

Finally, we used the additional code *implied* if, within a sentence, a technical term appears implicitly in a policy belonging to one corpus and explicitly in the other cor-

responding one. For example, in the sentence *"...il loro mancato conferimento (their failure to provide)..."* from the Italian version of Unicredit privacy policy, the term "*loro*" refers to the Tang et al. [159] technical term *personal information*.

We used other codes to estimate quality in the analysis of the replication corpus' privacy policies. They gauge the mapping of correspondence of the technical terms onto sentences of both versions of the same policy that belongs to distinct corpora. Specifically, we used the code *technical term* if a technical term that appears in a sentence of a privacy policy belonging to the replication corpus also occurs in the same position in the corresponding sentence of the same privacy policy of the source corpus. Suppose the technical term contained in a sentence of the source language of the privacy policy appears translated into a phrase or complex expression in the replication one. In that case, we use the code *paraphrase_rep*. Otherwise, we used the code *paraphrase_src*. In replication corpus policies' coding activity, we even used other codes (*to be discussed* and *new term*) to highlight particular conditions. We used the first to mark an entry that could not be determined the nature (a synonym, locution, or exact match?), requiring a discussion with the other coders to arbitrate the uncertain qualitative characteristics. We used the latter to code a term introduced in replication corpus policies to reflect some specific data protection aspects of the EU that Tang et al. did not consider because it is related to the US. We utilized this code in conjunction with *missing* code.

We have used other codes to track additional information related to the translation modalities of a technical term from the source language version of a privacy policy to the replication language one. When translating a technical term using a legal term from the EU legal system, we used the code *legal term*. In the absence of translation of a technical term, we used the code *no translation*.

Before starting coding, we ordered the lists of privacy policies so that the position of each English policy in the source corpus matched the same position as its corresponding Italian policy in the replication corpus. Later, we carried out the coding activities as follows. First, we manually coded the policies in the source corpus using Algorithm 2. After that, we coded the corresponding ones in the replication corpus according to Algorithm 3. We described the manual coding procedures as algorithms because they help readers understand the approach's key steps and, most importantly, help coders replicate a process where precision is essential.

Algorithm 2 analyzes all policies in the source corpus at the sentence level to see which technical terms from a list of terms belong there. If an exact occurrence of the term or its stem or root form is found in a sentence, the *exact* code is applied. Otherwise,

the code *synonym* is applied if the term appears as a synonym, or the codes *locution* and *implied* are applied if the term appears as a locution. Diversely, if the code does not appear explicitly in the sentence but has an implicit reference to concepts expressed, then the code *implied* is applied. In each of the previous cases, the sentence is also coded with a label named as the technical term's name. If none of the cases occur, the algorithm moves on to the following term. Algorithm 2 terminates after scanning all the terms in the list.

Algorithm 3 encodes each privacy policy of the replication corpus at a sentence level, looking for technical terms belonging to the same list used in Algorithm 2 and then comparing them with the corresponding privacy policy of the source corpus. The algorithm considers each technical term in the list and searches for it at the sentence level in the replication corpus policy. Let's the searched term is found in a sentence. In that case, the algorithm encodes the sentence using the term's name as a label (i.e., if the technical term account information is in the sentence, the sentence is marked with the code *account information*). Then, it parses the corresponding source corpus policy to search for a matching sentence. If a match is found, the algorithm considers whether, in the sentence of the replication corpus policy, the term appears exactly either as a steam or as a root form, or as a synonym. If so, the code *technical term* is applied. Otherwise, if the terms appear as a paraphrase, the codes *paraphrase* and *implied* are applied. If the term appears without a translation in the sentence, the code *no translation* is applied. Moreover, if the term appears with the legal, technical term used in GDPR, the code *legal code* is applied instead of the code *technical term*. Conversely, if the term does not explicitly appear in the sentence but has an implicit reference to the concept expressed, the code *implied* and the technical term name tag are applied to the sentence. Otherwise, the technical term is not in the sentence, and the algorithm looks for the following term.

Algorithm 2 It manually codes the privacy policies of the source corpus, looking for a technical term. It must be run before the algorithm 3. The variable *SourceTechnicalTermsList* contains the technical terms from Tang et al. [159].

```
procedure SOURCE POLICIES CODING(SourceTechnicalTermsList, SourcePolicy)
    Read SourcePolicy
    Read SourceTechnicalTermsList
    counter ← 0
    while counter < length(SourceTechnicalTermsList) do
        Read SourceTechnicalTermsList[counter]
        Policy ← SourcePolicy
        Term ← SourceTechnicalTermsList[counter]
        if ∃Term ∈ Policy then
            ∀Sentence ∈ Policy
            Search Term in Sentence
            if Term exactly in Sentence then
                Code Sentence as EXACT
            else if Term in Sentence as a stem or root form then
                Code Sentence as EXACT
            else if Term in Sentence as a synonym then
                Code Sentence as SYNONYM
            else if Term in Sentence as a locution then
                Code Sentence as LOCUTION
                Code Sentence as IMPLIED
            else if Term not in Sentence but implied then          ▷ Implicit reference to concepts expressed
                Code Sentence as IMPLIED
        Code Sentence as Term label
        counter = counter + 1
```

Algorithm 3 It manually codes a replication corpus privacy policy, looking for technical terms and comparing it with the corresponding privacy policy of the source corpus. It must be run after the algorithm 2.

procedure REPLICATION POLICIES CODING($ReplicationTechnicalTermsList, ReplicationPolicy, SourceTechnic$
  Read $ReplicationPolicy$
  Read $ReplicationTechnicalTermsList$
  $counter \leftarrow 0$
  while $counter < length(ReplicationTechnicalTermsList)$ **do**
    Read $ReplicationTechnicalTermsList[counter]$
    $PolicyRep \leftarrow ReplicationPolicy$
    $TermRep \leftarrow ReplicationTechnicalTermsList[counter]$
    if $\exists TermRep \in PolicyRep$ **then**
      $\forall SentenceRep \in PolicyRep$
      Search $TermRep$ in $SentenceRep$
      if $\exists TermRep \in SentenceRep$ **then**
        Code $SentenceRep$ as $TermRep$ label ▷ The label of TermRep and TermSrc is the same
        Read $SourcePolicy$
        Read $SourceTechnicalTermsList[counter]$
        $PolicySrc \leftarrow SourcePolicy$
        $TermSrc \leftarrow SourceTechnicalTermsList[counter]$
        Read $PolicySrc$
        if $\exists SentenceSrc$ corresponding to $SentenceRep$ **then** ▷ Before coding, it looks for a match between sentences
          Search $TermSrc \in SentenceSrc$
          if $\exists TermSrc \in SentenceSrc$ **then**
            if $TermRep$ in $SentenceRep$ as (exactly OR as a stem OR as a root form OR as a synonym) **then**
              Code $SentenceRep$ as $TermRep$ TECHNICAL TERM
            else if $TermRep$ in $SentenceRep$ as a paraphrase **then**
              Code $SentenceRep$ as PARAPHRASE
              Code $SentenceRep$ as IMPLIED
          if in $SentenceRep$ ($TermRep = TermSrc$) **then**
            Code $SentenceRep$ as NO TRANSLATION
          if in $SentenceRep$ ($TermRep$ = Legal technical term used in GDPR) **then**
            Code $SentenceRep$ as LEGAL CODE
      else if $TermRep$ not in $SentenceRep$ but implied **then** ▷ Implicit reference to concepts expressed
        Code $SentenceRep$ as IMPLIED
        Code $SentenceRep$ as $TermRep$ label
    $counter = counter + 1$

Table M.1: Code groups

Codes used to mark the precision and the quality with which a technical term appears in the analyzed privacy policies. All the code groups described below refer to the three Ciclosi's corpora: (i) the *original corpus*, (ii) the *source corpus*, and (iii) the *replication corpus*.

Code groups	**Description of the code groups**	**Type**
Exact	A term is present in a sentence of an English privacy policy of source corpus and is an exact occurrence of a term in the original corpus.	Precision
Synonym	A term is present in the English privacy policy of source corpus that is a synonym of a term in the original corpus.	Precision
Locution	A term comes into view indirectly in a sentence of an English privacy policy of source corpus, and it is a locution of a term in the original corpus.	Precision
Missing	A term is present in a sentence of an English privacy policy of source corpus; however, it is not an occurrence of a term in the original corpus because it reflects particular EU's data protection aspects.	Precision
Ambiguous	A term is present in a sentence of an English privacy policy of source corpus and has an ambiguous classification concerning occurrences of a term in the original corpus.	Precision
Technical term	A term is present in the same position in a sentence of an Italian privacy policy of replication corpus, and a sentence of the corresponding English privacy policy of source corpus, and in parallel is an occurrence of a term in the original corpus.	Quality
Legal term	A term is present in the same position in a sentence of an Italian privacy policy of replication corpus and a sentence of the corresponding English privacy policy of source corpus. Moreover, in parallel, the term is an occurrence of a term in the original corpus that is translated using a legal term from the EU legal system. It is an additional code respecting the *technical term*, which marks the translation with typical EU legal terminology.	Quality
Paraphrase_it	A term that is an occurrence of a term in the original corpus comes into view indirectly, like a phrase or a complex expression in the Italian version of the privacy policy belonging to the replication corpus and, in parallel, is present in the corresponding English privacy policy of the source corpus.	Quality
Paraphrase_en	A term that is an occurrence of a term in the original corpus comes into view indirectly, like a phrase or a complex expression in the English version of the privacy policy belonging to the source corpus and, in parallel, is present in the corresponding Italian privacy policy of the replication corpus.	Quality
New term	A term is not an occurrence of a term in the original corpus but is present in the Italian version of the privacy policy belonging to the replication corpus. This term is interesting because it reflects particular EU's data protection aspects.	Quality
To be discussed	A term is present in the Italian version of the privacy policy belonging to the replication corpus, in which is part of a sentence with uncertain quality characteristics, and in parallel is an occurrence of a term in the original corpus.	Quality
No translation	A term is present in the same position in a sentence of an Italian privacy policy of replication corpus, and a sentence of the corresponding English privacy policy of source corpus, without being translated from English to Italian, and in parallel it is an occurrence of a term in the original corpus.	Quality

Appendix N

Regular Expressions

We used regular expressions to automatically investigate the presence of a technical term. For example, to look for the technical term *device signal* we used the following expression:

```
\b(device[:space:]((signal(s)*))|
(connectivity[:space:]data))
```

The system carries out a GREP search looking for an empty string at a word boundary containing strings *device signal* or *device signals*. Otherwise, after the empty string, it looks for the string *data* preceded by the one *connectivity* and any white-space character. The GREP search feature is called Expert Search with Regular Expressions (regex) in Atlas.ti and is detailed at `https://manuals.atlasti.com/Win/en/manual/SearchAndCode/SearchAndCodeRegularExpressions.html`.

We have tried to use all sufficiently generic synonyms while reducing the possibility of generating false positives. For example, the text "*the transferred business assets*" is a particular synonym of the technical term "*data transfer*" found specifically in the Dutch policy (written in English) of Amazon.

Table N.1: Sample synonyms used in privacy policies source and replication corpora

Examples of technical terms (or their synonyms) found in the source and replication corpora (respectively in English and Italian) of Ciclosi's privacy policies.

Codebook	**English Original Term**	**English Synonyms**	**Italian "Equivalent" term**	**Italian Synonyms**	**GDPR IT legal term**	**GDPR EN legal term**
location-related information	location	location data, location information	posizione	ubicazione	dati relativi all'ubicazione	location data
general location-related information (*)	general location	IP address, geolocation cookies	posizione approssimativa	ubicazione generale, indirizzo IP	dati relativi all'ubicazione	location data
precise location-related information (*)	precise location	location of your device (GPS case), geolocation information	posizione esatta	posizione del dispositivo (GPS case), informazioni di geolocalizzazione, ubicazione esatta	dati relativi all'ubicazione	location data
device signal (**)	**no unique term**	connectivity data, Wi-Fi information, bluetooth signals, device information	segnali del dispositivo	dati di connettività, Informazioni WiFi, segnali Bluetooth, informazioni sul dispositivo		

* Some terms are not present across the original corpus by Ciclosi et al. [34] but exist in the analyzed privacy policies of the source corpus. Legally, a data subject must know the precision of the location-related information. So, we considered this for the analysis.

** Sometimes there is no unique technical term in the corpora of privacy policies from Ciclosi et al. [34]. The reported term have been found in the source and replication policies of *Facebook*, *Amazon US*, and *Amazon NL*, as well in the *Moonactive*'s replication policy.

Appendix O

Refinement process

The software Atlas.ti dynamically generates smart codes as an output of a query by distinguishing them from the simple codes (e.g., the following query isolates a technical term's false positives).

```
"false positive CODE" AND
NOT("technical term manual CODE") AND
"technical term AUTOCODE"
```

To isolate false negatives, we looked for all cases in which a sentence is coded as *false negative* and, contextually, is manually coded with the code of a technical term but not by "auto-coding". Instead, to isolate false positives, we looked for all cases in which a sentence is coded as *false positive* and contextually is auto-coded with the code of a technical term but not manually coded as such.

Sometimes the replication corpus privacy policies use source language technical terms or their translation indifferently. We adopted a method to test if looking for source language technical terms in the replication corpus privacy policies is helpful, too. We executed the queries with the regular expressions arranged for source corpus privacy policies on those in replication one. If these queries provided results, we added those source language terms to regular expressions arranged for the replication corpus privacy policies.

For example, to understand if the terms PII (*personally identifiable information*) and PI (*personal information*) are equivalent, we executed specific queries looking for widows' or orphans' presence (see Table O.1 for an example on the replication language version of the restricted set of privacy policies from the replication corpus.

We aimed to look for widows' or orphans' presence related to these technical terms. If a term appears in a policy of the replication corpus without a match in the corre-

sponding clauses or sections of the source corpus policy there is an orphan. While, if there is an occurrence in a policy of the source corpus in which the term PII appears without the corresponding PI term in the replication corpus policy there is a widow. Table O.1 summarizes some examples of these queries.

Table O.1: Queries used to look for widows or orphans

Example (related to the technical term *Personal information*) of the queries used to look for the presence of a technical term's widows or orphans.

Objective	**Query**
Searching for a technical term's widows	(NOT Personal information AND (Personally identifiable information OR Sensitive personal information)) AND (Technical term OR Legal term)
Searching for a technical term's orphans	(Personal information AND (NOT (Technical term OR Legal term)))

Appendix P

Top lists of technical terms

This section contains the top 26 lists of more frequent terms in privacy policies that belong to source and replication corpora (see Table P.1).

Table P.1: top 10 and top 26 lists of more frequent terms in privacy policies of source and replication corpora

These side-by-side tables describe the top 10 and the total 26 terms of Tang et al. [159] that appear most frequently in English privacy policies of the source corpus (left) and the Italian privacy policies of the replication corpus (right). The table uses boldface to represent the terms in the top 10 list.

Top 10-26 technical terms in the *English* privacy policies.

Position	Term	#Occurr.
1	**Personal information**	1219
2	**Cookies**	686
3	**Privacy policy**	437
4	**Account information**	280
5	**Third parties**	278
6	**IP address**	131
7	**Targeted ads**	130
8	**Encryption**	104
9	**Track/tracking**	95
10	**Keys**	71
11	Aggregate	67
12	Affiliates	64
13	Personally identifiable information	53
13	Unique identifiers	53
15	Correct (amend & revise)	44
16	Payment information	43
17	API/SDK	41
18	Location-related information	40
19	Web beacons	49
20	Usage data	38
21	Infer	29
22	Device attributes	28
22	End-to-end encrypted	28
24	Advertising identifier	22
25	Device identifier	21
26	Deactivate	18

Top 10-26 technical terms in the *Italian* privacy policies.

Rank	Term	#Occurr.
1	**Personal information**	1524
2	**of which PII**	1013
3	**Cookies**	650
4	**Third parties**	621
5	**Privacy policy**	473
6	**Correct (amend & revise)**	172
7	**IP address**	143
8	**Deactivate**	134
9	**Encryption**	87
10	**Account information**	60
11	Cryptographically hashed	54
12	Data transfer	50
13	Affiliates	47
13	Targeted ads	47
15	Unique identifiers	45
16	API/SDK	44
17	Location-related information	42
17	Payment information	42
19	Device operations	41
20	Anonymize	40
21	Sensitive personal information	36
22	Usage data	28
23	Infer	27
24	Profiling cookies	26
25	End-to-end encryption	25
26	Public information	24

Appendix Q

Technical terms choice for a replications study

Table Q.3 lists some of the Tang et al. [159] study's technical terms that are unusable in a hypothetical Italian version of the survey. We also highlight the cases in which some technical terms are frequent for Italian-speaking users (and therefore usable) in the Italian survey, even if they are uncommon. We found that before distributing the Italian version of the survey, it would be necessary to modify and arrange it. Table Q.2, column *New Terms* shows the list of technical terms that it is necessary to substitute with other terms that fit better and belong to the list of Tang et al. study [159] because they are rare (and thus insignificant for the broader task of understanding a privacy policy) for the Italian-speaking users. We extracted the list of terms to replace in the Italian language survey from the list of 58 technical terms in Tang's et al. [159] study that best fit the Italian context, as found by analyzing the results of "auto-coding" on the replication corpus in [34]. In addition, using the new terms in Table Q.2 allows us to refer to technical terms' categories previously excluded from the survey in the study of Tang et al. [159]. Indeed in this study's survey, the used technical terms did not belong to any of the following categories: *third-parties*, *advertising*, and *identifiers-2*.

Moreover, Table Q.1 summarizes the number of occurrences and the frequency position of the 26 technical terms appearing in the survey of Tang et al. [159] in the source, replication, and original corpora of Ciclosi et al. [34].

Table Q.1: Number of occurrences and position Frequency position of Tang's 26 survey technical terms

This table summarizes the number of occurrences and the frequency position of the 26 technical terms appearing in the survey of Tang et al. [159] in the source, replication, and original corpora of Ciclosi et al. [34].

Technical term	**EN Pos.**	**EN Occ.**	**IT Pos.**	**IT Occ.**	**ORIG. Pos.**	**ORIG. Occ.**
Account information	4	280	10	60	5	231
Aggregated	11	67	27	23	11	64
Anonymize	37	12	20	40	35	7
Browser web storage	48	4	49	4	46	3
Cookie	2	686	3	650	3	285
Cryptographic hash	57	0	11	54	29	15
De-identified	54	2	42	9	53	1
Device attributes	22	28	30	22	20	26
Do Not Track	37	12	57	1	29	15
Encryption	8	104	9	87	8	89
End-to-end encryption	22	28	25	25	18	29
Fingerprinting	46	5	59	0	40	5
Keys	10	71	38	13	13	60
Local storage	46	5	43	8	42	4
Metadata	36	13	30	22	31	13
Permanent cookie	29	17	46	7	42	4
Personal information	1	1219	1	1524	1	835
Personally identifiable information (PII)	13	64	2	1013	14	51
Pixel tag	43	6	43	8	35	7
Privacy policy	3	437	5	473	2	388
Private browsing	26	18	34	18	59	0
Public information	37	12	26	24	33	12
Sensitive personal information	31	16	21	36	46	3
Session cookie	29	17	40	10	42	4
Tracking	9	95	28	23	10	76
Web beacons	19	39	37	15	19	27

Table Q.2: Technical terms to substitute to run a replication in Italian of the English survey of [159]

This Table summarizes the technical terms of Tang et al. [159] that might need to be replaced to run the same study in the Italian version of the survey to reflect the corresponding context of Italian policies.

Original Term	**Category**	**New Term**	**Category**
track/tracking	tracking	usage data	meta/logging
private browsing	tracking	third parties	third-parties
personally identifiable information (PII)	information	location-related information	information
fingerprinting	identifiers-1	unique identifiers	identifiers-2
do not track	tracking	targeted ads	advertising
cryptographically hashed	crypto	affiliates	third-parties

Table Q.3: Technical terms whose use could be problematic in an Italian language survey

This table lists some technical terms from Tang et al. [159] whose use in an Italian-language survey version should be carefully considered. For example, because they may be uncommon or because they are meaningful to Italian-speaking users despite being uncommon. If applicable, the column *Position* reports the list name to which the technical terms belong (such as Italian top 10 or top 26). Otherwise, it notes the technical term's frequency's position number in the replication corpus privacy policies of Ciclosi et al. [34]. Because this list is sorted decreasingly by the number of occurrences of the technical term, a frequency's more considerable position number implies that the technical term is uncommon. Column *Decision* summarizes the rationale for using or not using the term in the Italian version of the survey.

Technical term	**Position**	**Decision**	**Considerations**
Browser web storage	Position 49	To use	The technical term is uncommon in the replication corpus' privacy policies. Still, it is significant for Italian-speaking users and can be used in the Italian version of the survey.
Cryptographically hashed	top 26	To exclude	There is an excessive difference between the number of occurrences of this technical term found in the source corpus' privacy policies (in English) and the number found in the replication corpus (in Italian). Hence, there is probably an "auto-coding" problem, and it is better not to use this term in the Italian version of the survey.
Do not track	Position 57	To exclude	Please refer to the considerations related to the *Cryptographically hashed* term.
Fingerprinting	Position 59	To exclude	The technical term is insignificant for the Italian version of the survey because we did not find it in the privacy policies of the replication corpus by searching for it with "auto-coding".
Personally identifiable information (PII)	top 10	To exclude	According to EU legislation, there is a correspondence between this technical term and *personal information*. Hence, in the Italian version of the survey, both can be referred to with the term *dati personali*.
Private browsing	Position 34	To exclude	The technical term is insignificant for the Italian survey version because it is strongly linked to a specific policy. We found all occurrences of the technical term in the source and replication corpora only in a section of the *Firefox* privacy policy dedicated to anonymous browsing.
Public information	Position 26	To use	The technical term is significant for Italian-speaking users and can be used in the Italian version of the survey.
Track/tracking	Position 27	To exclude	Please refer to considerations related to the *Cryptographically hashed* term.

Appendix R

Method for searching orphans and widows

Looking for orphans means searching in the privacy policies of the replication corpus for any occurrences in which the term *personal information* appeared in a privacy policy without a match in the correspondent version of the policy belonging to the source corpus. As an outcome, we found very few occurrences related to the presence of implicit terms or referring to pronouns. Therefore, it is possible to exclude these results because they are insignificant.

Looking for widows means a double search. The first investigation is that we checked if, in the privacy policies of the replication corpus, there were any occurrences in which the term PII (or the term SPI, *sensitive personal information*) appears without the term PI and a match in the corresponding policy of the source corpus. In parallel, a second investigation checked if, in the privacy policies of the source corpus, there were any occurrences in which the term PII (or the term SPI) appeared without the term PI. This double search aims to make up for the lack of markers for the correspondence of technical terms of Tang et al. existing in a privacy policy in the replication corpus to the correspondent policy in the source corpus. The only existing markers are those for reverse correspondence. Namely, markers that reveal if a technical term in a privacy policy (written in English) belonging to the source corpus is also part of the corresponding privacy policy version (in Italian) belonging to the replication corpus.

Appendix S

Technical terms occurrences details

Comparing the composition of the lists of the top 10 most frequently used terms in Italian and English, we discover a good correspondence. 70% of the technical terms in the English top 10 list also belong to the Italian one (with a peak of 80%, which also belongs to the Italian top 26). Analogously, comparing the lists of the 26 most frequently used terms, we found that it is also good. 73,08% of the technical terms in the English top 26 list belong to the Italian one, and they also cover the whole top 10 Italian list. Indeed, the number of occurrences of the five most frequent terms in each corpus ranges from 61.96% (in the original corpus) to 73.29% (in the replication corpus) of the total ones. However, if we consider the top ten most frequent technical terms, there is an increase in these values, ranging from 76.54% (in the original corpus) to 83.49% (in the replication corpus).

Appendix T

Technical terms meaningfulness details

Considering the 26 technical terms appearing in Tang's survey, thirteen (50,00%) appear in the top 26 list of the source corpus, while eleven (42,30%) appear in the top 26 list of the replication corpus. These results align with those obtainable considering the original corpus: twelve terms (46,15%) are present in the top 26 list of most frequent terms. The gap becomes more significant when we consider the placement of these 26 technical terms in the top 10 lists of more frequent terms in the source and replication corpus, where they occupy 26,92 percent (7 terms presence) and 23,08 percent (6 terms presence), respectively.

Appendix U

Questions from the first survey

This subsection contains the complete first survey questions. The reported text is in dual languages, Italian and English. We used Italian text in the field and offered English text to benefit non-Italian-speaking readers. Unless otherwise specified, each question accepts only one correct answer. For each question, the correct responses (or the correct response) are indicated using bold font. For each multiple-response question, we specified in italics the criterion we used to code a response as correct or incorrect. In each survey's questions, we asked the users to respond using their prior knowledge without consulting search engines such as Google.

First survey's Italian version

The Italian version of the survey questions is as follows:

1. Quale di queste opzioni descrive meglio lo scopo principale di un'**informativa sulla privacy**?

    - Spiega come saranno protetti i tuoi dati
    - **È un documento legale che descrive come i dati degli utenti verranno raccolti e utilizzati**
    - Spiega come l'azienda mantiene riservate le informazioni che raccoglie sugli utenti
    - Afferma che l'azienda non condividerà i dati degli utenti con altri siti o società senza autorizzazione
    - Non lo so

2. Quale di queste opzioni descrive meglio lo scopo principale di una **politica sull'uso dei dati**?

   - Spiega come saranno protetti i tuoi dati
   - **È un documento legale che descrive come i dati degli utenti verranno raccolti e utilizzati**
   - Spiega come l'azienda mantiene riservate le informazioni che raccoglie sugli utenti
   - Afferma che l'azienda non condividerà i dati degli utenti con altri siti o società senza autorizzazione
   - Non lo so

3. Immagina che un'azienda affermi che *"Memorizziamo le tue **chiavi** per un massimo di 30 giorni"*. Qual è la definizione più accurata di "**chiavi**" in questa affermazione?

   - Passwords
   - **Informazioni necessarie per leggere/decodificare i messaggi**
   - File archiviati in Internet
   - Informazioni che ti consentono di accedere al tuo account senza dover effettuare il login ogni volta
   - Non lo so

4. Immagina che un'azienda affermi che *"Utilizziamo lo spazio di **archiviazione web del browser** per conservare informazioni su di te"*. Dove possono essere archiviate le informazioni su di te?
   To know the question's compilation notes and evaluation criteria, please refer to its English translation in Section U.

   - **Sul tuo computer o dispositivo**
   - Su computer o server che si trovano all'interno dell'azienda
   - Su computer o server nella tua zona
   - Su qualsiasi computer o server
   - Non lo so

5. Immagina che un'azienda affermi che *"Utilizziamo l'**archiviazione locale** per conservare informazioni su di te"*. Dove possono essere archiviate le informazioni su di te?
   To know the question's compilation notes and evaluation criteria, please refer to its English translation in Section U.

   - **Sul tuo computer o dispositivo**
   - Su computer o server che si trovano all'interno dell'azienda
   - Su computer o server nella tua zona
   - Su qualsiasi computer o server
   - Non lo so

6. Immagina che la politica di utilizzo dei dati di un'azienda affermi che *"Memorizziamo i **metadati** dei tuoi post sui social media"*. Qual è il significato più accurato di "**metadati**" in questa affermazione?

   - Dati combinati da diverse fonti
   - Un modo migliore di utilizzare i dati
   - Un insieme di dati o informazioni
   - Una grande raccolta di dati
   - **Dati su dati o file (ad es. ora, dimensione, posizione)**
   - Non lo so

7. Quale delle seguenti opzioni potrebbe essere considerata un esempio di **metadato**?
   To know the question's compilation notes and evaluation criteria, please refer to its English translation in Section U.

   - **Dimensione di un file**
   - **Ora in cui è stato inviato un messaggio**
   - **A chi è stata inviata un'e-mail o un messaggio**
   - **Luogo in cui è stata scattata una foto**
   - **Chi ha caricato un file condiviso**
   - Non lo so

8. Immagina che un'azienda affermi che *"Usiamo **tag di pixel** e altre tecnologie simili"*. Qual è la definizione più accurata di "**tag di pixel**" in questo contesto?

   - Un modo per identificare un'immagine
   - Un piccolo pezzo di colore in un'immagine
   - **Un modo per monitorare quali siti gli utenti visitano utilizzando una piccola immagine**
   - Un modo per identificare quali persone o oggetti si trovano in un'immagine
   - Non lo so

9. Immagina che un'azienda affermi che *"Utilizziamo **web beacon** e altre tecnologie simili"*. Qual è la definizione più accurata di "**web beacon**" in questo contesto?

   - **Un modo per monitorare quali siti gli utenti visitano utilizzando una piccola immagine**
   - Uno strumento che consente alle aziende di accedere ai dati sul dispositivo dell'utente
   - Uno strumento che trasporta i dati tra i server
   - Identifica i dati che appartengono a un utente specifico
   - Non lo so

10. Quando vengono eliminati i **cookie persistenti**?
    To know the question's compilation notes and evaluation criteria, please refer to its English translation in Section U.

    - **È possibile eliminarli manualmente e vengono eliminati a quel punto**
    - **Hanno una data di scadenza e vengono eliminati in quel momento**
    - Quando si chiude il browser
    - Quando si esce dal sito web
    - Fino a quando non si chiude la scheda o si esce dal sito Web
    - Mai
    - Non lo so

11. Quali delle seguenti sono affermazioni accurate sulla **crittografia end-to-end**?
    To know the question's compilation notes and evaluation criteria, please refer to its English translation in Section U.

    - **Le applicazioni che utilizzano la crittografia end-to-end sono sicure**
    - **Le applicazioni che utilizzano la crittografia (ma non la crittografia end-to-end) sono sicure**
    - Un'azienda può leggere/decodificare i miei messaggi se la sua applicazione utilizza la crittografia end-to-end
    - **Un'azienda può leggere/decodificare i miei messaggi se la sua applicazione utilizza la crittografia (ma non la crittografia end-to-end)**
    - Il governo può ottenere l'accesso ai miei messaggi tramite una richiesta legale se un'applicazione utilizza la crittografia end-to-end
    - **Il governo può ottenere l'accesso ai miei messaggi tramite una richiesta legale se un'applicazione utilizza la crittografia (ma non la crittografia end-to-end)**
    - Nessuna delle precedenti
    - Non lo so

12. I **dati de-identificati** possono mai essere ricondotti a te come individuo?

    - **Si**
    - No
    - Non lo so

13. I **dati resi anonimi** possono mai essere ricondotti a te come individuo?

    - **Si**
    - No
    - Non lo so

14. I **dati aggregati** possono mai essere ricondotti a te come individuo?

    - Si
    - **No**

- Non lo so

15. Immagina che un'azienda affermi che *"Condividiamo le tue **informazioni personali** con i nostri partner commerciali"*. Elenca tre tipi di dati che potrebbero essere condivisi in base a questa politica.
    To know the question's compilation notes and evaluation criteria, please refer to its English translation in Section U.

16. Immagina che un'azienda affermi che, *"Quando qualcuno visita, interagisce con o utilizza i nostri Servizi, possiamo raccogliere o generare dati tecnici su tale visitatore. Tali dati consistono in dati di connettività, **dati** tecnici o **di utilizzo** aggregati, dati non identificativi"*. Qual è il significato più accurato del termine **dati di utilizzo** in questo contesto?

    - Dati combinati della cronologia di navigazione
    - Un'ampia raccolta di dati
    - **Dati derivati dall'uso di un dispositivo o servizio**
    - Un set di dati personali
    - Non lo so

17. Immagina che un'azienda affermi che, *"Mettiamo a tua disposizione servizi, prodotti, applicazioni o skill forniti da **terze parti** per l'utilizzo dei, o attraverso i, nostri Servizi"*. Qual è il significato più accurato del termine **terze parti** in questo contesto?

    - **Enti, diversi dall'utente che crea i dati e dall'azienda che li usa**
    - Istituzioni politiche coinvolte nel trattamento di dati
    - Aziende che lavorano nel settore della tecnologia
    - Persone, istituzioni o aziende che forniscono servizi agli utenti
    - Non lo so

18. Immagina che un'azienda affermi che, *"Alcuni annunci visualizzati nei nostri Prodotti sono forniti dalle nostre Società **affiliate** all'interno del Gruppo. Come parte della condivisione dei dati, i partner raccoglieranno dati su di te e li useranno in conformità con le proprie informative sulla privacy e con i nostri contratti"*. Qual è il significato più accurato del termine **affiliate** in questo contesto?

- Aziende che operano nell'ambito del marketing
- Persone che fanno parte di un'associazione
- Entità che vendono i propri prodotti su una piattaforma online
- **Organizzazioni controllate dal fornitore del servizio**
- Non lo so

19. Immagina che un'azienda affermi, *"Usiamo le **informazioni relative alla posizione** per fornire, personalizzare e migliorare i nostri Prodotti, comprese le inserzioni, per te e le altre persone"*. Qual è il significato più accurato del termine **informazioni relative alla posizione** in questo contesto?

    - Dati relativi alla posizione dell'azienda
    - **Informazioni sui luoghi nei quali l'utente si trova ora, o è stato in passato**
    - Informazioni relative alla residenza dell'utente
    - Dati sulla posizione dei prodotti dell'azienda
    - Non lo so

20. Immagina che un'azienda affermi, *"Utilizziamo fornitori di marketing, pubblicità e servizi di analisi indipendenti per fornire pubblicità e marketing dei Prodotti, compresa la **pubblicità mirata** basata sull'utilizzo da parte degli utenti del nostro sito Web"*. Qual è il significato più accurato del termine **pubblicità mirata** in questo contesto?

    - **Pubblicità personalizzata in base ai dati e alle attività online degli utenti**
    - Pubblicità diretta alla vendita di un prodotto specifico
    - Annunci creati appositamente per una categoria di utenti
    - Annunci che compaiono in un punto specifico di un sito web
    - Non lo so

21. Immagina che un'azienda affermi che, *"Le informazioni che raccogliamo quando usi i nostri servizi includono **identificatori univoci**"*. Qual è il significato più accurato del termine **identificatori univoci** in questo contesto?

- Sistemi di identificazione utilizzabili su di un solo dispositivo
- Informazioni tecniche relative a un dispositivo
- Credenziali di accesso utilizzabili una sola volta
- **Sistemi per identificare in modo univoco un utente o dispositivo**
- Non lo so

22. Immagina che un'azienda affermi che, *"quando l'utente utilizza i nostri prodotti o interagisce con gli stessi, possiamo raccogliere o trattare, per conto dei nostri clienti, le* ***informazioni sull'account****"*. Qual è il significato più accurato del termine **informazioni sull'account** in questo contesto?

    - **Dati relativi alla fatturazione e al piano dell'account**
    - Dati derivati dall'utilizzo di un dispositivo o di un servizio
    - Un insieme di dati combinati delle attività svolte online dall'utente
    - Un insieme di dati tecnici
    - Non lo so

23. Quali dei seguenti potrebbero essere esempi concreti di **attributi del dispositivo?**
    To know the question's compilation notes and evaluation criteria, please refer to its English translation in Section U.

    - **Dati sull'app del tuo dispositivo, sui nomi e sui tipi di file**
    - Il testo dei messaggi inviati o ricevuti tramite il dispositivo
    - **Dati sul livello della batteria del tuo dispositivo, sulla potenza del segnale, sullo spazio di archiviazione disponibile**
    - Immagini e foto memorizzate sul tuo dispositivo
    - Non lo so

24. Quali delle seguenti sono affermazioni accurate sulla **crittografia**?
    To know the question's compilation notes and evaluation criteria, please refer to its English translation in Section U.

    - **Le applicazioni che utilizzano la crittografia sono sicure**

- L'uso della crittografia in un servizio cloud assicura che il fornitore di servizi non possa accedere ai miei dati
- Il termine crittografia è un sinonimo di crittografia end-to-end
- **Un'azienda può leggere/decifrare i miei messaggi se la sua applicazione utilizza la crittografia**
- Il governo non può avere accesso ai miei messaggi tramite una richiesta legale se un'applicazione utilizza la crittografia
- Nessuna delle precedenti
- Non lo so

25. Immagina che un'azienda affermi che, *"Tu, altre persone che usano i nostri servizi e noi possiamo fornire l'accesso o inviare* ***informazioni pubbliche*** *a chiunque all'interno o all'esterno dei nostri Prodotti, tra cui in altri Prodotti offerti dalle nostre aziende, nei risultati della ricerca o attraverso strumenti e API"*. Qual è il significato più accurato del termine **informazioni pubbliche** in questo contesto?

    - **Le informazioni che condividi pubblicamente su di un social network**
    - Le informazioni che possono essere liberamente utilizzate senza bisogno di autorizzazioni e di corrispondere un compenso
    - Le informazioni che riguardano personaggi pubblici
    - Qualsiasi informazione su di me presente in Internet
    - Non lo so

26. Quali dei seguenti potrebbero essere esempi concreti di **informazioni personali sensibili**?
    To know the question's compilation notes and evaluation criteria, please refer to its English translation in Section U.

    - **Dati personali che rivelano l'origine razziale o etnica, le opinioni politiche, le convinzioni religiose o filosofiche**
    - **Dati relativi all'appartenenza a un sindacato**
    - **Dati relativi ai vaccini covid-19 ricevuti**
    - Dati personali che non voglio che siano divulgati

- **Dati relativi alla vita sessuale o all'orientamento sessuale di una persona**
- Non lo so

27. Quali dei seguenti potrebbero essere esempi reali di **categorie particolari di dati personali**?
    To know the question's compilation notes and evaluation criteria, please refer to its English translation in Section U.

    - **Dati personali che rivelano l'origine razziale o etnica, le opinioni politiche, le convinzioni religiose o filosofiche**
    - **Dati relativi all'appartenenza a un sindacato**
    - **Dati relativi ai vaccini covid-19 ricevuti**
    - Dati personali che non voglio che siano divulgati
    - **Dati relativi alla vita sessuale o all'orientamento sessuale di una persona**
    - Non lo so

28. Quando vengono eliminati i **cookie di sessione**?
    To know the question's compilation notes and evaluation criteria, please refer to its English translation in Section U.

    - Devono essere eliminati manualmente
    - Dopo 30 minuti di inattività
    - **Quando chiudi il browser**
    - **Quando esci dal sito web**
    - Mai
    - Non lo so

First survey's English version

The English version of the survey questions is as follows:

1. Which of these best describes the main purpose of a **privacy policy**?

    - It explains how your data will be protected

- **It is a legal document that says how users' data will be collected and used**
- It explains how the company keeps confidential the information it collects on users
- It says that the company will not share users' data with other sites or companies without permission
- I don't know

2. Which of these best describes the main purpose of a **data use policy**?

   - It explains how your data will be protected
   - **It is a legal document that says how users' data will be collected and used**
   - It explains how the company keeps confidential the information it collects on users
   - It says that the company will not share users' data with other sites or companies without permission
   - I don't know

3. Imagine that a company's data use policy states that, *"We store your **keys** for up to 30 days"*. What is the most accurate definition of **keys** in this statement?

   - Passwords
   - **Information necessary to read/decode messages**
   - Files stored on the Internet
   - Information that allows you to access your account without logging in every time
   - I don't know

4. Imagine that a company's data use policy states that, *"We use **browser web storage** to store information about you"*. Where does this mean that information about you can be stored?
   More than one answer is possible for this question. Nevertheless, "I don't know" is a response that excludes all others. In our analysis, a response will be counted as correct if the respondent has selected the right answer (designated below in bold) and none of the others.

- **On your machine or device**
- On machines or servers that are onsite to the company
- On machines or servers in your local area
- On any machine or server
- I don't know

5. Imagine that a company's data use policy states that, *"We use **local storage** to store information about you"*. Where does this mean that information about you can be stored?
   More than one answer is possible for this question. Nevertheless, "I don't know" is a response that excludes all others. In our analysis, a response will be counted as correct if the respondent has selected the right answer (designated below in bold) and none of the others.

   - **On your machine or device**
   - On machines or servers that are onsite to the company
   - On machines or servers in your local area
   - On any machine or server
   - I don't know

6. Imagine that a company's data use policy states that, *"We store **metadata** from your social media posts"*. What is the most accurate meaning of **metadata** in this statement?

   - Combined data from different sources
   - A better way of using data
   - A set of data or information
   - A large collection of data
   - **Data about data or files (e.g., time, size, location)**
   - I don't know

7. Which of the following might be examples of **metadata**?
   More than one answer is possible for this question. Nevertheless, "I don't know" is a response that excludes all others.

- **Size of a file**
- **Time a message was sent**
- **Who an email or message was sent to**
- **Location where a photo was taken**
- **Who uploaded a shared file**
- I don't know

8. Imagine that a company says, *"We use **pixel tags** and other similar technologies."* What is the most accurate definition of **pixel tag** in this context?

   - A way to identify an image
   - A small piece of color in an image
   - **A way to monitor which sites users visit using a small image**
   - A way to identify which people or objects are in an image
   - I don't know

9. Imagine that a company says, *"We use **web beacons** and other similar technologies."* What is the most accurate definition of **web beacons** in this context?

   - **A way to monitor which sites users visit using a small image**
   - A tool that allows companies to access data on a user's device
   - A tool that transports data between servers
   - Identifies data that belongs to a specific user
   - I don't know

10. When are **persistent cookies** deleted?
    More than one answer is possible for this question. Nevertheless, "I don't know" and "Never" are a response that excludes all others. In our analysis, a response will be counted as correct if the respondent has selected all the right answers (designated below in bold) and none others.

    - **You can manually delete them, and they are deleted then**
    - **They have an expiration date, and they are deleted then**
    - When you close your browser

- When you sign out of the website
- Until you close the tab or navigate away from the website
- Never
- I don't know

11. Which of the following are accurate statements about **end-to-end encryption**? *More than one answer is possible for this question. Nevertheless, "I don't know" and "None of the above" are a response that excludes all others. According to Tang et al., [159] approach, we broke this question into two parts in our analysis. The first part studies misconceptions about end-to-end encryption, while the second part explores misconceptions in the point-to-point one. From the perspective of end-to-end encryption analysis, we marked a response as correct if the two following conditions are together correctly. The first one is that the respondent has not marked the answer choices "A company can read/decrypt my messages if their application uses end-to-end encryption" and "The government can get access to my messages through a legal request if an application uses end-to-end encryption." The second one is that the respondent has marked answer choices "A company can read/decrypt my messages if their application uses encryption (but not end-to-end encryption)" and "The government can get access to my messages through a legal request if an application uses encryption (but not end-to-end encryption)." From the perspective of point-to-point encryption analysis, we marked a response as correct if the respondent has marked together the answer choices "Applications that use end-to-end encryption are secure" and "Applications that use encryption (but not end-to-end encryption) are secure."*

    - **Applications that use end-to-end encryption are secure**
    - **Applications that use encryption (but not end-to-end encryption) are secure**
    - A company can read/decrypt my messages if their application uses end-to-end encryption
    - **A company can read/decrypt my messages if their application uses encryption (but not end-to-end encryption)**
    - The government can get access to my messages through a legal request if an application uses end-to-end encryption

- **The government can get access to my messages through a legal request if an application uses encryption (but not end-to-end encryption)**
- None of the above
- I don't know

12. Can **de-identified information** ever be tied to you as an individual?
    - **Yes**
    - No
    - I don't know

13. Can **anonymized information** ever be tied to you as an individual?
    - **Yes**
    - No
    - I don't know

14. Can **aggregated information** ever be tied to you as an individual?
    - Yes
    - **No**
    - I don't know

15. Imagine that a company says, *"We share your **personal information** with our business partners."* Please list three types of data that could be shared under this policy.
    Users were provided with three textboxes to enter their responses.

16. Imagine that a company's data use policy states that, *"When someone visits, interacts with or uses our Services, we may collect or generate technical data about them. Such data consists of connectivity, technical or aggregated **usage data**, non-identifying data"*. What is the most accurate meaning of **usage data** in this statement?
    - Combined browsing history data

- A large collection of data
- **Data derived from the use of a device or a service**
- A set of personal data
- I don't know

17. Imagine that a company's data use policy states that, *"We make available to you services, products, applications, or skills provided by **third parties** for use on or through our services."* What is the most accurate meaning of **third parties** in this statement?

    - **Entities other than the user providing the data and the company using it**
    - Political institutions involved in data treatment
    - Companies that work in the field of technology
    - People, institutions, or companies that provide services for users
    - I don't know

18. Imagine that a company's data use policy states that, *"Information may also be shared with other services and **affiliates**. Our **affiliates** may use the information in a manner consistent with their privacy policies"*. What is the most accurate meaning of **affiliates** in this statement?

    - Companies operating in the field of marketing
    - People that are registered with an association
    - Entities selling their products on an online platform
    - **Organizations controlled by the service provider**
    - I don't know

19. Imagine that a company's data use policy states that, *"We use **location-related information** to provide, personalize and improve our products, including ads, for you and others."* What is the most accurate meaning of **location-related information** in this statement?

    - Data concerning the location of the company
    - **Information about the places the user is or went to in the past**

- Information about the location of the user's residence
- Data about the location of the company's products
- I don't know

20. Imagine that a company's data use policy states that, *"We use third-party marketing, advertising, and analytics providers to provide advertising and marketing for our products, including* ***targeted advertising*** *based on your use of our website."* What is the most accurate meaning of **targeted advertising** in this statement?

    - **Advertising personalized on the basis of users' data and online activity**
    - Advertising targeted to a specific user
    - Commercials created specifically for a category of users
    - Ads that appear in a specific place on the website
    - I don't know

21. Imagine that a company's data use policy states that, *"The information we collect as you use our services includes* ***unique identifiers****."* What is the most accurate meaning of **unique identifiers** in this statement?

    - Identification systems that can be used on only one device
    - Technical information about a device
    - Access credentials that can be used only once
    - **Systems to identify a user or a device univocally**
    - I don't know

22. Imagine that a company's data use policy states that, *"When you use or interact with our products, we may collect, or process on behalf of our customers, the* ***account information****."* What is the most accurate meaning of **account information** in this statement?

    - **Data related to billing and account plan**
    - Data derived from the use of a device or a service
    - A combined data set of the user's online activities
    - A set of technical data

- I don't know

23. Which of the following might be real examples of **device attributes**?
More than one answer is possible for this question. Nevertheless, "I don't know" is a response that excludes all others. In our analysis, a response will be counted as correct if the respondent has selected the right answers (designated below in bold) and none of the others.

- **Data about your device app, file names, and types**
- The text of messages sent or received through the device
- **Data about your device battery level, signal strength, available storage space**
- Images and photos stored on your device
- I don't know

24. Which of the following are accurate statements about **encryption**?
More than one answer is possible for this question. Nevertheless, "I don't know" and "None of the above" are a response that excludes all others. In our analysis, a response will be counted as correct if the respondent has selected all the right answers (designated below in bold) and none others.

- **Applications that use encryption are secure**
- Using encryption in a cloud service ensures that the service provider cannot access my data
- Encryption is a synonym for end-to-end encryption
- **A company can read/decrypt my messages if their application uses encryption**
- The government cannot get access to my messages through a legal request if an application uses encryption
- None of the above
- I don't know

25. Imagine that a company's data use policy states that, *"You, other people using our services, and we can provide access to or send **public information** to anyone on or off our products, including in other our products, in search results, or through tools and APIs."* What is the most accurate meaning of **public information** in this statement?

    - **Information that you share publicly on a social network**
    - Information that can be freely used without permission or compensation
    - Information about public figures
    - Any information about me on the Internet
    - I don't know

26. Which of the following might be real examples of **sensitive personal information**?
    More than one answer is possible for this question. Nevertheless, "I don't know" is a response that excludes all others. It is a repetition of the next question that uses the technical term in the common language. In our analysis, a response will be counted as correct if the respondent has selected all the right answers (designated below in bold) and none others.

    - **Personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs**
    - **Trade-union membership**
    - **Data on covid-19 vaccines received**
    - Personal data that I do not want to be disclosed
    - **Data concerning a person's sex life or sexual orientation**
    - I don't know

27. Which of the following might be real examples of **sensitive personal information**?
    More than one answer is possible for this question. Nevertheless, "I don't know" is a response that excludes all others. It is a repetition of the previous question that uses the legal term of the GDPR. In our analysis, a response will be counted as correct if the respondent has selected all the right answers (designated below in

bold) and none others.

- **Personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs**
- **Trade-union membership**
- **Data on covid-19 vaccines received**
- Personal data that I do not want to be disclosed
- **Data concerning a person's sex life or sexual orientation**
- I don't know

28. When are **session cookies** deleted?
More than one answer is possible for this question. Nevertheless, "I don't know" and "Never" are a response that excludes all others. In our analysis, a response will be counted as correct if the respondent has selected all the right answers (designated below in bold) and none others.

- They must be deleted manually, and they are deleted then
- After 30 minutes of inactivity, and they are deleted then
- **When you close your browser**
- **When you sign out of the website**
- Never
- I don't know

Appendix V

Questions from the second survey

This subsection contains the complete first survey questions. The reported text is in dual languages, Italian and English. We used Italian text in the field and offered English text to benefit non-Italian-speaking readers. We expressed the comfort level values in a likelihood scale organized on five values: Very comfortable (+2), somewhat comfortable (+1), neither comfortable nor uncomfortable (0), somewhat uncomfortable (-1), and very uncomfortable (-2).

Second survey's Italian version

The Italian version of the survey questions is as follows:

1. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Utilizziamo l'**archiviazione web del browser** per conservare informazioni su di te"*?

2. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Memorizziamo le informazioni su di te sul tuo computer o dispositivo"*?

3. Quanto ti sentiresti a tuo agio nell'usare un prodotto o un servizio se l'informativa sul trattamento dei dati personali affermasse: *"Utilizziamo l'**archiviazione locale** per conservare informazioni su di te"*?

4. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Memorizziamo le informazioni su di te sul tuo computer o dispositivo"*?

5. Quanto ti sentiresti a tuo agio nell'usare un prodotto o un servizio se l'informativa sul trattamento dei dati personali affermasse: *"Utilizziamo l'**archiviazione su cloud** per conservare informazioni su di te"*?

6. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Memorizziamo le informazioni su di te sulle nostre macchine e server"*?

7. Quanto ti sentiresti a tuo agio nell'usare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Non condividiamo il contenuto dei tuoi messaggi, ma condividiamo alcuni **metadati** con terze parti"*?

8. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Non condividiamo il contenuto dei tuoi messaggi, ma condividiamo informazioni come la lunghezza dei tuoi messaggi, l'ora in cui i tuoi messaggi sono stati inviati e le persone a cui hai inviato messaggi con terze parti"*?

9. Quanto ti sentiresti a tuo agio nell'usare un prodotto o un servizio se l'informativa sul trattamento dei dati personali affermasse: *"Non offriamo la crittografia end-to-end, ma utilizziamo la **crittografia** per proteggere i tuoi messaggi"*?

10. Quanto ti sentiresti a tuo agio nell'usare un prodotto o un servizio se l'informativa sul trattamento dei dati personali affermasse: *"Utilizziamo una forma di crittografia per proteggere i tuoi messaggi che garantisce che hacker e intercettatori non possano leggere i tuoi messaggi. Decifriamo e condividiamo il contenuto dei tuoi messaggi solo se riceviamo una richiesta legale dal governo"*?

11. Quanto ti sentiresti a tuo agio nell'usare un prodotto o un servizio se l'informativa sul trattamento dei dati personali affermasse: *"Utilizziamo la **crittografia end-to-end** per proteggere i tuoi messaggi"*?

12. Quanto ti sentiresti a tuo agio nell'usare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Utilizziamo una forma di crittografia per proteggere i tuoi messaggi che garantisce che solo tu (e il tuo destinatario) possiate decodificare il contenuto dei tuoi messaggi"*?

13. Quanto ti sentiresti a tuo agio nell'usare un prodotto o un servizio se l'informativa sul trattamento dei dati personali affermasse: *"Utilizziamo i **web beacon** per raccogliere informazioni"*?

14. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Utilizziamo piccole immagini per monitorare quali siti visita"*?

15. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Utilizziamo i* ***pixel tag*** *per raccogliere informazioni"*?

16. Quanto ti sentiresti a tuo agio nell'usare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Monitoriamo i comportamenti degli utenti su diversi siti web utilizzando immagini invisibili e altri elementi del sito web"*?

17. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Utilizziamo i* ***cookie di sessione*** *per memorizzare informazioni su di te"*?

18. Quanto ti sentiresti a tuo agio nell'usare un prodotto o un servizio se l'informativa sul trattamento dei dati personali affermasse: *"Memorizziamo le informazioni su di te in piccoli file che rimangono sul tuo disco rigido fino a quando non ti disconnetti"*?

19. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Utilizziamo i* ***cookie persistenti*** *per memorizzare informazioni su di te"*?

20. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Memorizziamo le informazioni su di te in piccoli file che rimangono sul tuo disco rigido fino a quando non vengono eliminati o scadono"*?

21. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Raccogliamo informazioni sugli* ***attributi del*** *tuo* ***dispositivo*** *per fornire i nostri servizi principali"*?

22. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Raccogliamo informazioni sulla tua app e sui nomi e tipi di file per fornire i nostri servizi principali"*?

23. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Raccogliamo informazioni sulle applicazioni che hai installato sul tuo dispositivo per fornire i nostri servizi principali"*?

24. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Possiamo raccogliere o generare* ***dati di utilizzo*** *aggregati"*?

25. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Possiamo raccogliere o generare dati come indirizzo IP, ID pubblicità univoci, dati non identificativi relativi a un dispositivo, a un sistema operativo, a un browser o alla versione di un'app, al gestore di telefonia mobile, alle impostazioni internazionali e di lingua, alle attività dell'Utente sui nostri Servizi, dati di diagnostica"*?

26. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Le* ***chiavi*** *di decriptazione sono divise in parti e non sono mai tenute insieme ai dati che proteggono"*?

27. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Le stringhe alfanumeriche che implementano l'algoritmo di decodifica dei dati sono divise in parti e non sono mai tenute insieme ai dati che proteggono"*?

Second survey's English version

The English version of the survey questions is as follows:

1. How comfortable would you be using a product or service if the data use policy says, *"We use* ***browser web storage*** *to store information about you"*?

2. How comfortable would you be using a product or service if the data use policy says, *"We store information about you on your machine or device"*?

3. How comfortable would you be using a product or service if the data use policy says, *"We use* ***local storage*** *to store information about you"*?

4. How comfortable would you be using a product or service if the data use policy says, *"We store information about you on your machine or device"*?

5. How comfortable would you be using a product or service if the data use policy says, "*We use **cloud storage** to store information about you*"?

6. How comfortable would you be using a product or service if the data use policy says, "*We store information about you on our machines and servers*"?

7. How comfortable would you be using a product or service if the data use policy says, *"We do not share the content of your messages, but we do share some **metadata** with third parties"*?

8. How comfortable would you be using a product or service if the data use policy says, *"We do not share the content of your messages, but we do share information such as the length of your messages, the times of your messages were sent, and the people you sent the messages to with third parties"*?

9. How comfortable would you be using a product or service if the data use policy says, *"We do not offer end-to-end encryption, but we use **encryption** to secure your messages"*?

10. How comfortable would you be using a product or service if the data use policy says, *"We use a form of encryption to secure your messages that ensures that hackers and eavesdroppers cannot read your messages. We only decrypt and share the contents of your messages if we receive a legal request from the government"*?

11. How comfortable would you be using a product or service if the data use policy says, *"We use **end-to-end encryption** to secure your messages"*?

12. How comfortable would you be using a product or service if the data use policy says, *"We use a form of encryption to secure your messages that guarantees that only you (and whoever you are talking to) can decode the content of your messages"*?

13. How comfortable would you be using a product or service if the data use policy says, *"We use **web beacons** to collect information"*?

14. How comfortable would you be using a product or service if the privacy policy says, *"We use small images to monitor which sites you visit"*?

15. How comfortable would you be using a product or service if the data use policy says, *"We use **pixel tags** to collect information"*?

16. How comfortable would you be using a product or service if the data use policy says, *"We monitor user behaviors across different websites using invisible images and other website elements"*?

17. How comfortable would you be using a product or service if the data use policy says, *"We use **session cookies** to store information about you"*?

18. How comfortable would you be using a product or service if the data use policy says, *"We store information about you in small files that remain on your hard drive until you log out"*?

19. How comfortable would you be using a product or service if the data use policy says, *"We use **persistent cookies** to store information about you"*?

20. How comfortable would you be using a product or service if the data use policy says, *"We store information about you in small files that remain on your hard drive until deleted or expired"*?

21. How comfortable would you be using a product or service if the data use policy says, *"We collect information about your **device attributes** to provide our Core Services"*?

22. How comfortable would you be using a product or service if the data use policy says, *"We collect information about your app and file names and types to provide our Core Services"*?

23. How comfortable would you be using a product or service if the data use policy says, *"We collect information about applications you have installed on your device to provide our Core Services"*?

24. How comfortable would you be using a product or service if the privacy policy says, *"We may collect or generate aggregated **usage data**"*?

25. How comfortable would you be using a product or service if the privacy policy says, *"We may collect or generate data such as your IP address, unique advertising, non-identifying data regarding a device, operating system, browser or App version, mobile carrier, locale and language settings, user activity on our Services, diagnostic data"*?

26. How comfortable would you be using a product or service if the privacy policy says, *"The decryption **keys** are split into parts and are never kept in the same place as the data they protect"*?

27. How comfortable would you be using a product or service if the privacy policy says, *"The alphanumeric strings that implement the data decryption algorithm are split into parts and are never kept in the same place as the data they protect"*?

Appendix W

Questions from the third survey

This subsection contains the complete first survey questions. The reported text is in dual languages, Italian and English. We used Italian text in the field and offered English text to benefit non-Italian-speaking readers. We express the comfort level values in a likelihood scale organized on five values: Very comfortable (+2), somewhat comfortable (+1), neither comfortable nor uncomfortable (0), somewhat uncomfortable (-1), and very uncomfortable (-2).

Third survey's Italian version

The Italian version of the survey questions is as follows:

1. Quanto ti sentiresti a tuo agio nell'usare un prodotto o un servizio se l'informativa sul trattamento dei dati personali affermasse: *"Raccogliamo* ***informazioni*** *su di te* ***dal tuo account*** *per fornire i nostri servizi principali"*?

2. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Raccogliamo informazioni come il tuo nome utente, il codice fiscale, o indirizzo e-mail per fornire i nostri servizi principali"*?

3. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Raccogliamo* ***informazioni personali*** *su di te per fornire i nostri servizi principali"*?

4. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o un servizio se l'informativa sul trattamento dei dati personali affermasse: *"Raccogliamo informazioni come il tuo nome completo, indirizzo ed e-mail per fornire i nostri servizi principali"*?

5. Quanto ti sentiresti a tuo agio nell'usare un prodotto o un servizio se l'informativa sul trattamento dei dati personali affermasse: *"Raccogliamo **informazioni pubbliche** su di te per fornire i nostri servizi principali"*?

6. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Raccogliamo informazioni come il tuo nome utente, immagine del profilo o post sui social media per fornire i nostri servizi principali"*?

7. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Raccogliamo **informazioni personali sensibili** su di te per fornire i nostri servizi principali"*?

8. Quanto ti sentiresti a tuo agio nell'usare un prodotto o servizio sse l'informativa sul trattamento dei dati personali affermasse: *"Raccogliamo informazioni come la tua razza ed etnia, opinioni politiche e religiose o informazioni sulla tua salute per fornire i nostri servizi principali"*?

9. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Come parte della condivisione dei dati, le società **affiliate** raccoglieranno dati su di te e li useranno in conformità con le proprie informative sulla privacy e con i nostri contratti"*?

10. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Come parte della condivisione dei dati con la nostra società, le società che sono controllate dalla nostra società o che la controllano, oppure che sono sotto il controllo comune con una o più altre società, raccoglieranno dati su di te e li useranno in conformità con le proprie informative sulla privacy e con i nostri contratti"*?

11. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Condividiamo solo **informazioni aggregate** con terze parti"*?

12. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Condividiamo con terzi solo elementi informativi raccolti su un certo numero di persone"*?

13. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Condividiamo solo* ***informazioni anonimizzate*** *con terze parti"*?

14. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Condividiamo con terzi solo le informazioni precedentemente identificabili che sono state de-identificate e per le quali non esiste più un codice o un'altra associazione per la re-identificazione"*?

15. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Condividiamo solo* ***dati de-identificati*** *con terze parti"*?

16. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Condividiamo con terzi solo i documenti in cui le informazioni di identificazione personale sono state rimosse o oscurate in modo tale che le informazioni rimanenti non identifichino una persona fisica, e non c'è una ragionevole base per credere che le informazioni possano essere utilizzate per identificare un interessato"*?

17. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Usiamo le* ***informazioni relative alla posizione*** *per fornire, personalizzare e migliorare i nostri Prodotti, comprese le inserzioni, per te e le altre persone"*?

18. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Usiamo alcune informazioni su di te, come la posizione attuale, il luogo di residenza, i luoghi che ami visitare e le aziende e le persone nelle vicinanze, per fornire, personalizzare e migliorare i nostri Prodotti, comprese le inserzioni, per te e le altre persone"*?

19. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio che ha un'**informativa sul trattamento dei dati personali**?

20. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio che ha un documento che descrive quali sono i dati personali raccolti, le finalità del trattamento, i diritti degli interessati, e tutti i servizi di terzi con cui tali dati vengono condivisi?

21. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Utilizziamo fornitori di marketing, pubblicità e servizi di analisi indipendenti per fornire la* ***pubblicità mirata*** *basata sull'utilizzo da parte degli utenti del nostro sito Web"*?

22. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Noi utilizziamo fornitori di marketing, pubblicità e analisi di terze parti per fornire pubblicità basata sull'utilizzo da parte degli utenti del nostro sito web. Questa pubblicità è diretta a un pubblico con determinate caratteristiche, che possono essere demografiche (come: razza, condizione economica, sesso, età, generazione, livello di istruzione, livello di reddito e occupazione) o psicografiche (incentrate sui valori, sulla personalità, sull'atteggiamento, sull'opinione, sullo stile di vita e sull'interesse del consumatore)"*?

23. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Mettiamo a tua disposizione servizi, prodotti, applicazioni o skill forniti da* ***terze parti****"*?

24. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Mettiamo a tua disposizione servizi, prodotti, applicazioni o skill forniti da una persona fisica o giuridica, da un'autorità pubblica, da un servizio o da un altro organismo differente dal titolare del trattamento, o dal responsabile del trattamento, o dalle persone autorizzate al trattamento dei dati personali che operano sotto l'autorità diretta di questi ultimi"*?

25. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Le informazioni che raccogliamo includono* ***identificatori univoci****"*?

26. Quanto ti sentiresti a tuo agio nell'utilizzare un prodotto o servizio se l'informativa sul trattamento dei dati personali affermasse: *"Le informazioni che raccogliamo includono una stringa di caratteri utilizzabile per identificare un browser, un'app o un dispositivo in modo univoco"*?

Third survey's English version

The English version of the survey questions is as follows:

1. How comfortable would you be using a product or service if the data use policy says, *"We collect **Account Information** about you to provide our Core Services"*?

2. How comfortable would you be using a product or service if the data use policy says, *"We collect information such as your username, tax code, or email address to provide our Core Services"*?

3. How comfortable would you be using a product or service if the data use policy says, *"We collect **Personal Information** about you to provide our Core Services"*?

4. How comfortable would you be using a product or service if the data use policy says, *"We collect information such as your full name, address, and email to provide our Core Services"*?

5. How comfortable would you be using a product or service if the data use policy says, *"We collect **Public Information** about you to provide our Core Services"*?

6. How comfortable would you be using a product or service if the data use policy says, *"We collect information such as your username, profile picture, or social media posts to provide our Core Services"*?

7. How comfortable would you be using a product or service if the data use policy says, *"We collect **Sensitive Personal Information** about you to provide our Core Services"*?

8. How comfortable would you be using a product or service if the data use policy says, *"We collect information such as your race and ethnicity, political and religious views, or information about your health to provide our Core Services"*?

9. How comfortable would you be using a product or service if the privacy policy says, *"Your information may also be shared with others our services and **affiliates**, that may use them in a manner consistent with their privacy policies"*?

10. How comfortable would you be using a product or service if the privacy policy says, *"Your information may also be shared with others our Services or with any other company that controls or is controlled by our company, or otherwise that it is under common control with another one or with more companies. These companies may use that information in a manner consistent with their privacy policies"*?

11. How comfortable would you be using a product or service if the data use policy says, *"We only share **aggregated information** with third parties"*?

12. How comfortable would you be using a product or service if the privacy policy says, *"We share with third parties only information elements collated on a number of individuals"*?

13. How comfortable would you be using a product or service if the data use policy says, *"We only share **anonymized information** with third parties"*?

14. How comfortable would you be using a product or service if the privacy policy says, *"We only share with third parties previously identifiable information that has been de-identified and for which a code or other association for re-identification no longer exists"*?

15. How comfortable would you be using a product or service if the data use policy says, *"We only share **de-identified information** with third parties"*?

16. How comfortable would you be using a product or service if the privacy policy says, *"We only share with third parties records that have had enough personally identifiable information removed or obscured such that the remaining information does not identify an individual, and there is no reasonable basis to believe that the information can be used to identify an individual"*?

17. How comfortable would you be using a product or service if the privacy policy says, *"We use **location-related information** to provide, personalize and improve our products, including ads, for you and others"*?

18. How comfortable would you be using a product or service if the privacy policy says, *"We use information about you, such as where you live, the places you like to go, and the businesses and people you're near, to provide, personalize and improve our Products, including ads, for you and others"*?

19. How comfortable would you be using a product or service that has a **privacy policy**?

20. How comfortable would you be using a product or service that has a document describing which personal data are collected, the purposes of the processing, the rights of the data subjects, and all the third-party services with which such personal data are shared?

21. How comfortable would you be using a product or service if the privacy policy says, *"We use third-party marketing, advertising, and analytics providers to provide* ***targeted advertising*** *based on your use of our website"*?

22. How comfortable would you be using a product or service if the privacy policy says, *"We use third-party marketing, advertising, and analytics providers to provide advertising based on your use of our website. This advertising is directed towards an audience with certain traits that can either be demographic (such as: race, economic status, sex, age, generation, level of education, income level, and employment), or psychographic (focused on the consumer values, personality, attitude, opinion, lifestyle and interest)"*?

23. How comfortable would you be using a product or service if the privacy policy says, *"We make available to you services, products, applications, or skills provided by* ***third parties****"*?

24. How comfortable would you be using a product or service if the privacy policy says, *"We make available to you services, products, applications, or skills provided by a natural or legal person, by a public authority, by an agency or by a body other than the controller, or than the processor, or than persons who are authorized to process your personal data, operating under the direct authority of the latters"*?

25. How comfortable would you be using a product or service if the privacy policy says, *"The information we collect includes* ***unique identifiers****"*?

26. How comfortable would you be using a product or service if the privacy policy says, *"The information we collect includes a string of characters that can be used to uniquely identify a browser, app, or device"*?

Appendix X

GDPR in the Small: Appendix

X.1 Processing allowed in schools

Table X.1 lists some processing that commonly occurs in the school setting, indicating whether or not they are permitted and indicating any limitations, if any.

X.2 Preliminary survey text

The preliminary survey administered to the participants. The survey consists of data protection and demographic questions.

Data protection questions

This subsection lists the questions related to understanding the organizational structure for the data protection that we administered to the staff of the ICs participating in the study. At the bottom of each question, we specified whether it was always administered to participants or only following their responses that required further study.

PQ1 Do you know who a data protection officer (DPO) is?

(a) I do not know who the DPO is

(b) I know who a DPO is, but I do not think there is one in my educational institution

(c) I know who a DPO is, and I think there is one in my educational institution

This question was always administered.

Table X.1: Types of data and data processing carried out at schools

Type of data	**Processing description**	**Data owners**	**Allowed (Yes / No)**
Grades	Online publication	Students	No
Grades	Posting in the restricted area of the electronic registry with access limited to the student or family	Students	Yes
Personal data of pupils	Preparation of school communications not addressed to specific recipients / services related to the school activities (e.g. school bus, canteen).	Students	No
Health data	Dissemination of pupil health data (e.g., including the names of students with disabilities in a circular published online).	Students	No
Data on educational, intermediate, and final outcomes	Communication to third parties to facilitate career guidance, training, and job placement.	Students	Yes, by informing the data owners in advance.
Audio /photos /videos recorded at school or during school related activities	Dissemination on the Internet (website / social networks and so on)	Students / School staff / Parents	Yes, only with the explicit consent of the data owners.
Personal data	Use of distance education systems within institutional purposes.	Students, parents, and teachers	Yes, by informing the data owners in advance.
Audio/video recordings of lectures	Record the lecture for personal purposes only (e.g., for self-study).	Students and teachers	Yes, if it is allowed by school rules
Audio/video shooting	Dissemination on the web and social networks of audio/video footage of a previously recorded lecture for personal purposes.	Students and teachers	No, unless the exercisers of parental responsibility and other persons present give explicit informed consent
Video Shooting	Video recording of the lesson in which class dynamics are manifested.	Students and teachers	No, even if distance learning platforms are used.
Class composition data	Publication of data on the institutional website.	Students	No
Class composition data	Posting data in the restricted area of the electronic school register with access limited to the student or family.	Students	Yes (only first and last names).
Rankings of teachers and administrative technical and auxiliary staff (ATA)	Online publication of rankings for selection procedures	Teachers and ATA staff	Yes, but only data necessary to identify the candidate.
Video recordings	A video surveillance system inside the building protects the building and school property.	Members of the school community	Yes, but only at the end of class time and extracurricular activities.

PQ2 Do you know who is the DPO of your educational institution and how to contact him or her?

(a) I do not know who is and I do not know how to contact him/her

(b) I do not know who is, but I know how to contact him/her

(c) I know who is, but I do not know how to contact him/her

(d) I know who is and I know how to contact him /her

This question was administered only if the user selected option (c) in question PQ1.

PQ3 Do you know what a data processing notice related to processing is?

(a) I do not know what a data processing notice is

(b) I know what a data processing notice is, but I do not think we have one at our educational institution

(c) I know what a data processing notice is, and I think we have one at our educational institution

The question was always administered.

PQ4 Do you know (and where to find) the particular data processing notices related to each processing?

(a) I do not know them and I do not know where to find them

(b) I do not know them, but I know where to find them

(c) I know them, but I do not remember where to find them

(d) I know them, and I know where to find them

(e) I know them, but they exist only for some data processing

This question was administered only if the user selected option (c) in question PQ3.

PQ5 Do you know what a data protection policy is?

(a) I do not know what a data protection policy is

(b) I know what a data protection policy is, but I do not think we have one at our educational institution

(c) I know what a data protection policy is, and we have one at our educational institution

This question was always administered.

PQ6 Do you know (and do you know where to find) your organization's data protection policy?

(a) I do not know it and do not know where to find it

(b) I do not know it, but I know where to find it

(c) I know it, but I do not remember where to find it

(d) I know it, and I know where to find it

(e) I do not know, but the corporate data protection rules are based on practice

The question was administered only if the user selected option (c) in question PQ5.

PQ7 Do you know what a record of processing activities is?

(a) I do not know what a record of processing activities is

(b) I know what a record of processing activities is, but I do not think we have it

(c) I know what a record of processing activities is, and I think we have it

The question was always administered.

PQ8 Do you know (and do you know where to find) your organization's record of processing activities?

(a) I do not know it and do not know where to find it

(b) I do not know it, but I know where to find it

(c) I know it, but I do not remember where to find it

(d) I know it, and I know where to find it

(e) I do not know it, but the knowledge of each processing is based on practice

The question was administered only if the user selected option (c) in question PQ7.

PQ9 Do you know what a data protection impact assessment (DPIA) is?

(a) I do not know what a DPIA is

(b) I know what a DPIA is, but I do not think we have conducted it

(c) I know what a DPIA is, and I think we have conducted it

The question was always administered.

PQ10 Do you know (and do you know where to find) your organization's data protection impact assessment (DPIA)?

(a) I do not know it and do not know where to find it

(b) I do not know it, but I know where to find it

(c) I know it, but I do not remember where to find it

(d) I know it, and I know where to find it

(e) I do not know, but I know that some assessment has been conducted informally

The question was administered only if the user selected option (c) in question PQ9.

PQ11 Do you know what a personal data breach reporting procedure is?

(a) I do not know what a personal data breach reporting procedure is

(b) I know what a personal data breach reporting procedure is, but I do not think we have it

(c) I know what a personal data breach reporting procedure is, and I think we have it

The question was always administered.

PQ12 Do you know (and do you know where to find) your organization's personal data breach reporting procedure?

(a) I do not know it and do not know where to find it

(b) I do not know it, but I know where to find it

(c) I know it, but I do not remember where to find it

(d) I know it, and I know where to find it

(e) I do not know, but the charged staff informally knows what to do

The question was administered only if the user selected option (c) in question PQ11.

PQ13 Do you know what a formal personal data protection organizational chart is?

(a) I do not know what it is

(b) I know what it is, but I do not think we have it

(c) I know what it is, and I think we have it

The question was always administered.

PQ14 Do you know (and do you know where to find) your organization's formal personal data protection organizational chart?

(a) I do not know it and I do not know where to find it

(b) I do not know it, but I know where to find it

(c) I know it, but I do not remember where to find it

(d) I know it, and I know where to find it

(e) I do not know it, but the charged staff informally knows it

The question was administered only if the user selected option (c) in question PQ13.

PQ15 Do you know what technical and organizational measures for processing security are?

(a) I do not know what they are

(b) I know what they are, but I do not think we have them

(c) I know what they are, and I think we have them

The question was always administered.

PQ16 Do you know (and do you know where to find) the technical and organizational measures for processing security formally adopted by your organization?

(a) I do not know them and I do not know where to find them

(b) I do not know them, but I know where to find them

(c) I know them, but I do not remember where to find them

(d) I know them, and I know where to find them

The question was administered only if the user selected option (c) in question PQ15.

PQ17 Do you know what a formal procedure for data subjects' rights management is?

(a) I do not know what it is

(b) I know what it is, but I do not think we have it

(c) I know what it is, and I think we have it

The question was always administered.

PQ18 Do you know (and do you know where to find) your organization's formal procedure for data subjects' rights management?

(a) I do not know it and I do not know where to find it

(b) I do not know it, but I know where to find it

(c) I know it, but I do not remember where to find it

(d) I know it, and I know where to find it

(e) I do not know, but the charged staff informally knows how to handle a data subject request

The question was administered only if the user selected option (c) in question PQ17.

PQ19 What is a formal designation of natural persons (other than the processor) to handle the processing of personal data?

(a) I do not know what it is

(b) I know what it is, but I do not think we have it

(c) I know what it is, and I think we have it

The question was always administered.

PQ20 Do you know (and do you know where to find) the formal designations of natural persons (other than the processor) to handle processing personal data adopted from your organization?

(a) I do not know them and I do not know where to find them

(b) I do not know them, but I know where to find them

(c) I know them, but I do not remember where to find them

(d) I know them, and I know where to find them

(e) I do not know them, but the charged staff are informally designated

The question was administered only if the user selected option (c) in question PQ19.

PQ21 Do you know the training in data protection for natural persons in charge of processing?

(a) I do not know what it is

(b) I know what it is, but I do not think we do it

(c) I know what it is, and I think we do it

The question was always administered.

PQ22 How does your organization train in data protection the natural persons in charge of processing? *[Choose all applicable options]*

(a) Attendance of internal training courses

(b) Attendance of external training courses

(c) Self-instruction courses

(d) Information received from the DPO

(e) Guidance received from superiors

(f) Consultation of institutions (e.g., competent Supervisory Authority, EDPS, EDPB) informational materials

(g) Consultation of non-institutional informational material

(h) I have yet to be trained

The question was administered only if the user selected option (c) in question PQ21.

Demographic questions This subsection lists the demographic questions we administered to the ICs' staff participating in the study.

DQ1 In which country do you currently reside?

(a) *[Select one of the countries from the proposed list]*

This question was always administered.

DQ2 In what type of educational institution do you primarily work?

(a) Preschool school (private or public)

(b) Primary school (private or public)

(c) Lower secondary school (private or public)

(d) Upper secondary school (private or public)

(e) Higher school (private or public)

This question was always administered.

DQ3 Which size is the institution you work for?

(a) Micro (less than 10 employees)

(b) Small (less than 50 employees)

(c) Medium (less than 250 employees)

(d) Large (250 employees or more)

This question was always administered.

DQ4 Which of the following roles most closely matches the one in which you are employed?

(a) School principal/director of general administrative services

(b) Secretarial staff

(c) School aides

(d) Preschool teacher

(e) Elementary school teacher

(f) Secondary school teacher

(g) Other

This question was always administered.

DQ5 How many years of working experience do you have?

(a) 0 - 5

(b) 6 - 10

(c) 11 - 15

(d) 16 - 20

(e) 21 - 25

(f) 26 - 30

(g) 31 - 35

(h) 36 - 40

(i) 41 or more

This question was always administered.

X.3 Annotations' template fields

Researcher:

[To complete by entering the text]

Date and time of observation:

[To complete by entering the text]

Place of observation:

[To complete by entering the text]

Organization involved:

[To complete by entering the text]

The researcher should begin by standing back and letting the scene unfold. Since capturing every bit of what is going on in a setting is impossible, the observer should start by jotting a few general notes describing the area under observation, as follows.

Necessary: Should the researcher accidentally become aware of health-related data, data belonging to special categories, or judicial data during the observation activities, they should keep these data private by not recording them. The researcher will note the observed fact but will do so in such a way that the reported data are anonymous and not attributable to a specific respondent.

General notes

Description of the area under observation
[Some rows to complete by entering the text of the description]

Please describe the following
What routine activities are going on
[Some rows to complete by entering the text of the description]

The personnel present
[Some rows to complete by entering the text of the description]

The pace and timing of activities
[Some rows to complete by entering the text of the description]

The interactions taking place
[Some rows to complete by entering the text of the description]

Other relevant elements
[Some rows to complete by entering the text of the description]

Interesting incidents or happenings

In this section, researchers must note incidents or happenings that seem particularly interesting. Please specify what is happening, what is being said and done, by whom, where, and with what consequences.

[Some rows to complete by entering the text of the description]

Follow up

The researcher can then follow up with questions about the incident with those involved and perhaps even come up with a concept or two to describe what was happening in the situation. This concept can later be followed up on in future observations in the same or another setting.

[(Eventually) Some rows to complete by entering the text of the description]

X.4 Units' coding

Table X.2 shows how it is possible to code similar units differently. More specifically, the table reports some annotations related to staff's members leaving a room without locking a computer. Although all the units share some codes (e.g., *A* (administrative staff), *DI* (digital data), and *F* (compliance failure)), different codes were used to specify the context in which observations about data were performed. Thus, for example, the *TP* (third parties) code was used if (i) third parties were in the staffed room (Table X.2, QIDs: 129 and 982), or (ii) the room was unmanned (Table X.2, QIDs: 402 and 1557), allowing third parties entering undisturbed, as our investigator did.

X.5 Codebook template

Code: A label (4-12 characters) describing the code content.

Long code: (Optional) A short sentence (20-80 characters) that captures the essence of the theme or theme component(s) that the code is intended to signify.

Description: A paragraph that details the key features of the theme or theme component(s) that the code is intended to signify.

When to use it (examples): It provides guidelines / examples on when the code should be used.

When not to use it (examples): It provides caveats / examples on when the code should not be used.

Table X.2: Examples of similar units coded in a different way

QID	Annotations	Coding
19	OP02 exits the room, leaving documents on the desk and the computer unlocked, showing some content on the screen. [...] Only OP01 remained in the room. [...] Data is exposed for 4 minutes. [...] No third parties are present in the room.	A-IA-II-DI-PH-M-F
91	OP02 exits the manned room, leaving the computer unlocked and the monitor turned on. [...] Data is exposed for 8 minutes. No third parties are present in the room [...].	A-IA-II-DI-M-F
129	OP04's work is often interrupted by parental visits [...]. OP04 performs many tasks quickly and often has to get up and leave; in doing so, they leave the computer unlocked and the monitor turned on. During their absence, however, the room remained staffed.	A-TP-IA-II-DI-M-F
982	OP03's computer is turned on and unlocked. There is an application (a spreadsheet) opened on the screen from which I can read data. The workstation remained unattended even while OP03 is not in the room, which is never left empty. [...] Some third parties routinely enter the room to perform certain activities.	A-TP-IA-II-DI-M-F
402	OP01 exits the room, leaving the door and the computer unlocked (with applications opened). For the first time since the observation began, the room is unmanned. Data is exposed for 2 minutes. [...] No third parties are in the room.	A-TP-IA-II-DI-M-F
1557	In Lab 01 [...], a technical assistant's computer is turned on and unlocked. The lab has been unattended for a long time.	A-TP-IA-II-DI-M-F

X.6 Codebook

Tables X.3, X.4, X.5, X.6, and X.7 contain the entire codebook researchers used in this study. Specifically, Table X.5 shows the types of interactions between actors, Table X.3 shows the types of data that actors processed, Table X.4 shows the types of actors involved, Table X.6 shows the processing flow actions, and Table X.7 shows the possible type of failure in processing.

X.7 Day-to-day contribution to each code

In Table X.8, codes are grouped according to the type of entities in the data path and ordered from top to bottom in decreasing order based on the number of contributing days of observation. In parallel, the observation days are in order of date, leftmost first. The colored cells show the observation's day-to-day contribution to each code. The red color shows the days when observations occurred predominantly in administrative offices, while the blue color shows the days when observations occurred predominantly in classrooms, laboratories, or teaching rooms. We used the following codes to represent the entities in the data path: 1 - *Types of data*, 2 - *People accessing and manipulating data*, 3 - *Interactions between people processing data*, 4 - *Types of data processing*, and 5 - *Types of failure in processing.*

X.8 Violation with unclear impact

Table X.9 shows some examples of security violations related to a staff's incorrect use of tools that results in an unclear privacy impact.

X.9 Excerpt of quotes' text

Table X.10 shows the text of the quotations referred to in the article in Subsections: 5.6.2, 5.6.4, and 5.6.5.

Table X.11 shows the text of the quotations referred to in the article in Subsections: 5.6.8, 5.6.6, and 5.6.10.

Table X.12 shows the text of the quotations referred to in the article in Sections 5.7 and 5.10.

Table X.3: Codebook: Types of data

The table shows the codebook's portion related to the types of data processed by actors.

Code	**Long code**	**Description**	**When to use it (examples)**	**When not to use it (examples)**
PH	Physical data	It indicates physical data (e.g., paper documents).	A person is writing a paper document.	A person is writing a text on a file.
DI	Digital data	It indicates digital data (e.g. files).	A person is writing text on a file. A person forwards a WhatsApp voice message to another person.	A person is writing a paper document.
AU	Audio data	It indicates data reproduced in audible form, and available to people only for a certain amount of time and in a certain location.	An person verbally communicates data to another operator. A person listen to a WhatsApp voice message.	A person forwards a WhatsApp voice message to another person.

Table X.4: Codebook: People accessing and manipulating data

The table shows the codebook's portion related to the types of actors involved.

Code	**Long code**	**Description**	**When to use it (examples)**	**When not to use it (examples)**
T	Teaching staff	It indicates that a member of the teaching staff is involved in data processing.	A teaching staff member is present on the field.	Only an administrative or managerial staff member is present on the field.
A	Administrative staff	It indicates that a member of the administrative staff (e.g., janitors, administrative, technical, and auxiliary staff) is involved in data processing.	An administrative staff member is present on the field	Only a teaching or management staff member is present on the field.
H	Head	It indicates that a member of the the management staff (e.g., the school's principal or the general and administrative services director) is involved in data processing.	A management staff member is present on the field.	Only an administrative or teaching staff member is on the field.
TP	Third Party	It indicates that someone else does not belonging to the previous categories is involved in data processing	Someone not authorized for processing, such as a parent or student, is on the filed.	A person authorized for processing, such as a staff member, is on the field.

Table X.5: Codebook: Interactions between people processing data

The table shows the codebook's portion related to types of interactions between actors.

Code	**Long code**	**Description**	**When to use it (examples)**	**When not to use it (examples)**
IA	Individual action	It indicates if a person takes an individual action on data	A person leaves a document in a room without interacting with others (excluding, for example, simple courtesy greetings).	A person leaves a document in a room, discussing its content with others there.
I	Interaction	It indicates if there is a deliberate interaction between two or more people processing data.	Two or more people talk about data.	One person knocks on the door of a room where there is another person and then, without interacting, leaves a document on the table. Two or more people greet each other without talking about or manipulating data.
II	Indirect interaction	It indicates if interactions between two or more people occurs at the presence of a third party not directly involved in data processing. **Jointly use:** If the indirect interaction occurs due to a compliance failure, Code *II* should be used jointly with Code *F*. Otherwise, if the indirect interaction occurs due to a resource failure, code *II* should be used in conjunction with codes *F* and *R*.	A third party overhears a confidential conversation in another room.	Two persons talk to each other about data or data processing.
MI	Missing interaction	It indicates a lack of interaction between two or more people processing data, i.e., interactions during which some people involved in data processing do not behave as they should.	The *MI* code will be used for interactions between people only. A person asks another one a data processing-related question without having a response. A persons call another one to ask something, but the other person is missing. A person can access a protected area freely because the front desk staff is absent.	It must not used for interactions between humans and machines. A person wants to print a document, but the printer jams, causing the action to end because the person moves on, giving up printing. A person looks for another person to talk to them but does not find them

Table X.6: Codebook: Types of data processing

The table shows the codebook's portion related to the processing flow actions.

Code	**Long code**	**Description**	**When to use it**	**When not to use it**
C	Creating data	It indicate a new data (physical or digital) is created **Jointly use:** Code *C* must be used with codes *DI* or *PH* to specify the data type.	A person is filling out a paper form. A person is writing an email.	A person is photocopying a document. A person is copying a file.
2P	To physical	It indicates that data is converted from a digital to a physical form or, if already physical, duplicated.	A person prints a PDF file. A person photocopies a paper document.	A person sends a file as an attachment to an email message. A person scans a paper document.
2D	To digital	It indicates that data is is converted from a physical to a digital form or, if already digital, duplicated.	A person scans a paper document, transforming it into a PDF file. A person attaches a file to an email.	A person photocopies a paper document. A person transfers a copy of a file into a USB flash drive.
Q	Query on digital or physical data	It indicates data consultation. **Jointly use:** This code must be used together with *DI* or *PH* to indicate digital or physical data usage, respectively.	A person performs a search in the document management system. A person takes a hard copy file and consults it.	A person opens a digital file to print it without reading its content. A person takes a hard copy file without consulting its content.
SP	Store Physical	It indicates that a physical data is voluntarily stored somewhere (or it is permanently archived). **Jointly use:** If the procedure leads a person to store data in a folder on a table or a desk, codes *F*, *R*, and *SP* will jointly mark the resource failure in addition to the physical data store. Let's suppose the procedure leads a person to store data into an unlocked closet. In that case, the *F* and *SP* codes will jointly mark the failure of compliance in addition to the physical storage of the data.	A person places a document in a locked closed. A person picks up a document and then puts it on a table.	A person temporarily leaves on a table a document to be photocopied.
DTT	Data transfer transient	It indicates a transient transfer of data from a person to another one without that the first person looses the possibility to process data.	A person shows a document to another person but does not hand it over. A person shows the content of an on-screen file to another person.	A person delivers a document into the hands of another person.
DTR	Data transfer real	It indicates that a data is transferred (for a while / definitely) from a person to another one. **Alternative coding:** In case of a digital file transfer (e.g., via email), *2D* will be used (because data is not transferred but merely duplicated).	A person gives a document to another person. A person moves a file to a thumb drive and gives it to another person.	A person downloads a file from the document management system.
M	Move on	It indicates that a person processing data breaks its link with it. **Jointly use:** When the persons breaks its link with data due to an interaction with another person (that is disruptive), this code should be used in conjunction with the *I* code. For example, a person sends a document to a printer. However, before they pick it up from the printer, another person stops them to ask something.	When a person leaves a document on a table and then exits the room.	We should not use it when the person's link with data fades without a violation resulting from this. When a person leaves a document on a room's table and then moves in the room for performing other tasks, but no third party is present in the room. A person leaves a document on a room's table and exits the room to make a photocopy.

Table X.7: Codebook: Types of failure in processing

The table shows the codebook's portion related to the possible types of failure in processing.

Code	**Long code**	**Description**	**When to use it**	**When not to use it**
F	Failure	It indicates if something within the processing procedure does not work as expected and generates an error or failure (lack of compliance).	A printer does not print a document because of a paper jam. A person cannot access an online service because of an application error. A person leaves a room without locking the computer's operating system. A person exits a room, leaving open documents on the desk. A person cannot consult a document because of an error generated by the document management system. Two or more people discuss processing in a publicly accessible place like a hallway. A person leaves a document unattended in the printer. While two or more people talk about processing in a publicly accessible place, there is at least one unauthorized person that could overhear what is said.	A person exits a room and leaves open documents on the desk because another person is processing them. An operator leaves an unlocked document on their desk in an access-controlled room where only another operator authorized to process the data in the document is present.
R	Resources failure	It indicates if a procedure or process fails due to a lack of resources. Thus, it is used also when two or more procedures have conflicting requirements that fail. **Jointly use:** This code must be used with the *F* code, as it specifies a particular failure circumstance.	In a room, an unauthorized person can listen to talks about data because of the physical layout of the premises and compliance with the organization's procedures. A single janitor has to operate the reception desk and simultaneously be absent to perform activities elsewhere. An unauthorized person can hear the conversation in another room because of the unsound-proofed walls.	Because of non-compliance with the organization's procedures in a room, there are several people inside, not all authorized, who can listen to talks about data.

Table X.8: The contribution that each day of observation makes to each code

ID	Codes (text)	Entity	01	02	03	04	05	06	07	08	09	10	11	12	13	Observa
PH	Physical data	1														
AU	Audio data	1														
DI	Digital data	1														
T	Teaching staff	2														
A	Administrative staff	2														
TP	Third Party	2														
H	Head	2														
IA	Individual action	3														
II	Indirect interaction	3														
I	Interaction	3														
MI	Missing interaction	3														
C	Creating data	4														
Q	Query on digital or physical data	4														
SP	Store Physical	4														
DTT	Data transfer transient	4														
M	Move on	4														
DTR	Data transfer real	4														
2P	To physical	4														
2D	To digital	4														
F	Failure	5														
R	Resources failure	5														

Table X.9: Examples of security violations with unclear privacy impact

Table X.11 of Section X.9 contains quotes' text.

Quotes IDs	**Premises type**	**Violation type**	**Setup**						Risk
			Room empty	**Door open**	**Staff present**	**3rd parties present**	...	**Info**	
248 & 245	Classroom	Unlocked computer	No	No	Yes	Yes	...	unknown	Low
816 & 832	Teachers' lounge	Unlocked computer	Yes	Yes	No	Yes	...	possibly sensitive	Medium-High
1611 & 1617	Computers lab	Unlocked computer	No	Yes	Yes	Yes	...	unknown	Low
291 & 299	Teachers' lounge	Unlocked computer	Yes	Yes	No	No			
330 & 329	Labroom	Missing anti-theft cable	Yes	Yes	No	No			
1568 & 1567	Classroom	Missing anti-theft cable	No	Yes	Yes	No			
742 & 734	Classroom	Missing anti-theft cable	Yes	Yes	No	No	...	unknown	investigate

Table X.10: Excerpt of quotes' text on IT physical security, bulletin boards, and staff habits

IDs	Topic	Annotations	Coding
794	Doors left open	The door to the server room and the rack cabinets inside (containing the institute's server and network equipment) were not locked. It happened that someone entered the room without anyone noticing the presence.	A-TP-IA-DI-M-F
1589	Doors left open	A teacher consulted with parents in one classroom by holding the door open. While pausing in the hallway, someone could hear what the parents and the teacher were saying to each other.	T-TP-I-II-AU-DTT-F
1596	Doors left open	Before moving on to another premise, the janitor failed to lock the guardhouse door with video surveillance monitors.	A-TP-IA-II-DI-M-F
234	Unmanned classrooms	Teachers did not lock the front door of classrooms before going outside. During the lunch break, they were all open and deserted.	T-TP-IA-MI-DI-PH-M-F-R
734	Unmanned classrooms	The doors of the classrooms on the second floor are all open, and the classrooms are unmanned.	T-TP-IA-MI-DI-PH- M-F-R
236	Unlocked computers	The computer in classroom 20 is turned on and unlocked, with documents opened on the screen.	T-TP-IA-II-DI-M-F
880	Unlocked computers	The computer in faculty room 2 is always turned on and unlocked with the Gmail screen open.	T-TP-IA-II-DI-M-F
1597	Unlocked computers	The laptop computer in the classroom is turned on and unlocked. A teacher does not man the classroom.	T-TP-IA-II-DI-M-F
1156	Staff Habits	OP01 stops to talk with a colleague in the secretarial corridor. Because third-party subjects habitually walk in that corridor, unauthorized listening may occur.	A-TP-I-II-AU-DTT-F
68	Staff Habits	Many post-its with e-mail addresses or phone numbers (e.g., from technical support or suppliers) exist on the workstations' monitors.	A-IA-PH-SP-F
189	Staff Habits	On OP06's desk are post-it notes containing first and last names, or company names, with corresponding phone numbers.	A-IA-PH-SP-F
522	Staff Habits	OP03 ends a personal call on Google Meet and disconnects from the application. The operator used the school's computer equipment for personal reasons.	A-IA-DI-Q-F
962	Information on bulletin board	Inside the classroom, the children's names are written on a document on the door, along with their ways of leaving school and transportation (e.g., alone on foot or by school bus).	T-IA-PH
970	Information on bulletin board	Behind the faculty room door, a sheet containing all pupils' "sensitive" data is hanging. The data include whether they do or do not attend the Catholic religion hour, their first and last name, the class attended, whether they eat in the cafeteria, and whether they use bus transportation. The room is accessible by third parties, and the data are exposed.	T-TP-IA-II-PH-F-R
946	Information on bulletin board	The annual duties sheet with the children's names is on the classroom wall.	T-IA-PH
973	Information on bulletin board	Some documents, such as the substitution, hearing, and committee attendance sheets, have teachers' first and last names written on them on the faculty room bulletin board.	T-IA-PH
908	Information on bulletin board	On the laboratory wall bulletin boards are sheets with the computers' accounts and passwords.	A-TP-IA-II-DI-PH-M-F

Table X.11: Excerpt of quotes' text on staff's incorrect use of tools, shadow IT, or mistakes in the organizational procedures

IDs	Topic	Annotations	Coding
245	Unlocked computer	The computer is unlocked in the classroom, and the big screen monitor is turned on.	T-IA-II-DI-M-F
248	Third party presence	During the lesson in the classroom, a university intern was present who observed the children working and taking notes on a tablet.	TP-IA-DI-C
816	Third party presence	An educator from an outside cooperative enters the unmanned faculty room and consults the absent staff substitution book (which notes the teachers' first and last names, class, schedule, and handwritten signatures). After one minute, the educator leaves the room.	TP-IA-PH-Q
832	Unlocked computer	The second computer in the faculty room is unlocked and allows free access to the IC's network and the Internet.	T-TP-IA-II-DI- M-F
1611	Unlocked computer	In a computer lab where third parties are present, a computer is turned on and left unattended and unlocked, with a PDF viewer window open and a salary slip screen visible on the monitor. The data remained exposed for the duration of the observation.	T-TP-IA-II-M-F
1617	Third party presence	The student's school representative entered a lab and shared some information about a school trip. After 2 minutes, the representative left the lab.	TP-I-AU-DTT
291	Unlocked computer	In the faculty room, which is unmanned and with the door open, a computer workstation is turned on, unlocked, and with the monitor turned on.	T-TP-IA-II-DI- M-F
299	Unmanned and open room	Due to an incorrectly designed business procedure, the faculty room was unmanned, open, and empty.	T-TP-IA-II-M- F-R
329	Unmanned and open room	The great hall door was open, and the great hall was unmanned and empty.	T-TP-IA-MI-DI-M-F-R
330	Missing anti-theft cable	A laptop computer is turned off and not connected to an anti-theft cable in the unmanned great hall.	T-TP-IA-II-DI- M-F-R
1567	Unlocked computer	A computer is turned on and unlocked in a classroom. The teacher is present here, but he/she is not using it.	T-IA-II-DI-M-F
1568	Missing anti-theft cable	In classroom 17 there is a laptop computer without an anti-theft cable.	T-IA-DI-F-R
734	Unmanned and open room	The classrooms on the second floor have open doors and are unmanned.	T-TP-IA-MI-DI-PH-M-F-R
742	Missing anti-theft cable	The laptop computer connected to the classroom's big screen is turned on and not attached to an anti-theft cable.	T-TP-IA-II-DI- M-F-R
928	Wrong procedure	The entrance door is broken and does not automatically close by itself; hence, a janitor usually tells the visitors, asking them to wait for him to come and close it. The janitor acts this way because the IC management did not fix the entrance door.	A-TP-I-AU- DTT-F-R
929	Wrong procedure	When a visitor left the building, no one was manning the entrance. Because the door's automatic locking mechanism is broken, if the visitor did not close the entrance door manually, it would have remained open, allowing unauthorized persons access.	A-TP-IA-MI-F-R
777	Unmanaged PC	In the faculty room the teacher MT12 uses a personal laptop to access his e-mail.	T-IA-DI-Q-F
780	Unmanaged PC	In addition to the classroom desktop computer, on the desk, there is a laptop computer that is personal to the teacher. The laptop is turned on and unlocked; sometimes, the teacher uses it to interact with applications.	T-IA-DI-Q
1574	Unmanaged PC	ST03 turns off the personal laptop computer, picks up a folder with students' homework and the same laptop computer, and then leaves the classroom going away.	T-IA-DI-PH-M

Table X.12: Excerpt of quotes' text on failures

IDs	Topic	Annotations	Coding
626	Unattended/unlocked room	There are some empty classrooms with computers turned off. Laboratory 01 is empty, with the door open and the computers turned off.	T-TP-IA-MI-DI-PH-M-F-R
879	Unattended/unlocked room	All the second-floor classrooms are open and empty because teachers and pupils have recess in the courtyard.	T-TP-IA-MI-DI-PH-M-F-R
1417	Unattended/unlocked room	The doors to five classrooms on the second floor, where no classes were in progress, were closed but not locked.	T-TP-IA-MI-DI-PH-M-F-R
234	Unattended/unlocked room	Teachers did not lock the doors of classrooms before going outside for the lunch break.	T-TP-IA-MI-DI-PH-M-F-R
309	Unattended/unlocked room	After the class, everyone left the laboratory at 11:00 a.m., which remained empty and with the entrance door unlocked	T-TP-IA-MI-DI-PH-M-F-R
680	Unattended/unlocked room	The door to laboratory room 02 was left open when it was not in use because no classes were in progress.	A-IA-DI-PH-2P-M-F
1444	Unattended/unlocked room	Laboratory 01 was open and unmanned. Someone entered it without anyone noticing and stayed there.	A-T-TP-IA-MI-M-F-R
484	Unattended/unlocked room	The teachers who left did not lock the door to the faculty room, hence now the room is unmanned.	T-TP-IA-MI-PH-M-F-R
339	Unattended/unlocked room	The last teacher to leave the faculty room left the door unlocked, so the room is now unmanned.	T-TP-IA-II-MI-M-F-R
628	Unattended/unlocked room	The door to the faculty room was open, and the room was empty, so someone went to the unmanned room and stayed there for five minutes.	T-TP-IA-MI-M-F-R
884	Unattended/unlocked room	The door to the janitorial office is open and the room is empty because janitor JAN6 is busy on another floor of the building, so the office is unattended.	A-TP-IA-MI-M-F-R
27	Unattended/unlocked room	The room is sometimes left unattended and open, with documents on the table in plain view and computers with the operating system unlocked.	A-TP-IA-II-MI-DI-PH-M-SP-F-R
567	Unattended/unlocked room	On two occasions, the operators were absent at the same time, and the room was left open and unattended for a short period of time (about 1-2 minutes).	A-H-TP-IA-MI-M-F-R
623	Lack of manning	Because the janitors' number was inadequate, someone freely walked around the institution for ten minutes, taking advantage of the fact that no one was at the school entrance.	A-TP-IA-MI-M-F-R
1414	Lack of manning	Because no one was at the reception desk, someone entered the school undisturbed and walked down the corridor until, on the first floor, they found a cage halfway down the corridor, where JAN2 was manning the corridor with a colleague.	A-TP-IA-MI-M-F-R
834	Lack of manning	Between 10:50 a.m. and 11:15 a.m., the front desk was never manned, even though third parties routinely went to the reception desk. Contextually, the janitor was intent on cleaning the hallways and the outside of the building.	A-TP-IA-MI-M-F-R
923	Lack of manning	The janitor's desk in the hallway is almost always unmanned because JAN7 has to move between floors.	A-TP-IA-MI-M-F-R
988	Lack of manning	JAN1 was called to clean some rooms and left, so both the reception desk and the corridor in front of the secretariats remained unmanned. JAN1 returned after 9 minutes to clean the corridor in front of the secretariats. Meanwhile, the reception desk remained unmanned even if third parties went there.	A-TP-IA-MI-M-F-R
568	Excessive turnover	Because of the high staff turnover (often on fixed-term contracts), a change in the composition of the working group happened in conjunction with a regulatory deadline, resulting in missing the deadline because it was not known who was supposed to do what.	A-H-IA-MI-M-F-R
577	Excessive turnover	In the preceding months, a significant change in administrative staffing (including the top management figure) occurred close to a major regulatory deadline.	A-H-IA-MI-M-F-R
85	Unattended backup data	The IC's server system is in an unsuitable room (it is an administrative office accessible from third parties, with the door always left open and sometimes unattended). The server is locked and requires a login, but the two hard drives with the backup data are easily removable, as they are external USBs.	A-TP-IA-MI-DI-M-F-R
203	Staff absence	OP03 leaves the room for 5 minutes to retrieve a document from the printers, leaving the room unattended with the door open and the computer's operating system unlocked, but the monitor turned off.	A-TP-IA-II-MI-DI-PH-2P-M-F
498	Staff absence	OP03 exits the room for 2 minutes to go to the bathroom, leaving the room unattended, with the door open, the computer unlocked, and applications open on the screen.	A-IA-II-MI-DI-M-F
513	Staff absence	All the operators went outside the room for 2 minutes to scan a document, leaving the door open. While the room was unattended, the computers remained unlocked, with applications open.	A-H-IA-I-II-MI-DI-PH-2D- M-F
889	Staff absence	Teacher PT06 left classroom 24 for less than a minute to fetch something from another classroom. Only the children were left in the classroom, and the desktop computer remained unlocked.	T-TP-IA-II-MI-DI-M-F
971	Wrong procedure	The faculty room, open to third parties, has a blackboard where the WiFi password is written.	T-TP-IA-II-DI-PH-M-F
908	Wrong procedure	Attached to the laboratory's bulletin boards are sheets holding the login credentials for accessing the computers.	A-TP-IA-II-DI-PH-M-F
300	Wrong procedure	In the faculty room, all teachers use a shared workstation. They share the same Windows account (autologin is enabled), and only then do they use their management application account.	T-TP-IA-II-DI-C-Q-F-R
332	Wrong procedure	The computers in laboratory 03 used Windows autologin. Auto-logoff from applications is not enabled, so browsing mail from users who have not logged off is possible.	A-T-TP-IA-II-DI-F-R
1249	Inadequate premises	While OP02 is discussing a student's education plan with a support teacher, OP10 and another teacher enter the room to call other parents. Since all of them are talking simultaneously, unauthorized people may learn confidential information.	A-T-I-II-AU-PH-C-DTT-F-R

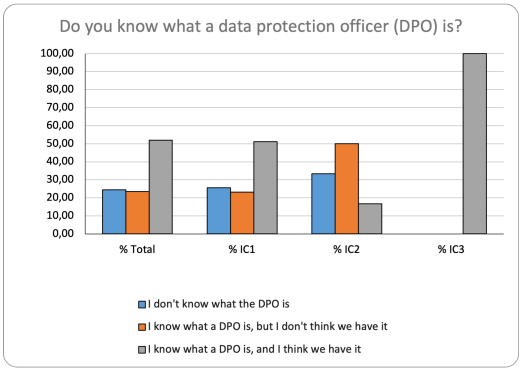

Figure X.1: Results (in percentage) of staff's knowledge about the role of a DPO

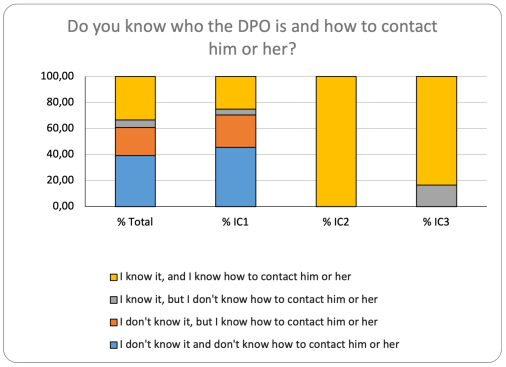

Figure X.2: Results (in percentage) of staff knowledge about who the DPO is and how to contact them

X.10 Survey's results

Figures X.1 and X.2 illustrate the preliminary survey results about staff knowledge of the DPO role. Figures depict the results about staff knowledge of the DPIA (Figure X.3), the record of processing activities (Figure X.4), the data breach management (Figure X.5), and the data subject's rights management (Figure X.6), respectively. Finally, Figures X.7 and X.8 illustrate the results about staff training in data protection.

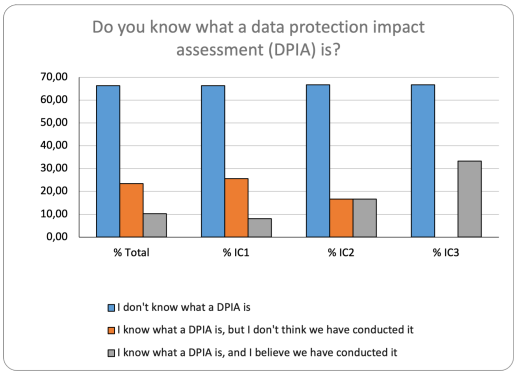

Figure X.3: Results (in percentage) of staff's knowledge of a data protection impact assessment (DPIA)

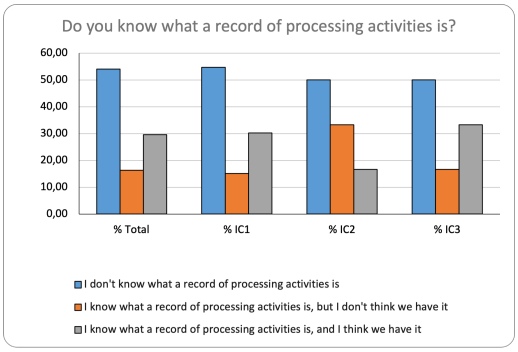

Figure X.4: Results (in percentage) of staff's knowledge of a record of processing activities

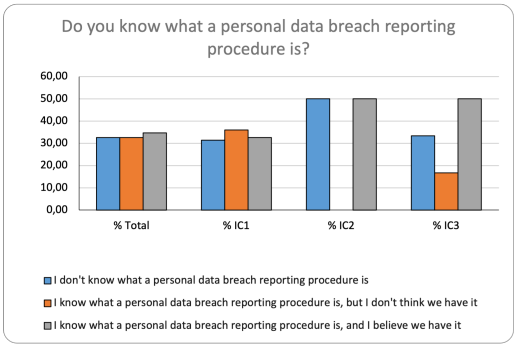

Figure X.5: Results (in percentage) of staff's knowledge of personal data breach reporting procedure

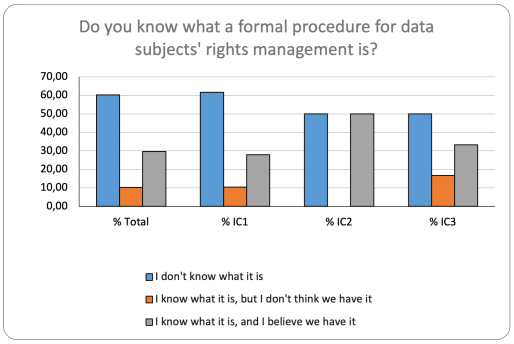

Figure X.6: Results (in percentage) of staff's knowledge of a formal procedure for data subject's rights management

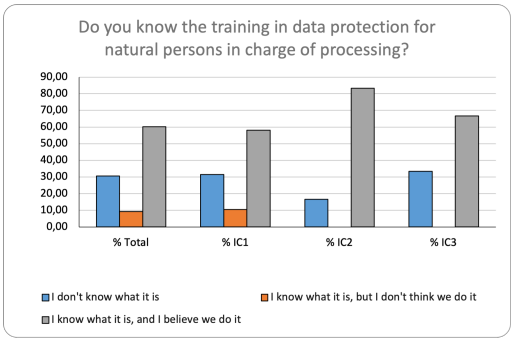

Figure X.7: Results (in percentage) of staff's knowledge of a training in data protection

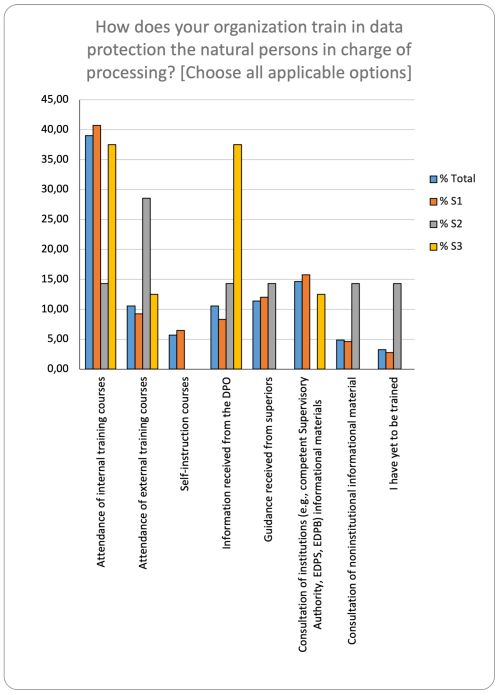

Figure X.8: Results (in percentage) of staff's members that received a specific training in data protection

Appendix Y

GDPR in the Small: Reference legislation

Listed below are references to Italian legislation dealing with the protection of personal data in the school setting.

Personal data protection legislation:

- 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). [Online]. Available: `https://eur-lex.europa.eu/eli/reg/2016/679/oj`

- Codice in materia di protezione dei dati personali - d.lgs.30/06/2003, n. 196. [Online]. Available: `https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9042678`

Opinions, guidelines, and ad hoc measures of the Italian supervisory authority:

- "Linee guida in materia di trattamento di dati personali dilavoratori per finalità di gestione del rapporto di lavoro in ambito pubblico", del 14 giugno 2007. [Online]. `https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/1417809`

- Provvedimento in materia di videosorveglianza - 8 aprile 2010. [Online]. `https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/1712680`

- Trattamento di dati personali per l'iscrizione dei bambini all'asilonido comunale - 6 giugno 2013. [Online]. https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/2554925

- Parere su una bozza di circolare relativa alle iscrizioni alle scuoledell'infanzia e alle scuole di ogni ordine e grado, cartacee eonline, per l'anno scolastico 2014/2015 - 12 dicembre 2013. [Online]. https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/2894420

- Linee guida in materia di trattamento di dati personali,contenuti anche in atti e documenti amministrativi, effettuatoper finalità di pubblicità e trasparenza sul web da soggettipubblici e da altri enti obbligati - 15 maggio 2014. [Online]. https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/3134436

- Parere su uno schema di decreto legislativo concernente larevisione e semplificazione delle disposizioni di prevenzionedella corruzione, pubblicità e trasparenza - 3 marzo 2016. [Online]. https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/4772830

- Parere su uno schema di decreto volto a disciplinare il periododi conservazione di alcune tipologie di dati personali relativi aglistudenti, acquisiti all'Anagrafe nazionale degli studenti (ANS) - 21 aprile 2016. [Online]. https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/5029548

- Parere sullo schema di regolamento del Ministero dell'Istruzionedell'università e della ricerca relativo all'integrazionedell'Anagrafe nazionale degli studenti con i dati sulla disabilitàdegli alunni - 15 ottobre 2015. [Online]. https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/4448995

- Parere sullo schema di decreto recante la "regolamentazioneper la realizzazione e consegna della Carta dello Studentedenominata 'IoStudio' - 28 maggio 2015. [Online]. https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/4070802

- Parere sullo schema di regolamento predisposto dall'Istitutonazionale per la valutazione del sistema educativo di istruzionee di formazione, concernente le modalità di svolgimento delleprove Invalsi del terzo anno della scuola secondaria di

primogrado - Parere del 15 febbraio 2018, n. 76. [Online]. https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/8081291

- Parere su uno schema di regolamento predisposto dall'Istitutonazionale per la valutazione del sistema educativo di istruzionee di formazione, concernente le modalità di svolgimento delle prove Invalsi dell'ultimo anno della scuola secondaria disecondo grado - 14 febbraio 2019, n. 44. [Online]. https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9102421

- Provvedimento recante le prescrizioni relative al trattamento dicategorie particolari di dati, ai sensi dell'art. 21, comma 1 deld.lgs. 10 agosto 2018, n. 101 provv. n. 146 del 5 giugno 2019. [Online]. https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9124510

- Provvedimento 5 marzo 2020, n. 45 (doc web 9365147) relativoalla messa disposizione sul registro elettronico di informazionianche relative alla salute di un alunno. [Online]. https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9365147

- Provvedimento del 26 marzo 2020 - "Didattica a distanza: primeindicazioni" del 26 marzo 2020. [Online]. https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9300784

- Ordinanza ingiunzione nei confronti di Istituto Comprensivo diUggiano La Chiesa relativo alla affissione sul portone d'ingressodi un istituto scolastico di dati personali relativi a minori - Provv.2 luglio 2020, n. 117. [Online]. https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9445324

- Ordinanza ingiunzione nei confronti di Istituto ComprensivoStatale Crucoli Torretta riguardante la pubblicazione sul sitoweb di un istituto scolastico di una graduatoria di studentirecanti dati personali anche relativi alla disabilità - Provv. 9 luglio2020, n. 140. [Online]. https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9451734

- Ordinanza ingiunzione nei confronti di Ministero dell'Istruzione,Ufficio Scolastico Regionale per il Lazio relativo allacomunicazione di dati personali anche relativi alla salute di unalunno - Provv. 25 febbraio 2021 n. 67. [Online]. https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9565218

- Ordinanza ingiunzione nei confronti di Università Commerciale“Luigi Bocconi” di Milano - Provv. 16 settembre 2021, n. 317. [Online]. https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9703988

- Parere sullo schema di decreto del Ministro dell'Istruzionerecante la Definizione dei criteri e delle modalità di realizzazionee distribuzione della Carta dello Studente denominata “IoStudio”- 30 settembre 2021. [Online]. https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9713787

- Parere su uno schema di decreto del Ministro dell'Istruzionerecante il regolamento sulle modalità di attuazione efunzionamento dell'Anagrafe Nazionale dell'Istruzione (ANIST) -24 marzo 2022. [Online]. https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9767057

- Parere sullo schema di decreto del Ministro dell'università e della ricerca, previsto dall'art. 62-quinquies, concernente l'Anagrafe nazionale dell'istruzione superiore - 21 luglio 2022. [Online]. https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9806759

- Ordinanza ingiunzione nei confronti di Direzione Didattica Statale 1° Circolo - Eboli relativo alla comunicazione di datipersonali anche relativi alla disabilità di taluni alunni - Provv. n.148 del 28 aprile 2022. [Online]. https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9777156

- Provvedimento n. 185 del 12 maggio 2022 relativo allavalutazione d'impatto sulla protezione dei dati riguardantel'iniziativa “Io Studio - Carta dello Studente” trasmessa dalMinistero, ai sensi dell'art. 11 del d.m. del Ministerodell'Istruzione del 30 settembre 2021 e degli artt. 36, par. 5 e 58,par. 3, lett. c) del Regolamento. [Online]. https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9782000

- Ordinanza ingiunzione nei confronti di Conservatorio di MusicaS. Cecilia di Roma Provvedimento n. 367 del 10 novembre 2022 relativo all'illecito trattamento di dati personali di studenti contenuti in un dispositivo USB contenente la registrazione audio/video di un'assemblea studentesca e all'utilizzo degli stessi per avviare un procedimento disciplinare nei confronti di uno studente. [Online]. https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9835095

- Ordinanza ingiunzione nei confronti di Scuola Statale Secondaria di primo grado "Bianco-Pascoli", di Fasano (BR) Provvedimento n. 421 del 15 dicembre 2022 relativo alla messa a disposizione sul registro elettronico di un istituto scolastico di dati personali, e dati relativi salute di alunni. [Online]. `https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9852255`

Italian legislation applicable in the field of education:

- Legge 07/08/1990, n. 241 e ss.mm. - Nuove norme in materia di procedimento amministrativo e di diritto di accesso ai documenti amministrativi. [Online]. `https://www.normattiva.it/eli/id/1990/08/18/090G0294/CONSOLIDATED/20240612`

- Legge 05/02/1992, n. 104 - Legge-quadro per l'assistenza,l'integrazione sociale e i diritti delle persone handicappate. [Online]. `https://www.normattiva.it/eli/id/1992/02/17/092G0108/CONSOLIDATED/20240612`

- D.Lgs. 16/04/1994, n. 297 - Approvazione del testo unico delle disposizioni legislative vigenti in materia di istruzione, relative alle scuole di ogni ordine e grado Regolamento recante norme in materia di autonomia delle istituzioni scolastiche, ai sensi dell'art. 21 della Legge 15/03/1997, n. 59. [Online]. `https://www.normattiva.it/eli/id/1994/05/19/094G0291/CONSOLIDATED/20240612`

- D.P.R. 24/06/1998, n. 249 e ss.mm. - Regolamento recante lo Statuto delle studentesse e degli studenti della scuola secondaria. [Online]. `https://www.normattiva.it/eli/id/1998/07/29/098G0305/CONSOLIDATED/20240612`

- D.P.R. 08/03/1999, n. 275 - Regolamento recante norme in materia di autonomia delle istituzioni scolastiche, ai sensi dell'art. 21 della Legge 15/03/1997, n. 59. [Online]. `https://www.normattiva.it/eli/id/1999/08/10/099G0339/CONSOLIDATED/20240612`

- Legge 10/03/2000, n. 62 - Norme per la parità scolastica e disposizioni sul diritto allo studio e all'istruzione. [Online]. `https://www.normattiva.it/eli/id/2000/03/21/000G0099/CONSOLIDATED/20240612`

- Direttiva MPI n. 104 del 30/11/2007 - Linee di indirizzo e chiarimenti interpretativi ed applicativi in ordine alla normativa vigente posta a tutela della privacy con particolare riferimento all'utilizzo di telefoni cellulari o di altri dispositivi

elettronici nelle comunità scolastiche allo scopo di acquisire e/o divulgare immagini, filmati o registrazioni vocali. [Online]. https://archivio.pubblica.istruzione.it/normativa/2007/allegati/dir104_07.pdf

- D.P.R. n. 122 del 22/05/2009 - Regolamento recante coordinamento delle norme vigenti per la valutazione degli alunni e ulteriori modalità applicative in materia. [Online]. https://www.normattiva.it/eli/id/2009/08/19/009G0130/CONSOLIDATED/20240612

- Legge 08/10/2010, n. 170 - Nuove norme in materia di disturbi specifici di apprendimento in ambito scolastico. [Online]. https://www.normattiva.it/eli/id/2010/10/18/010G0192/ORIGINAL

- Decreto Ministeriale MIUR del 12/07/2011 - Linee guida per il diritto allo studio degli alunni e degli studenti con disturbi specifici di apprendimento. [Online]. https://www.miur.gov.it/documents/20182/187572/Linee+guida+per+il+diritto+allo+studio+degli+alunni+e+degli+studenti+con+disturbi+specifici+di+apprendimento.pdf/663faecd-cd6a-4fe0-84f8-6e716b45b37e?version=1.0&t=1495447020459

- Direttiva del Ministero dell'Istruzione, dell'università e dellaricerca del 27 dicembre 2012 "Strumenti d'intervento per alunni con bisogni educativi speciali e organizzazione territoriale per l'inclusione scolastica". [Online]. https://www.miur.gov.it/documents/20182/0/Direttiva+Ministeriale+27+Dicembre+2012.pdf/e1ee3673-cf97-441c-b14d-7ae5f386c78c?version=1.1&t=1496144766837

- D.Lgs. 14/03/2013, n. 33 e ss.mm. - Riordino della disciplina riguardante gli obblighi di pubblicità, trasparenza e diffusione di informazioni da parte delle pubbliche amministrazioni. [Online]. https://www.normattiva.it/eli/id/2013/04/05/13G00076/CONSOLIDATED/20240612

- Legge 13/07/2015, n. 107 - Riforma del sistema nazionale diistruzione e formazione e delega per il riordino delle disposizioni legislative vigenti. [Online]. https://www.normattiva.it/eli/id/2015/07/15/15G00122/CONSOLIDATED/20240612

- Legge 07/08/2015, n. 124 - Deleghe al Governo in materia di riorganizzazione delle amministrazioni pubbliche (cosiddetta Riforma della pubblica amministrazione - ddl Madia). [Online]. https://www.normattiva.it/eli/id/2015/08/13/15G00138/CONSOLIDATED/20240612

- Linee di orientamento per azioni di prevenzione e contrasto al bullismo e al cyberbullismo - MIUR, del 13/04/2015). [Online]. https://www.istruzione.it/allegati/2015/2015_04_13_16_39_29.pdf

- D.Lgs. 13/04/2017, n. 62 - Norme in materia di valutazione ecertificazione delle competenze nel primo ciclo ed esami di Stato, a norma dell'articolo 1, commi 180 e 181, lettera i), della legge 13 luglio 2015, n. 107. [Online]. https://www.normattiva.it/eli/id/2017/05/16/17G00070/CONSOLIDATED/20240612

- D.Lgs. 13/04/2017, n. 66 - Norme per la promozione dell'inclusione scolastica degli studenti con disabilità, a norma dell'articolo 1, commi 180 e 181, lettera c), della legge 13 luglio 2015, n. 107. [Online]. https://www.normattiva.it/eli/id/2017/05/16/17G00074/CONSOLIDATED/20240612

- D.Lgs. 07/08/2019, n. 96 - Disposizioni integrative e correttive al decreto legislativo 13 aprile 2017, n. 66, recante: "Norme per la promozione dell'inclusione scolastica degli studenti con disabilità, a norma dell'articolo 1, commi 180 e 181, lettera c), della legge 13 luglio 2015, n. 107". [Online]. https://www.normattiva.it/eli/id/2019/08/28/19G00107/ORIGINAL

- Decreto del Ministero dell'Istruzione di concerto con il Ministro dell'Economia e delle finanze n. 182 del 29 dicembre 2020 "Adozione del modello nazionale di piano educativoindividualizzato e delle correlate linee guida, nonché modalità di assegnazione delle misure di sostegno agli alunni con disabilità, ai sensi dell'articolo 7, comma 2-ter del decreto legislativo 13 aprile 2017, n. 66" e correlate "Linee Guida concernenti la definizione delle modalità, anche tenuto conto dell'accertamento di cui all'articolo 4 della legge 5 febbraio 1992,n. 104, per l'assegnazione delle misure di sostegno di cui all'articolo 7 del D.Lgs 66/2017 e il modello di PEI, da adottare da parte delle istituzioni scolastiche". [Online]. https://www.istruzione.it/inclusione-e-nuovo-pei/allegati/m_pi.AOOGABMI.Registro%20Decreti(R).0000182.29-12-2020.pdf

- "Linee guida per la redazione della certificazione di disabilità in età evolutiva ai fini dell'inclusione scolastica e del profilo di funzionamento" del 14 settembre 2022 del Ministro della Salute di concerto con il Ministro dell'Istruzione, il Ministro del Lavoro e delle Politiche Sociali, il Ministro dell'Economia e delle Finanze, il Ministro per gli Affari regionali e le Autonomie, il Ministro per le Disabilità. [Online]. https://www.salute.gov.it/imgs/C_17_pubblicazioni_3276_allegato.pdf

- Nota del Ministero dell'Istruzione n. 9168 del 9 giugno 2020 "Ulteriori precisazioni e chiarimenti sull'applicazione dell'O.M. n.11 del 16 maggio 2020 concernente la "valutazione finale degli alunni per l'anno scolastico 2019/2020 e prime disposizioni per il recupero degli apprendimenti". [Online]. https://www.miur.gov.it/

- Nota del Ministero dell'Istruzione n. 13914 del 11 giugno 2021 "Pubblicazione online degli esiti degli scrutini - indicazioni operative". [Online]. `https://www.istruzione.it/archivio-usr-molise/images/m_pi.AOODGOSV.REGISTRO_UFFICIALEU.0013914.11-06-2021.pdf`

- Nota del Ministero dell'Istruzione e del merito - Dipartimento per il sistema educativo di istruzione e formazione - Direzione generale per gli ordinamenti scolastici, la valutazione e l'internazionalizzazione del sistema nazionale di istruzione del 30 novembre 2022 relativa alle Iscrizioni alle scuole dell'infanzia e alle scuole di ogni ordine e grado per l'anno scolastico 2023/2024. [Online]. `https://www.miur.gov.it/documents/20182/6739250/Nota+iscrizioni+anno+scolastico+2023-2024.pdf/d3fcc568-5fdd-fb3a-a3ee-ec8e36798234?t=1669917232803`

- Nota del Ministro dell'Istruzione e del merito del 19 dicembre 2022 avente ad oggetto "Indicazioni sull'utilizzo dei telefoni cellulari e analoghi dispositivi elettronici in classe". [Online]. https://www.miur.gov.it/