Johanna Johansen

Towards Making Privacy Usable

Thesis submitted for the degree of Philosophiae Doctor

Department of Informatics
The Faculty of Mathematics and Natural Sciences

2022
To my family that has always been there caring for me – Cristi, Elena, Silviu, and Meluzin – and to the people that have positively influenced my life and have been an inspiration for me – Olaf Owe, Hanne Baadsgaard Utigard, Tore Pedersen, and Simone Fischer-Hübner.
Preface

This thesis is submitted in partial fulfillment of the requirements for the degree of Philosophiae Doctor at the University of Oslo. The research presented here was conducted at the University of Oslo, part of the Gemini Centre on IoT, which was established as a collaboration between the University of Oslo, the Norwegian University of Science and Technology (NTNU), and the research institute SINTEF. My main supervisor has been Olaf Owe (professor at University of Oslo), and co-supervisors have been Tore Pedersen (professor at Bjørknes Høyskole), Simone Fischer-Hübner (professor at Karlstad University, Sweden), Frank Alexander Kraemer (associate professor at Norwegian University of Science and Technology), and Amela Karahasanovic (senior researcher at SINTEF). The project IoTSec – Security in IoT for Smart Grids – has supported my research stay at Karlstad University.
Preface

Acknowledgements

I would like to thank my supervisors for being true role models for me as early-stage researcher, for their good guidance and for sharing with me their knowledge, experience, and time (which I learned is very scarce and precious in academia). The meetings with you have been something to look forward too. I would particularly like to thank my main supervisor Olaf for his support, patience and understanding, and for being always available when I needed advice and help with both research as well as more practical administrative matters. I would also like to thank my co-supervisor Tore for introducing me to the field of Psychology, and for being an inspiration in many ways, especially with his true passion for research, his modesty, and his fantastic way of collaborating with people, bringing forth the best in them, and making them feel valued and appreciated for their efforts. Last, but not least, I would like to thank co-supervisor Simone, which is an example of what I would like to achieve in my career as a researcher, and especially as a woman researcher. Because of Simone I was given the possibility to visit her group in Karlstad and take part in the Privacy & Us Network. In these very positive, inclusive, and friendly environments I met highly skilled people, with knowledge and experience that were very relevant for my research. Simone has been a door-opener for my research, guiding me towards the right venues, topics and people.

I am also grateful to the Faculty of Mathematics and Natural Sciences and the Norwegian Research Council for funding my PhD.

Johanna Johansen
Oslo, February 2022
List of Papers

**Paper I**: “A Multidisciplinary Definition of Privacy Labels”

is conditionally accepted, subject to revisions, for the journal *Information and Computer Security*,


**Paper II**: “Making GDPR Usable: A Model to Support Usability Evaluations of Privacy”


A longer version supporting this paper was published on the ArXiv (arXiv:1908.03503), with the same title.

**Authors**: Johanna Johansen and Simone Fischer-Hübner

This work was presented in several places:

- at the *16th Symposium on Usable Privacy and Security (SOUPS 2020)* as a lightning talk in August 2020;
- at the *Gemini Centre on IoT annual PhD Seminar* in May 2020;
- at the *IFIP Summer School on Privacy and Identity Management* in August 2019;
- at the *Open Day for Privacy, Usability, and Transparency Workshop (PUT 2019)* part of the *19th Privacy Enhancing Technologies Symposium (PETS)* as a Poster presentation in July 2019;
List of Papers

**Paper III**: “Expert opinions on making GDPR usable”

is in preparation to be submitted,
with a longer version published as Technical Report nr. 501 in the series with ISSN 0806-3036 managed by the University of Oslo Library.

*Authors*: Johanna Johansen.

**Paper IV**: “Revealing Human-to-Computer Bias Transference”

to appear in Springer’s journal *AI & Society: Knowledge, Culture and Communication*,


*Authors*: Johanna Johansen, Tore Pedersen, and Christian Johansen.
# Contents

Preface iii

List of Papers v

Contents vii

I Overview 1

1 Introduction 3

1.1 Motivation 4
1.2 Research Goals 6
1.3 Research Methods 7
1.4 Structure of the Thesis 9

2 Background 11

2.1 Privacy Labels 11
2.2 Usability and Privacy Certifications 14
2.3 Privacy Biases and their Transference 16

3 Contributions 19

3.1 A Definition of Privacy Labels 20
3.2 Support for Usability Evaluations of Privacy 22
3.3 Validations with Experts 23
3.4 A Method for Revealing the Transference of Privacy Biases 24

4 Overview of the Research Papers 27

4.1 Paper I: A Multidisciplinary Definition of Privacy Labels: The Story of Princess Privacy and the Seven Helpers 27
4.2 Paper II: Making GDPR Usable: A Model to Support Usability Evaluations of Privacy 27
4.3 Paper III: Expert opinions on making GDPR usable 28
4.4 Paper IV: Revealing Human-to-Computer Bias Transference 28

5 Conclusion 31

5.1 Summary of Contributions 31
5.2 Future Work 33

Bibliography 35
## Contents

<table>
<thead>
<tr>
<th>Papers</th>
<th>44</th>
</tr>
</thead>
<tbody>
<tr>
<td>I A Multidisciplinary Definition of Privacy Labels</td>
<td>45</td>
</tr>
<tr>
<td>II Making GDPR Usable: A Model to Support Usability Evaluations of Privacy</td>
<td>93</td>
</tr>
<tr>
<td>III Expert opinions on making GDPR usable</td>
<td>155</td>
</tr>
<tr>
<td>IV Revealing Human-to-Computer Bias Transference</td>
<td>201</td>
</tr>
</tbody>
</table>
Part I

Overview
Chapter 1

Introduction

This work has the general goal of making privacy easier to understand and work with for all entities involved, considering the context of the modern computerized society. Special consideration is given to the lay people as being the subject of private information collection, but who, at the same time, have privacy as a basic right. Since privacy is a complex concept that is exacerbated by the complexity of the technology that is using private data, it is paramount to make privacy more usable, e.g., to make the laws and regulations regarding privacy easier to understand by lay people, so to know their rights better; or to make the information about the privacy protection that a product offers more easy to evaluate, so to be able to make better informed decisions when buying a new device or registering for an online service.

A second category of stakeholders that this work is concerned with are those producing the multitude of software products and services that use private information. The programmers, e.g., need also to understand privacy regulations such as the GDPR, in order to be able to implement their software systems so to respect these, and be certified accordingly. At the same time, lack of education and knowledge about privacy among these stakeholders gives way to preconceptions and biases regarding privacy (e.g., considering privacy only a matter of security) which can then be reproduced in their software artifacts.

More generally, the industry actors need to put more effort into making the privacy of their products more usable, and these endeavours should also be rewarded, not only in the long run through more satisfied clients, but also in the short run through being certified by accredited certification bodies that consider the usability of privacy when certifying IT products or services. Therefore, we need to make the idea of usable privacy manageable (and measureable) also for the certification experts so that they have models and processes in place to help them evaluate the usability of privacy. The same information would then be used by the industry as a guide for making the privacy of their products more usable.

In this work we bring concepts and ideas from the field of Human-Computer Interaction (HCI) that has long been developing methods for making complex systems (e.g., airplane controls or user interfaces) easier to use for the respective target users (e.g., pilots, who are highly specialized experts, but also common people that use a software system through a user interface). We argue in this thesis that the same principles from HCI can be used to make the complexity of privacy more usable, i.e., easier to understand, to argue about, to control, and to use in people’s benefit. More motivation for this work is given in Section 1.1, whereas Chapter 2 provides more background to help understand better the goals of this work, which are listed in Section 1.2.
1. Introduction

The contributions of this thesis are thus meant to push the research community into the direction outlined above. Chapter 3 details these contributions, which have been published in the papers that are included in this thesis, and which are summarized in Chapter 4.

1.1 Motivation

Privacy is a human right and an essential prerequisite for protecting fundamental human values, such as dignity and autonomy. However, the goal of protecting the privacy of the individuals and public has proved to be challenging to achieve. Currently, people have difficulty in understanding how their privacy is affected by the current practices in the technological world and are often unaware of its value and the implications of its loss.

In its current state, privacy does not always serve the people, but mainly a few very wealthy and influential actors prosper from its misuse. Loss of privacy has both micro implications, at a personal level, e.g., people being influenced to buy what they do not want or need, to vote for extremists, or to develop antisocial behavior), but also macro implications, at a societal level (e.g., a society living in fear of being watched by surveillance capitalists or manipulated on social media. The lack of privacy literacy can be partly attributed to commercial entities that often, while profiting from handling data, work hard to keep privacy “out-of-sight”– like a sleeping princess locked in a tower – e.g., telling people “You have zero privacy anyway. Get over it.”

In the current digital society, privacy has even stronger forces compounding its complexity, coming from, e.g., technological advances in miniaturizing hardware that enabled cheap privacy-invasive gadgets, powerful algorithms that can make inconceivable inferences, or supercomputing in ‘invisible’ clouds; all too complex for laypeople to grasp.

The complexity of the privacy concept as such and of digital data and technology, makes it difficult for one to evaluate the privacy properties of a specific piece of technology (e.g., web service, Internet of Things (IoT) product, or communication device). The difficulty is not only for average people, but also for regulators to check compliance, and for developers to be able to provide privacy-aware digital services/products/systems. Indeed, there are multiple concepts

1 We will be liberal in our use of the word “privacy”, including the concept of personal (digital) data protection as used by GDP16; however, we are aware that “privacy” has different formulations and scopes in different European or international laws Lyn14.

2 The “right to privacy” emerged in the Universal Declaration of Human Rights, adopted in 1948, as one of the fundamental human rights. Shortly after, this right was reaffirmed in the European Convention on Human Rights (ECHR), drafted in 1950.

3 https://www.wired.com/1999/01/sun-on-privacy-get-over-it/

4 Note that system/product/service are used interchangeably throughout the thesis.
Motivation

involved in digital privacy, like data sharing (which for normal business practices nowadays can form a highly intricate network of relationships), ownership and control of data, accountability or transparency (both towards the regulators as well as the users). Many of the privacy concepts are even a challenge by themselves when it comes to their evaluation, since they are difficult to measure or to present/explain. Our contribution in Section \ref{sec:3.2} proposes a way to measure the usability of privacy, to make it easy for the users to understand how well their privacy is respected by a respective company or a product. Furthermore, in Section \ref{sec:3.3} we also validate the concepts and methods that we introduce in the \cite{JF20}.

To solve the intricacies of privacy around the use of personal data in technologies, collaboration between specialists from different fields – such as computer science, interaction design, human ergonomics, law, and cognitive science – is needed. When a common understanding has been reached this should be translated into a form that can be grasped by regular people with limited time and level of expertise in fields such as technology or law. The concept of Privacy Labels is our contribution in this direction, detailed in Section \ref{sec:3.1}. Another useful work is \cite{Nis+18}, which is meant to bridge the computer science and legal approaches to privacy.\footnote{For this work the authors have received the 2019 PET Award for Outstanding Research in Privacy Enhancing Technologies.}

In this thesis we adopt a multidisciplinary approach to understand not only the needs and goals of laypeople when it comes to privacy, but also of other stakeholders involved such as businesses, programmers, or lawyers. Being privacy aware could be a competitive advantage \cite{MM17}, especially in markets such as Europe, where privacy protections are required by law. Even in privacy unregulated markets, a company that has visible and easily understandable statements of compliance will be perceived by the customers as one that takes privacy seriously. This has the potential to increase the customer’s trust in the company, which is seen in \cite{BLT20} as essential for data-driven businesses.

At the same time, one of the fastest spreading types of data-driven software is the AI/Deep learning based software (see the 2019 Turing award laureates excellent overview \cite{LBH15}), with a considerable number of programmers actively involved. In his recent call for AI regulations, Etzioni urges regulators to focus on five critical areas “no killing, responsibility, transparency, privacy, and bias” \cite{Etz18}. Privacy has earned a forth place on this list of concerns for AI software because AI is data-hungry and much of this data will presumably come from IoT systems close to humans – of course, for those AI applications that are interacting in some form with people and the society at large (e.g., in decision support or smart systems). Privacy labels could help in these regulating endeavors as well, this time not addressed only to the convenience users but more to the businesses, e.g., for allowing well informed AI software purchases, as well as to AI engineers, e.g., to guide their choices of libraries and software components.

Privacy has long been an important concern for software developers, e.g., the (then) President of the ACM, David Parrerson in \cite{Pat05} put forward the “SPUR
manifesto” which was placing (P)rivacy as one of the four main focus areas for software engineering, along with (S)ecurity, (U)sability, and (R)eliability. On the contrary, biases in software (chiefly in AI-based decision systems, as Etzioni describes) is a rather new concept for software developers.

Biases in AI systems (the fifth area of concern for Etzioni) normally come from improper use of training data, i.e., the system is trained on a data set that is not representative of the population/problem that it is applied to (or does predictions about). AI biases can be about gender, race, or other social aspects, but also about privacy. This last form of bias, which we call privacy biases, is largely not investigated because it is not seen as a machine bias, i.e., it does not appear from data or the software code. Privacy biases are human biases, in line with the traditional bias mechanisms studied in psychology – see also a nice account of how cognitive and behavioral biasing mechanisms (such as the anchoring heuristic or framing effect) influence privacy behaviors in [Acq+17, Sec.2.3]. Quite a number of privacy biases could fall in the class that we would call I-have-nothing-to-hide, with a large collection of such privacy attitudes nicely presented in [Sol11]. A privacy bias that programmers often fall prey to can be called privacy=security. Our contribution detailed in Section 3.4 shows that human biases can be transferred from the programmer into the software that she is building.

Bias transference is thus an additional mechanism to the standard one studied in AI biases, through which human biases can manifest into the software that we build. Therefore, privacy biases are elevated to being a serious threat to the software that programmers develop as it can incorporate the privacy biases of their creators. Privacy biases have as a root cause the lack of adequate knowledge, either that the person is time constrained and cannot gather or infer the needed knowledge for the decision task at hand, or that simply the person is inexperienced for the new task. Programmers often find themselves in such uncertainty situations, e.g., when faced with incomplete specifications or vague requirements. This is even more so in the case of understanding privacy policies. As a result, programmers are mostly left to their own means and judgment when implementing privacy features or requirements; and any privacy bias or neglect can reflect on the users of the resulting software. This happens because of the transfer of the programmers privacy views and biases into the software artifact when privacy aspects are not easy to comprehend.

1.2 Research Goals

As already stated above the overall goal is to make privacy more usable for lay people, in the sense that laws and regulations regarding privacy, as implemented by organizations, to be easier to understand, so that lay people can better know their rights, and so that it is easier to evaluate the privacy protection of an digital service or product. This will help end users to make better informed decisions when buying a new computerized device or registering for an online service. In addition, the thesis has three main subordinate goals:
1. The first goal is to define and analyze how the concept of Privacy Labeling/Labels, which hereafter we refer to as PL, could contribute to resolving several of the challenges that privacy is currently facing. Given the many facets of privacy and its society-wide implications, it is important to adopt a multidisciplinary approach. Therefore we brought together experts from different fields representing the stakeholders of PL to express their perspectives, needs, and goals when it comes to how PL should be.

2. The second goal focuses on one challenge, where we want to show how to achieve the usability goals that appear in privacy regulations, by including usability evaluations in privacy certification schemes. In addition to giving a solution to how one could evaluate and measure usability of privacy we also wanted to validate our approach with experts within the field of usable privacy, data protection law and certifications to find out if this is a viable solution for what it is needed for both practice and research communities, and how it could be improved or extended through further work.

3. The third goal studies a second challenge regarding the implementation of privacy aspects, where privacy biases of the developers can be a hinder to a good alignment between privacy regulations and agreements on one hand and software implementations on another hand. However, the topic of biases in software as coming from human programmers is little investigated in comparison with algorithmic bias. As such, our goal was to first find if bias transference exists at all, as an additional mechanism to the standard one studied in AI, before looking specifically for privacy biases.

These goals have driven the research behind this thesis and related results are published in dedicated papers as follows:

- **Paper I** deals with goal 1.
- **Paper II** and **Paper III** deal with goal 2 whereas
- **Paper IV** has laid the foundations for studying in more detail goal 3.

More details are given in Chapter 3 as well as throughout this overview chapter.

### 1.3 Research Methods

For the work in this thesis we had the goal to bring to discussion and represent in our research the different stakeholders of data protection and privacy in general, as denoting different domains of practice and research. With this purpose, we conducted empirical studies with the respective stakeholders, where we employed both qualitative and quantitative methods of research.
1. Introduction

For the work in [Paper I] in order to capture the multidisciplinary nature of privacy, we organized a workshop where experts from different fields of practice and research gathered to present and discuss their views on the topic of Privacy Labels. We have employed a focus groups method in order to allow the participants to interact, ask questions, challenge, or (dis)agree with each other. We thus brought in the following perspectives:

- Business (relevant topics including, e.g., market potential, incentives, social responsibility, added value);
- Law (e.g., compliance, privacy policies, audit);
- Regulations (e.g., national, European, implementations, domain-specific standards);
- Usability and human factors (e.g., personas, easy to understand, completeness, contextual);
- Education (e.g., psychology of people, of SMEs (Small and Medium-sized Enterprises), of CEOs (Chief Executive Officers), nudging for good, mental heuristics);
- Technology (e.g., AI, reasoning, automation, dynamic labels, verification);
- Multidisciplinary (e.g., communication across fields, people, or companies, making synergies).

The form of the workshop was thought to allow the experts in the respective fields to first present their view on the topic separately, and then discuss and collaborate with the other participants with the goal to ultimately settle on a common perspective. During the workshop, as reflected in the paper, we deliberate on how to combine the seven different perspectives, the roles and priorities of each of these in relation to PL, and point to the state of affairs in the respective fields. The discussion between the participants, as well their contribution to the topic continued, in an asynchronous way, while working with presenting the results from the workshop.

[Paper III] conducts a validation study for the concepts and methods introduced in [Paper II]. Building on our methodological goal to offer a multidisciplinary perspective on privacy, we interview experts working across fields of relevance to these concepts, including law and data protection/privacy, certifications and standardization, and usability (as studied in the field of Human-Computer Interaction). The study uses a critical qualitative research [BC13], where we take an interrogative stance towards the meanings and experiences expressed in the data we collect through interviews. We are also interested in how the individual meanings reflect how usability is understood by the broader communities that the participant’s expertise are representing. To achieve this we involve the three different theoretical perspectives in a “theory triangulation” manner [Pat99]. Special in our case is that the participants are not brought to discuss the data, but to discuss the theories and concepts introduced in [Paper II].
A slightly different methodological approach was taken in Paper IV where we single out one stakeholder group to focus on, i.e., the software programmers. In the first papers we conducted descriptive and relational types of empirical investigation, while in Paper IV, in order to determine the causal effect between bias in computer programs and bias in software programmers, we adopt an experimental research methodology. Every experiment starts with a hypothesis to be tested, which in our case is whether biases can be transferred from the programmer to the program. This hypothesis is tested with three experimental cohorts, created based on the educational and occupational background of the participants. In the analyses of both (i) the comparative aspect (i.e., differences between the three cohorts) and (ii) the experimental aspect (i.e., differences within each cohort), we employed both (a) inferential statistics, more specifically chi-square analyses of categorical data, as well as (b) descriptive statistics to report frequencies and percentages. We performed an experiment on each cohort, as well as compared the three cohorts to each other, regardless of the experimental manipulation.

In summary, we have made use of empirical studies when working towards the three goals of this thesis. However, the specific methods used were chosen to meet the needs of the goals being studied, ranging from a multidisciplinary approach, with participants from seven different fields, towards a more focused approach, where we singled out one stakeholder group to perform quantitative studies with ca. 300 participants.

1.4 Structure of the Thesis

This thesis document is structured in two parts, the first part contains an overview of the research, whereas the second part includes the published research papers. The overview part starts by giving a general introduction for the work, including motivations and presenting the goals of this research. We then set some background knowledge, needed to understand better the problems and follow easier the details of the contributions. The background includes aspects relating to Privacy Labels, usability evaluations and certifications of privacy, as well as concepts from psychology pertaining to biases and how these are formed and transferred between humans. The second part lists four research papers; for the first three we include both a short version, peer-reviewed or under review for a journal, and a longer version published as technical reports on ArXiv.
Chapter 2

Background

Before presenting in Chapter 3 the contributions that this thesis makes towards the goals that have already been outlined and motivated in the beginning of Chapter 1, we present here background on the main concepts used throughout this thesis. The purpose is to prepare the reader for the presentation of the contributions, thus making these easier to grasp. We structure this section following our goals and contributions, i.e.:

• We first talk about concepts related to Privacy Labels, which is our multidisciplinary definition described in the contributions Section 3.1. This is a rather broad topic, thus necessitating knowledge from several areas, including business, psychology, software engineering, and privacy.

• We then give relevant background for usable privacy certifications, which is our contribution detailed in Sections 3.2 and 3.3. Here we introduce more concepts coming from the diverse fields of Human-Computer Interaction, certifications and standardization, and privacy regulations.

• We end this section with terminology necessary for our third line of contributions, detailed in Section 3.4. For this we need more information from psychology, particularly concerning biases and heuristics, but also regarding AI and algorithmic biases, which is where this work has immediate impact, with privacy biases being only one form of bias that this work covers.

2.1 Privacy Labels

When it comes to privacy we are faced with a general tendency where people often confuse privacy with security, and in our papers we want to make clear the distinction between the two, even though they are interrelated, to avoid confusion, or to avoid affirmations where one might think that privacy aspects are been addressed by default when tackling security concerns only. Although multiple standards and certifications currently do not make such clear distinctions (and are looking mostly at security), security should only be a baseline, e.g., The General Data Protection Regulation (GDPR) considers security as one of its several data protection principles (see Art. 5 I (f)). Privacy protection goals include the

1GDPR – General Data Protection Regulation from European Union [GDP16] was adopted in April 2016, replacing the Data Protection Directive, as a way to modernize the European data protection legislation, making it more adapted to the new developments in the digital economy and society. GDPR is also meant, as an improvement from the Directive, to provide a single set of data protection rules, with similar levels of enforcement, severity of sanctions, and interpretations across Europe.
2. Background

classical security protection goals confidentiality, integrity and availability (CIA),
and in addition also privacy goals such as transparency [MF17], intervenability
and unlinkability that go beyond CIA [LJR15]. Moreover, when looking at the
attitudes of the users there are clear differences between security and privacy,
due to individual differences [EP15].

If in security it is often said that the weakest link is the user, in privacy we see
that the weakest link is the controller. Examples of privacy breaches for which
the controller is responsible can be: the controller “tricks” the users into giving
more data than the user is aware of, often through hiding information or by
using privacy-invasive approaches known as “dark patterns” [Bös+16; Mat+19;
Nou+20]; lack of legal competence when drawing contracts with third parties;
programming incompetence incurring leakage of data, e.g., usage of third party
libraries; or not investing in measures for preventing security leakages [Pal+20].

The Recital 100 of GDPR encourages “the establishment of certifications
mechanisms and data protection seals and marks [to allow] data subjects to
quickly assess the level of data protection of relevant products and services”
[GDP16]. While Art. 42(1) encourages the implementation of certification and
data protection seals for demonstrating compliance by accredited certification
bodies, PL should go beyond and measure on a scale how well the privacy is
respected and how easy is for a user to understand that. Compared, e.g., with
energy consumption where one number is enough, maybe placed on a graded
scale for visual comparison, and measured with well known instruments, privacy
involves multiple concepts spanning from social to technological, most of which
we do not know how to measure, definitely not in a universal way since they
are relative to a person’s view on privacy influenced by this person’s level of
knowledge about privacy.

Examples of existing certifications include: Datenschutzgutesiegel, granted to
systems and products by ULD (The Schleswig-Holstein Data Protection Author-
ity) [2], EuroPriSe [3], Common Criteria (ISO/IEC 15408) including a Privacy Class
meant for defining privacy functionality, focusing on aspects such as anonymity
[Ell+18], pseudonymity, unlinkability [MAS18], unobservability [PK01]. Some of
these are partly required by law, e.g., ULD Datenschutzgutesiegel is used for
public procurement in the Schleswig-Holstein German state, whereas Common
Criteria are taken up in eIDAS (electronic IDentification, Authentication and
trust Services) EU Regulation No 910/2014, and partly required for certain
procurements in certain public sectors.

For privacy certifications we also need evaluations. The challenge is that the
focus of schemes such as the above is much on security testing and penetration
testing. There is a need for more formal evaluation, verification, or testing
of privacy requirements. Aspects of usability should also be included in the
evaluation of privacy. Parallel this with security where the weakest link is often
the end-user. Nowadays communication protocols are formally proven secure, but

https://www.commoncriteriaportal.org/
still security breaches occur because the user interfaces are not usable, leading end-users to doing mistakes, e.g., the security warnings for SSL certificates or other types of security warnings where the end-users have to make decisions without good guidance or usable instructions, e.g., Whitten and Tygar already in 1999 tested Pretty Good Privacy and revealed several usability issues that lead to insecure decisions or that the encryption products/features were not used at all \cite{WT99}.

In addition, privacy (like security) is usually only a secondary task for the users \cite{WT99}, e.g., when buying train tickets with a ticket-app the primary goal of the user is not to check how well the app protects her privacy, but to reach a certain destination. In addition, it is arduous for a regular person to keep track of all the electronic data that she is generating, given that many activities nowadays are happening online. It is even more difficult to know exactly which effect this data has on our privacy, because of the modern machine learning algorithms that can make inferences based of apparently non-private pieces of data \cite{Acq17,RHM20}. To overcome such user/usability related challenges, one has to make the privacy related measures usable. The GDPR is a good place for finding examples of usability goals, e.g., “communication ... relating to processing [to be provided] to the data subject in a concise, transparent, intelligible and easily accessible form, using clear and plain language” (Art. 12); with 30 more such usable privacy goals identified in \cite{PaperII}. However, the usability goals appearing in GDPR are too general, given the inherent nature of the GDPR (and laws in general), which in this case allow for too much subjective interpretation by controllers for their own interest. One classical behavior is to make the privacy settings blend into the background, or even worse, the button for the privacy invasive option is highlighted by design; e.g., when emphasizing the ‘Accept’ button for the privacy policy, the users will give their consent without reading. This is the case with the privacy policies as well, which can containing information manipulated in a way that the user on average will not be well informed, though they formally meet the requirements of the GDPR \cite{KPF20,MC08}.

Usability and Human-Computer Interaction (HCI) techniques have mostly been developed for making technologies that are difficult to use, or made for highly specialized experts, more easy to understand and interact with, both for the expert user, but often also for novice users. Usability is even more important for privacy since privacy is a highly complex concept, related to complex technologies such as AI and Big Data, but which is especially addressed to the laypeople, not to experts, because privacy is a human right. As shown by \cite{Pat03,PK03}, it can however be challenging to map legal requirements into HCI requirements. For the programmers as well it can be difficult to implement the statements made in the privacy policies or regulations. Therefore, help is needed in bridging the gap between regulations or legal documents (such as privacy policies) and the software/technology that these talk about. Those needing support in this case being the lawyers, interaction designers, and the programmers.

One starting point for evolving certification schemes from seals and trust
marks towards privacy labels of the energy efficiency type is by measuring the usability of privacy using HCI methods. Paper II works in this direction by proposing a set of criteria thought to produce measurable evaluations of the effectiveness, efficiency, and satisfaction with which privacy goals of GDPR are reached. More details about how we propose usability to be measured in order to be later translated in scales is further discussed in 3.2.

2.2 Usability and Privacy Certifications

The present work can be placed in the research field called usable privacy and security, with seminal works s.a. [AS99; Cra18; GK03; WT99] and conference series s.a. the Symposium On Usable Privacy and Security (SOUPS). As we previously explained in Section 2.1 we consider that research on privacy requires, an interdisciplinary approach (encompassing the expertise coming from research fields such as Psychology, Law or Human-Computer Interaction). Therefore, in both papers Paper II and Paper III we continue to favor an interdisciplinary approach to privacy, combining law, technology, and usability.

There have been considerable efforts towards including specialist from different areas of research on issues related to privacy. Examples of such efforts are the constitution of The Privacy & Us Innovative Training Network (ITN) or the organization of the IFIP Summer School on Privacy and Identity Management.

Other examples of cross-disciplinary research efforts come from the automation of privacy agreements (or Terms of Services – ToS) where the goal is to presented ToS in an accessible way to the general user. Notable contributions in this regard are the endeavors of the LeDA network or the Usable Privacy Policy Project or the CLAUDETTE project.

Similarly to Paper I in the case of the Paper II as well we consider security as one integral aspect of privacy, where privacy implies security but not the other way around. We consider such a clarification necessary, as we have seen a tendency in the general public to equalize the meanings of the two terms in favor of security. In computer science, privacy research has been closely intertwined with security research, reflected e.g. in the contents and the structure of the book [CG05]. However, in this paper, we favor the term ‘usable privacy’, as it includes by default security, which is in accordance with the data protection

---

Footnotes:

5 This project has received funding from the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 675730, within the Marie Skłodowska-Curie Innovative Training Networks (ITN-ETN) framework; https://privacyus.eu/

6 https://www.ifip-summerschool.org/

7 The Legal Design Alliance (LeDA) is formed of lawyers, designers, technologists, academics, and other professionals who are committed to making the legal system more human-centered and effective, through the use of design. https://www.legaldesignalliance.org/


9 CLAUDETTE (automated CLAUse DETectEr), http://claudette.eui.eu/about/index.html. See also their tool http://www.claudette.eu/gdpr/
Usability and Privacy Certifications

legislation, where security (integrity and confidentiality) is specified as one of the several principles to abide by in order to assure the privacy of users’ data.

Having the goal to evaluate the usability of privacy in technological systems and products, makes our work part of the larger HCI research on privacy [AM05 KBK05 KKB12 Pat+03]. Following the classifications made by Iachello and Hong in their review [IH07], we approach privacy from a “data protection” perspective by extracting usability related goals from the GDPR. A similar approach is taken in [Pat+03 PK03], which translates legislative clauses of the Directive 95/46/EC (now replaced by GDPR) into interaction implications and interface specifications. Similarly, [Lan01] develops principles for guiding system design based on fair information practices found in the US Privacy Act of 1974 and the EU Directive 95/46/EC.

The model we propose, detailed in Section 3.2, integrates well with a user-centered design where HCI methods are applied to elicit requirements based on understanding the users, their needs and the context of use. We also base our conceptualization of ‘usable privacy’ on the definition of usability as presented in the ISO 9241-11:2018 [ISO18], which we adapt to include privacy. Traditionally, usability is a quality related to the use of a product. In our case, we are not interested in the usability of a product per se, but only in those aspects of a product that concern privacy. Once privacy measures and privacy enhancing technologies are integrated into the design of a product, it still remains to find out if (and how much or to what extent) those measures empower and respect the rights of their particular user as intended. In Human-Computer Interaction (HCI) this is determined based on user testing and usability evaluations. The criteria we propose presume the use of such established HCI methods for usability evaluations (e.g., [DR99]).

When it comes to the legislation part of our work, we exemplify our theories directly on the GDPR text. Though, there are several legislative texts adopted in different jurisdictions for explaining and regulating privacy rights and obligations, besides research articles and books [Eur18], GDPR in Europe makes a good effort in clarifying many aspects of data privacy, providing the legislative support to enforce better data protection practices on anyone (within its jurisdiction) collecting and processing personal data. However, these regulations, GDPR included, only specify the requirements on the data controllers (i.e., the organizations providing the service) in the form of basic principles, and the rights of the data subjects, but do not make any strict claims about the extent to which a controller (or processor) should go about implementing these requirements so that they are beneficial for the user, and to what degree. As such, the usability goals of GDPR – as we call them and define as Usable Privacy Goals in Paper II, s.a. “... any information ... and communication ... relating to processing [to be provided] to the data subject in a concise, transparent, intelligible and easily accessible form, using clear and plain language, ...” (Article 12 (1) of GDPR), are left open to the subjective interpretation of both evaluators and controllers. The provisions of GDPR regarding usability are too general and high-level to be suitable for a certification process [KD18]. To remedy this, as we describe in more detail in 3.2, we propose a set of criteria thought to
produce measurable evaluations of the usability with which privacy goals of data protection are reached.

Besides GDPR, for evaluating privacy we take as starting point the methodology developed by EuroPriSe [Eur17] that has as purpose to evaluate compliance with GDPR. Certification schemes (s.a. EuroPriSe) provide a seal showing compliance with data protection regulations and industry standards. In addition to such a certification, our evaluation measures on a scale how well data protection obligations are respected and how easy it is for a user to understand that. These measurements can be presented to the user in different ways, e.g., using “traffic light” scales, showing which level of usability has been reached by the privacy of a certain technological product. A “traffic light” presentation of privacy is recommended by [Hou16, Chapter 6:235] as a way to “foster competition” and “show good practice on privacy policies”.

The marketing motivation for the providers to address usability aspects of privacy is that the users tend to choose the product that answers best to their specific and real needs for privacy [Joh+20; RR20]. Only the lack of alternatives in the market today explains why the data subjects still accept detrimental privacy conditions that would rather fit the interest and attitudes of other type of stakeholders (controllers and processors) [Sch15]. The metaphor of the “dancing bear” of Alan Cooper [Coo04] illustrates well this situation.

Creating alternative products or features is a way for the businesses to differentiate themselves. Now a new way of differentiation is the level of privacy protection offered, beyond the minimum required by GDPR, as well as how usable this is. Certification schemes such as EuroPriSe will give a product the seal of GDPR compliance [CC18], but as the GDPR compliance is mandatory, all businesses will seek to conform. Beyond this, considering usability aspects when implementing measures for protecting privacy would facilitate further differentiation between services. Usability is known as a market differentiator, e.g., the ISO 9241-11:2018 standard asserts that designing for usability helps with marketing of a product and with offering the user better customized choices.

### 2.3 Privacy Biases and their Transference

Media and the general public seem to have the assumption that machines and algorithms are in themselves neutral and objective. However, it has been known for quite some time (and recently also came to the attention of several public actors) that complex algorithms, such as those from artificial intelligence among others, may exhibit biases s.a.: racial bias [SOT18], gender discrimination [ZS18] and other socially relevant types of biases [BC12; FN96; JIV19], when processing information in the support of human and institutional decision making [Cor+17; DF18]. This phenomenon is commonly labeled as machine biases or algorithmic bias, and has been confirmed in different areas (published in respective top venues), e.g., in big data [HBC16], web [Bae16; Bae18], autonomous systems [DL17]. Among institutions that have raised concerns about the existence
of “biased algorithms” are: the ACM US Public Policy Council\( ^{10} \) the EU Parliament\( ^{11} \) the New York City Council which passed a bill on “Accountability and transparency in algorithms for public agency support”\( ^{12} \) ERCIM RTG19; World Wide Web Foundation\( ^{13} \) and many more\( ^{14} \). An influential research report Bru+18 has raised even more concerns about harmful algorithms, and has recently been joined by articles in major publication venues such as Science and Nature Gia+18 Obe+19 ZS18 and by scholarly books Bod08 ONe16.

All the works above focus on the data that AI algorithms train on, and show how the data contains biases. We are not aware of works that study empirically the transfer of biases from the human programmer to the algorithm, although we have found related ideas mentioned in recent articles: SK19 describe nine types of biases (present at five different algorithmic stages: input, algorithmic operations, output, users, and feedback), some of which can be studied in conjunction with the general bias transfer that we demonstrate in Paper IV Bae18 brings up the users and producers of the web content as sources of bias related to the data, but also points out different forms of bias coming from the user interface made by interaction designers whom could be regarded as ‘programmers’. We consider it particularly useful to study empirically the different forms of biases described by Bae18 in the light of our hypothesis of ‘bias transfer’ and using methods similar to what we present in Paper IV especially so since the author recognizes in the conclusion the same general sources of biases as we study in our work, i.e.: “each program probably encodes the cultural and cognitive biases of their creators”, and points in the introduction “measuring bias” as a major challenge, which is what we do in Paper IV.

To state that algorithms are biased, or to assert that algorithms systematically produce an output that is biased, must as a consequence lead scholars to pose the question of whether a biased output could be proven and whether potential causal mechanisms leading to the bias could also be tested and studied. Although an awareness of biases in algorithms has arisen, including the awareness of biases originating from data, no research programs seem to have undertaken the aim to study empirically the (cognitive) mechanisms that may lead to biases in algorithms, i.e., biases that do not originate from data itself, but from the programmers’ cultural backgrounds or from contextual influences in the immediate programming environment.

Thus, rather than pointing once more to the problem itself, as we detail in 3.4 the Paper IV provides instead explanations and insights, derived from our scientific study containing empirical evidence of actual human programming


behaviour, into why this phenomenon may occur. We operate within the same paradigm and with the same agenda as those who study human behaviour, that is, we follow in the path of other multidisciplinary research themes such as Behavioral Economics [KKT91; TK74], Behavioral Transportation Research [GEF14; PFK11] and the recent contributions termed Behavioral Artificial Intelligence [PJ19] and Behavioural Computer Science [PJJ18].
Chapter 3

Contributions

First of all we provide a definition of Usable Privacy. In doing this we strive to stay close to existing standards on usability so to make our definition both adequate as well as easier to accept and adopt by the community. With the definition of Usable Privacy in place, we then work on identifying all the existing usability related statements appearing in the GDPR regulations text, which are termed Usable Privacy Goals. This exhaustive search has gathered 30 Usable Privacy Goals, which have been used to define a list of 25 Usable Privacy Criteria, all having multiple detailed sub-criteria. The development of the Usable Privacy Criteria has followed standard methods from the usability community and has particularly focused on producing measurable criteria, i.e., criteria that could produce measures of usability aspects of privacy that can be placed on various forms of scales. In order to make it easier for certification bodies to adopt our criteria, we have organized these into a Usable Privacy Cube model, building on one of the most advanced privacy certification scheme in Europe, called EuroPriSe.

This work on making GDPR usable, summarized in Section 3.2 is therefore of interest for all the stakeholders of privacy that have already been mentioned in Chapter 1. Besides the immediate impact on the certification processes, the Usable Privacy Goals can work as a gentle entry for the lay people in understanding the GDPR since all of these excerpts from the regulations regard them, i.e., the end user. At the same time, the industries also need to take into account these statements, but for a company the Usable Privacy Criteria are even more valuable because these offer tangible guidelines about what the company should strive for.

In consequence, we have gathered a large group of experts covering all these stakeholders and the relevant areas of certification/standardization, usability, law/GDPR, and technology, in order to validate the above four novel concepts that we have developed. This large study with experts is summarized in Section 3.3 The conclusions from this study show that the experts are quite positive and unanimously supportive.

The idea of measuring privacy, argued for by the above line of work, is even more important for, and indeed was motivated largely by, the second set of contributions that this thesis provides, namely constructing Privacy Labels. We have brought multiple perspectives into this discussion, including that of the end users for whom the Privacy Labels would be most useful for as well as the business for whom the Privacy Labels could constitute a marketing advantage similar to how the energy labeling is being used on home appliances. Privacy Labels extend the already existing concept of privacy icons, also mentioned in the privacy regulations, with the concept of measuring (different aspects of)
privacy on (different) scales and with the concepts of layering of information and personalization of the labels to different types of privacy interested persons. Details of this line of contributions are given in Section 3.1.

One last set of contributions focuses on the developers of software that is processing personal data, particularly data-intensive software such as the modern AI and decision support systems. Multiple privacy biases exist in the general population, and thus in the programmers population. This set of privacy stakeholders is rapidly expanding in size and diversity of skill and knowledge, because of the proliferation of various forms of ‘easy programming’ such as configurations of smart-* systems, graphical programming, or detailed use/configuration of technical software in almost all areas of work that are nowadays being supported by software. Our concern was whether privacy biases would be transferred from the programmer to the programming artifact.

However, there exists almost no evidence in the literature about such bias transference, as opposed to a large amount of evidence showing algorithmic biases coming from biased data sets and other forms of biases in different aspects of the AI algorithms. Therefore, our work (summarized in Section 3.4) has focused on developing a method for studying bias transference in a very general form, i.e., regardless of the type of (privacy) bias. All we have set out to do is to (dis)prove the concept of transference between humans and software. With this research question now answered, the community can work focused on specific types of biases (and particularly on privacy biases) and specific types of programmers working in specific programming environments developing specific types of software.

### 3.1 A Definition of Privacy Labels

The main contribution of the Paper is to show how the concept of Privacy Labeling/Labels could contribute to resolving several of the challenges that privacy is currently facing. It also brings a multidisciplinary approach to the problems discussed, by gathering experts from different fields of practice and research to present and discuss their views on the topic of PL. We thus bring in the following perspectives:

- **Business** (relevant topics including, e.g., market potential, incentives, social responsibility, added value);
- **Law** (e.g., compliance, privacy policies, audit);
- **Regulations** (e.g., national, European, implementations, domain-specific standards);
- **Usability and human factors** (e.g., personas, easy to understand, completeness, contextual);
- **Education** (e.g., psychology of people, of SMEs (Small and Medium-sized Enterprises), of CEOs (Chief Executive Officers), nudging for good, mental heuristics);
A Definition of Privacy Labels

- Technology (e.g., AI, reasoning, automation, dynamic labels, verification);
- Multidisciplinary (e.g., communication across fields, people, or companies, making synergies).

We elaborate on how to combine the seven different perspectives, the roles and priorities of each of these in relation to PL, and point to the state of affairs in the respective fields.

Another main contribution of the Paper I is that we define the concept of Privacy Labeling\(^1\) which can be read below. More details on each of its elements are given throughout the Paper I.

A Privacy Label is a legally binding label containing information about the privacy that a product or service provides. The labels may be physical or digital. They are defined, and are visually presented, in a layered manner, where one can drill-down from a simple overview to more complex information, to allow the user to focus on different levels of detail, depending on the intended use. The labels measure privacy using graded-scales to make it easy to compare two labeled products with respect to privacy aspects relevant for a particular (type of) user.

More specifically, one can imagine PL as being similar to both nutrition facts labels and energy consumption labels. To make PL legally binding one can tie it to a privacy policy text, so that it cannot become a means of deceit in the hands of product advertisers, thus going beyond, but not against, laws and regulations such as GDPR. PL are promoting “privacy as an added value” to a digital product\(^1\) allowing privacy conscious businesses to differentiate themselves from those market competitors that prefer to monetize on the big-data model at the expense of the privacy of the users\(^4\).

As such, Privacy Labels should:

(i) be educational (“Oh, there’s a notion of privacy for this TV-set!”),

(ii) be an incentive and promote business differentiation (business slogans could sound like: “We care about your privacy. So should you!”),

(iii) be legally conscious yet

(iv) be usable for the layperson (“Hey son, what’s all this writing about privacy here?”), hence with sufficient detail as needed, yet visual and simple,

(v) be taken up into regulations and

---

\(^1\)A 90 seconds ‘elevator-pitch’ video where we present Privacy Labeling for a general public can be viewed online at: [https://youtu.be/noE_vF2_GEs](https://youtu.be/noE_vF2_GEs).


3. Contributions

(vi) supported by technologically innovative tools.

Apart from defining PL and discussing it from all the seven relevant viewpoints, we propose along the way: goals, characteristics, open problems, and starting points for research. We believe that identifying all the characteristics that would form an ideal PL could be useful to attaining the goals that we point out in Paper I. To help the community work towards creating such a PL we outline several open problems and also identify good starting points among the existing works, along with drawing, in the end of the Paper I, a general roadmap to follow while taking one (or more) of the three approaches that we propose for managing PL.

Further contributions can be summarized as follows.

- We identify how the present landscape of privacy certifications (including privacy seals and marks [RP18]) could be improved by PL.
- We identify the stakeholders of PL, what are their needs and characteristics, as well as the relation between them.
- We present and discuss the following three important aspects of PL: its educational power to change people’s knowledge of privacy; tools useful for constructing PL; the possible visual appearances of PL.
- We present three approaches to obtaining a PL and propose a roadmap for achieving PL.

3.2 Support for Usability Evaluations of Privacy

The main contribution of the Paper II is a model meant for organizing the process and the criteria to use for evaluating and measuring on scales the usability of privacy of digital products or services. This model builds on the criteria proposed by the EuroPriSe certification scheme by adding usability criteria to it. EuroPriSe does not consider usability, which is the main focus of our work. As such, another contribution of this paper is to show how to add usability aspects to the existing evaluation criteria of EuroPriSe.

Our model is visually represented through a cube, called Usable Privacy Cube (or UP Cube), where each of its three axes of variability captures, respectively: rights of the data subjects, privacy principles, and Usable Privacy criteria. We slightly reorganize the criteria of EuroPriSe to fit with the UP Cube model, where we show how EuroPriSe can be viewed as a combination of only rights and principles, forming the two axes at the basis of our UP Cube. In this way we bring out two perspectives on privacy: that of the data subjects and, respectively, that of the controllers/processors.

We define 24 Usable Privacy criteria each with measurable subcriteria, based on 30 Usable Privacy goals that we have extracted from the whole text of the General Data Protection Regulation. The criteria are designed to produce measurements of the level of usability with which the goals are reached. Precisely,
we measure effectiveness, efficiency, and satisfaction, considering both the objective and the perceived usability outcomes, producing measures of accuracy and completeness, of resource utilization (e.g., time, effort, financial), and measures resulting from satisfaction scales.

We also propose a definition of ‘usable privacy’ based on the ISO 9241-11:2018, showing how the concepts in this standard can be applied to evaluate usability of privacy:

Usable privacy refers to the extent to which a product or a service protects the privacy of the users in an efficient, effective and satisfactory way by taking into consideration the particular characteristics of the users, goals, tasks, resources, and the technical, physical, social, cultural, and organizational environments in which the product/service is used.

Our long term goal is to create a methodology to support service providers to make the privacy of their products more usable. The UP Cube and the Usable Privacy criteria are the first building blocks of the methodology we are aiming for. They are meant as tools, for both usability engineering experts and certification bodies, to evaluate if a product was designed to respect and protect the privacy of its users in an usable way. Once privacy measures and privacy enhancing technologies are integrated into the design of a product, it still remains to find out if (and how much or to what extent) those measures empower and respect the rights of their particular user as intended. In Human-Computer Interaction (HCI) this is determined based on user testing and usability evaluations. The criteria we propose presume the use of such established HCI methods for usability evaluations (e.g., [DR99]).

3.3 Validations with Experts

The main contribution of Paper III is to validate concepts and methods that have been introduced in Paper II using experts working across fields of relevance to these concept, including law and data protection/privacy, certifications and standardization, and usability (also called human-computer interaction).

The expert opinions are varied, example-rich, and forward-looking, which gives an impressive list of open problems where the above four concepts can work as a foundation for further developments. We employed a critical qualitative research method, using theory triangulation applied to three groups of experts categorized as ‘certifications’, ‘law’, and ‘usability’, coming both from industry and academia. The results of our analysis show agreement among the experts about the need for evaluations and measuring of usability of privacy in order to allow for exercising data subjects’ rights and to evaluate the degree to which data controllers comply with the transparency principle. However, the community still needs to find archetypal usability thresholds to guide the businesses in measuring the usability of the privacy of their products in relation to expected competitive advantages.
3. Contributions

Regarding the four concepts mentioned in Section 3.2, these experts validation results first show that, while agreeing with the Usable Privacy definition, the experts are more often interested in finding (and giving) instances of this definition, i.e., examples of where and how it applies in practice. Similarly, the experts are happy with the list of Usable Privacy Goals, seeing these as instances of the Usable Privacy definition. However, even if the Usable Privacy Criteria are seen as a good solution for starting to evaluate usability of privacy, these are not giving enough detail and method for a data controller to be able to understand how (what usability techniques and processes to use) to meet these criteria. Finally, the Usable Privacy Cube model is seen as a good abstract representation of known, but maybe too implicit, principles from data protection evaluations. Therefore, organizing these aspects into a model can only be beneficial to certification bodies, and especially for the adoption of the new Usable Privacy Criteria.

3.4 A Method for Revealing the Transference of Privacy Biases

The main contribution of the Paper IV is to show that there exists another source of machine bias in addition to the well-studied cases of algorithmic bias resulting from characteristics within the dataset on which the algorithms are trained. Namely, we show that biases may originate from

(i) the programmers’ cultural background, such as education or line of work, or

(ii) the contextual programming environment, such as software requirements or developer tools.

Combining an experimental and comparative design, we studied the effects of cultural metaphors and contextual metaphors, and tested whether each of these would ‘transfer’ from the programmer to program, thus constituting a machine bias. The results show

(i) that cultural metaphors influence the programmer’s choices and

(ii) that ‘induced’ contextual metaphors can be used to moderate or exacerbate the effects of the cultural metaphors.

This supports our hypothesis that biases in automated systems do not always originate from within the machine’s training data. Instead, machines may also ‘replicate’ and ‘reproduce’ biases from the programmers’ cultural background by the transfer of cultural metaphors into the programming process.

More specific contributions are related to the methods that we devised to support our hypothesis that people may unknowingly and inadvertently transfer biases to computer programs that they build:
1. We develop a cognitive task intended to reveal biases originating from the cultural background of a programmer, such as education, line of work, and free-time activities.

2. We investigate users with different programming skill levels, i.e., from professional to amateur.

3. We investigate a very general form of bias, but well hidden inside a programming task. This is because the programming task needs to be appropriate also for people with little or no programming skills.

4. We investigate whether *inducing* a bias is effective.

5. We investigate whether people educated in programming exhibit less biases and are less prone to manipulation.

As one can see from the methods we employ, further detailed in [Paper IV](#), another contribution comes from showing how to develop ergonomically valid and reliable instruments, procedures and testing conditions to empirically study such biases transfer. As such, this paper is a foundation for future research endeavours to *improve* and *diversify* these instruments, procedures and testing conditions.

Seen from a wider perspective, the work reported in [Paper IV](#) is relevant for researchers from several fields. First of all, people working in AI and machine learning can be interested in our proposal that biases in machine learning can come not only from the data but also from the people programming the algorithms. Second, people working in psychology and cognitive sciences can be interested in this new application that we propose, where they can apply their skills and methods to study this new form of human bias and its transfer to machines. Third, practitioners working with software engineering or managing software development teams can be interested in studying more various programming environments and tools to see how much human bias is transferred to the programs in each situation. Finally, at a macro level, both governments, private business enterprises, and NGOs would become aware of machine bias originating from human programmers who unknowingly transfer the influences from their own cultural backgrounds to the machine programs. Thus, the target audiences are diverse and would benefit both on a micro level, e.g., in research and development, and in (computer science) education, as well as on a macro level, e.g., in issuing improved knowledge-informed national regulations on the domains where automated decision-support systems operate.

In addition, one can see several immediate benefits of the study presented in [Paper IV](#). For example, in education one could measure how well programming courses train the students, by measuring the bias transfer-rate at the start and end of the courses. Another example is to measure how effective some technology quality assurance method is at removing or identifying programmer’s biases, like testing frameworks, peer programming, abstract/detailed specifications, code generators, etc. Moreover, regarding the growing population of ‘lay’ programmers...
3. Contributions

in the smart-living and IoT-ubiquitous programming environments of today (i.e., almost everyone in technologically ‘modern’ societies) both business companies and consumers would benefit from more insight into the non-conscious influence of culture and context on the programming choices that are made by the ‘novice’ programmer that has no formal training. In terms of education and learning, we argue that this insight could be used to help consumers become more aware of the cultural and contextual influences that shape their cognitive tendencies when they are programming.
Chapter 4

Overview of the Research Papers

This section is only meant to give a quick overview of the research papers that are included in this thesis. This can very well be skipped by a reader that is keen on reading the respective papers, found further down in this thesis. Even more so since the previous Chapter 3 already gave much context about these papers, and therefore, here one would find only a very concise and more factual account of the contents of these papers.

4.1 Paper I: A Multidisciplinary Definition of Privacy Labels: The Story of Princess Privacy and the Seven Helpers

In this paper we employ storytelling and metaphors from fairytales to make reader-friendly and streamline our arguments about how a complex concept of Privacy Labeling (the ‘knight in shining armor’) can be a solution to the current state of Privacy (the ‘princess in distress’). We give a precise definition of Privacy Labeling (PL), painting a panoptic portrait from seven different perspectives (the ‘seven helpers’): Business, Legal, Regulatory, Usability and Human Factors, Educative, Technological, and Multidisciplinary. We describe a common vision, proposing several important ‘traits of character’ of PL as well as identifying ‘undeveloped potentialities’, i.e., open problems on which the community can focus. More specifically, this position paper identifies the stakeholders of the PL and their needs with regard to privacy, describing how PL should be and look like in order to address these needs. Throughout the paper, we highlight goals, characteristics, open problems, and starting points for creating, what we consider to be, the ideal PL. In the end we present three approaches to establish and manage PL, through: self-evaluations, certifications, or community endeavors. Based on these, we sketch a roadmap for future developments.

4.2 Paper II: Making GDPR Usable: A Model to Support Usability Evaluations of Privacy

In this paper we introduce a new model for evaluating privacy that builds on the criteria proposed by the EuroPriSe certification scheme by adding usability criteria. Our model is visually represented through a cube, called Usable Privacy Cube (or UP Cube), where each of its three axes of variability captures, respectively: rights of the data subjects, privacy principles, and usable privacy criteria. We slightly reorganize the criteria of EuroPriSe to fit with the UP Cube model, i.e., we show how EuroPriSe can be viewed as a combination of only rights and principles, forming the two axes at the basis of our UP Cube. In this way we also want to bring out two perspectives on privacy: that of the
4. Overview of the Research Papers

data subjects and, respectively, that of the controllers/processors. We define usable privacy criteria based on usability goals that we have extracted from the whole text of the General Data Protection Regulation. The criteria are designed to produce measurements of the level of usability with which the goals are reached. Precisely, we measure effectiveness, efficiency, and satisfaction, considering both the objective and the perceived usability outcomes, producing measures of accuracy and completeness, of resource utilization (e.g., time, effort, financial), and measures resulting from satisfaction scales. In the long run, the UP Cube is meant to be the model behind a new certification methodology capable of evaluating the usability of privacy, to the benefit of common users. For industries, considering also the usability of privacy would allow for greater business differentiation, beyond GDPR compliance.

4.3 **Paper III: Expert opinions on making GDPR usable**

In this paper we present the results of a study done in order to validate concepts and methods that have been introduced in **Paper II**. We use as respondents in our interviews experts working across fields of relevance to these concepts, including law and data protection/privacy, certifications and standardization, and usability (as studied in the field of Human-Computer Interaction). The expert opinions are varied, example-rich, and forward-looking, which gives an impressive list of open problems where the above four concepts can work as a foundation for further developments. We employed a critical qualitative research, using theory triangulation through involving three groups of experts categorized as ‘certifications’, ‘law’, and ‘usability’, coming both from industry and academia. The results of our analysis show agreement among the experts about the need for evaluations and measuring of usability of privacy in order to allow for exercising data subjects’ rights and to evaluate the degree to which data controllers comply with the transparency principle. However, the community still needs to find archetypal usability thresholds to guide the businesses in measuring the usability of the privacy of their products in relation to expected competitive advantages.

4.4 **Paper IV: Revealing Human-to-Computer Bias Transference**

In this paper we hypothesize that machine’s bias may originate from (i) the programmers’ cultural background, such as education or line of work, or (ii) the contextual programming environment, such as software requirements or developer tools. Combining an experimental and comparative design, we studied the effects of cultural metaphors and contextual metaphors, and tested whether each of these would ‘transfer’ from the programmer to program, thus constituting a machine bias. The results show (i) that cultural metaphors influence the programmer’s choices and (ii) that ‘induced’ contextual metaphors can be used to moderate or exacerbate the effects of the cultural metaphors. This supports our hypothesis that biases in automated systems do not always originate from
within the machine’s training data. Instead, machines may also ‘replicate’ and ‘reproduce’ biases from the programmers’ cultural background by the transfer of cultural metaphors into the programming process. Implications for academia and professional practice range from the micro programming-level to the macro national-regulations or educational level, and span across all societal domains where software-based systems are operating such as the popular AI-based automated decision support systems.
Chapter 5

Conclusion

5.1 Summary of Contributions

In all our papers we adopted a multidisciplinary perspective on privacy for the first three papers, and on biases in software engineering for the fourth. In Paper I and Paper III we bring to the discussion different experts representing the main stakeholder groups for privacy labels and for evaluating the usability of privacy respectively.

When dealing with a complicated concept such as privacy, which is faced with multiple long-standing problems as discussed in both Paper I and Paper II, then to develop a solution can be a matter of correctly representing and including all the sides concerned. In response, in Paper I, and as detailed in Section 3.1 of this thesis, we propose an all-encompassing definition of Privacy Labeling as a possible start on the road towards a solution to many of the current privacy problems our society is struggling with.

When developing a panoptic concept such as the Privacy Labels, we found it necessary to include discussion partners among the various stakeholder groups. Therefore, the ideas that we have presented have roots in our conversations with experts from the seven different fields that we considered relevant for privacy labeling. Our goal with bringing in all these views was to investigate the concept of privacy labeling (and its implications) from many different angles. No one single discussion partner has ‘the right answer’, but their collective opinions sum up to a result that is much more comprehensive and powerful than any one view could accomplish on its own.

Following the multidisciplinary approach, in the case of the Paper II we propose applying methods of usability evaluations from the Ergonomics of human-system interaction to data protection, more precisely to GDPR. We worked directly with the GDPR text, guided by [Eur18], which also inspired our structuring of the EuroPriSe criteria into rights and principles. Our HCI and usability perspective on privacy is influenced by the seminal works [AM05; AS99; Cra18; GK03; KBK05; KKB12; Pat+03].

To guide the usability evaluation of privacy we proposed the Usable Privacy Cube model that has multiple benefits:

(i) emphasizing both the perspectives of data subjects and of controllers;

(ii) representing visually on the three variability axes the existing rights and principles criteria from EuroPriSe, together with our new UP criteria;

(iii) visualizing intersections between the three axes;

(iv) allowing ordering of the criteria on each axis.
5. Conclusion

To build the UP Cube we have:

- identified from the GDPR text 30 UP goals,
- created 24 UP criteria, each with measurable subcriteria, and
- restructured the criteria of EuroPriSe, laid as the basis of the UP Cube.

In Paper I we present the main stakeholders for Privacy Labels, and we identify their needs, goals, and challenges in respect to privacy. One of the stakeholders are the programmers that have to implement the data protection requirements into their code. In Paper I we look at what are the characteristics of the ‘convenience users’ that might influence their choices and decisions when it comes to protecting their own privacy. Similarly, we think that programmers might be influenced as well by their cognitive styles or environment when implementing privacy related measures in their software. Our ultimate goal is to find out if privacy related biases that programmers exhibit are transferred to the programming artifact. However, we did not have a good place to start from as the research on the transfer of bias from human to machine was very scarce. Therefore, we had to find out first if bias transfer occurs at all between the human programmer and the software. This motivated the research in Paper IV.

The Paper IV is thus studying a hitherto not researched phenomenon, namely the transfer of human biases from the (not necessarily expert) programmer to the artifact that is developed (or configured). The main hypothesis of this paper is that the machine bias may originate not only from biased data, but also from the programmer’s biases in terms of influences from the cultural background as well as contextual influences from the programming environment.

Interestingly, under conditions of uncertainty (e.g., in the absence of instructions or specifications, something which is often the case for ubiquitous systems programming carried out increasingly by non-experts), we observe that the programmers’ cultural background influences the choices they make and are subsequently transferred from the programmer to the program artifact. Thus, cultural metaphors in terms of irrelevant and inappropriate influences on the programming task at hand, represent instances of biases that are being transferred from humans to machines. This implies that human culture ‘transfers’ to machines through the humans that program these, thus representing a strong source of bias.

Equally interesting, attempts to moderate the strong influence from the cultural metaphors by means of experimentally introducing ‘hidden’ (i.e., not consciously detected) contextual metaphors, were only successful to a certain extent. When the priming metaphor was chosen well (as in the case of ‘philosopher story’ related to the ‘balance’ rationale; with words that were easy to understand and rather common in a standard vocabulary;) we saw influences in both directions of strengthening the cultural background as well as moderating it, each time tipping the balance of answers in the direction of the metaphor. These findings are orthogonal to what traditional and current machine bias research suggests, i.e., that machine bias originates from data, and thus our findings
provide new insights into the origins of bias in the wide spreading AI and decision-support systems.

Having proven the existence of bias transfer between the programmer and the program, we can now further advance our research to identify more specific type of bias, such as privacy, as it was our initial intention.

## 5.2 Future Work

In [Paper I] the future work is reflected in the eleven ‘open problems’ that are marked throughout the text in the contexts that they are being relevant for – their respective research/practice domains represented in the paper. Some examples of such open problems are:

- Making the concept of PL flexible enough to accommodate different purposes and audiences, and for each such PL instance how should it be designed in order to convey the intended information to the intended audience.

- If PL use tools to translate/explain privacy policies to convenience users, we would like to investigate how can these same tools be useful to the programmers to understand how to implement the statements from privacy policies and law.

- This is an open problem that we set about solving in [Paper IV] How can PL prevent the transfer of the programmers’ privacy views and biases into a source of privacy problems in software?

- Finally, an open problem that we set about solving in [Paper II] and [Paper III]: Evaluating the degree of data protection is complicated, but needs to be made measurable, at least for some of the privacy aspects, and fit into a graded-scale system.

Working towards the open problem 5.2 above, in [Paper II] we propose a model, which is the groundwork for building a certification methodology, extending EuroPriSe to evaluate the usability of privacy. The proposed UP criteria are designed to produce measurable evaluations, useful for generating privacy labels in order to guide stakeholders when choosing technological products, by representing and visualizing the different levels of privacy. To achieve this larger goal, one needs to investigate which existing HCI methods for usability testing should be used for each of the UP criteria, and in what way.

One example of such a usability method for measuring the perceived usability of a system is the System Usability Scale (SUS) [Bro96], a ten-item attitude Likert scale questionnaire. The standard [ISO18] Annex B: Usability measurements also gives examples of methods that produce measurements relevant for our UP criteria, such as observing the user behavior to identify the actual usability problems, or asking the users to carry out tasks in a real or simulated context of use and measuring the outcomes. The experts can also run heuristic evaluations
following design principles, theories and standards from the design and cognitive fields. More concrete examples of HCI methods and how these could be used for privacy and security solutions can be found in [KBK05].

With the same goal of achieving a complete methodology that can be taken in use by the accreditation bodies, building on the present model, one could create a visual representation of the evaluation, i.e., a translation of the measurements of usability of privacy provided by the UP criteria into a visually appealing privacy label. The concept of Privacy Labels has been discussed more in detail in relation to [Paper I]. This should serve as a vertically graded scale to differentiate a customer product from another. According to ISO 9241-11:2018, “where usability is higher then expected, the system, product or service can have a competitive advantage (e.g. customer retention, or customers who are willing to pay a premium)”. The visuals will be thought to come in addition to the GDPR compliance seal and reflect the usability of the privacy implemented. The purpose will be the same as for the methodology, to help the businesses that have already achieved GDPR compliance to further differentiate themselves on the market. From the point of view of the user of the product, the visual scale would offer support for choosing the service or product that best respects her privacy expectations.

When it comes to [Paper IV] one venue for future research would be to refine our study design’s ability to elicit cultural or contextual influences in an even more fine-grained manner, specifically by improving our instruments and procedures.

One good source of alternative investigations – working towards the open problem 5.2 identified in [Paper I] – can be the study of specific biases in specific situations or social activities where software is paramount. One example can be biases related to privacy in the big data economy (sometimes called the ‘surveillance capitalism’ [Zub19]), e.g.: Are privacy related concepts or views from the cultural background – which is specific to the programmer – transferred to the software – which is used on an international scale? One can imagine a programmer coming from a cultural background that always promotes the slogan “You have zero privacy; get over it!”, or another programmer from a background that “is entrenched by rules and regulations about who/how any form of private electronic data can be used”. Are such different cultural views transferred to the software built by these two different programmers? What is the global influence of such bias transfers? In this setting, one could alternatively study biases coming from the user of the software (not the programmer) to see whether the user biases (call them ‘wishes’ or ‘needs’) are transferred to the software through specifications elicitation, user stories, and other interaction design methods [LFH17; RSP11] that are now a popular way of developing software systems.
Bibliography


Bibliography


Bibliography


[**KD18**] Kamara, I. and De Hert, P. “Data protection certification in the EU: Possibilities, Actors and Building Blocks in a reformed landscape”. In: *Privacy and Data Protection Seals*. Springer, 2018, pp. 7–34.


[Pat05] Patterson, D. A. “20th century vs. 21st century C&C: The SPUR manifesto”. In: *Communications of the ACM* vol. 48, no. 3 (2005), pp. 15–16.


Papers
Paper I

A Multidisciplinary Definition of Privacy Labels


Conditionally accepted subject to revisions for the journal Information and Computer Security.

# A Multidisciplinary Definition of Privacy Labels

<table>
<thead>
<tr>
<th><strong>Journal:</strong></th>
<th><em>Information and Computer Security</em></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Manuscript ID:</strong></td>
<td>Draft</td>
</tr>
<tr>
<td><strong>Manuscript Type:</strong></td>
<td>Original Article</td>
</tr>
<tr>
<td><strong>Keywords:</strong></td>
<td>privacy labels, General Data Protection Regulation, usability, Certification, automation, psychological models</td>
</tr>
</tbody>
</table>
A Multidisciplinary Definition of Privacy Labels

Authors Anonymous

Abstract

Purpose: We present arguments about how a complex concept of Privacy Labeling can be a solution to the current state of Privacy.

Design/methodology/approach: We give a precise definition of Privacy Labeling (PL), painting a panoptic portrait from seven different perspectives: Business, Legal, Regulatory, Usability and Human Factors, Educative, Technological, and Multidisciplinary. We describe a common vision, proposing several important ‘traits of character’ of PL as well as identifying ‘undeveloped potentialities’, i.e., open problems on which the community can focus.

Findings: This position paper identifies the stakeholders of the PL and their needs with regard to privacy, describing how PL should be and look like in order to address these needs. Main aspects considered are the PL’s educational power to change people’s knowledge of privacy, tools useful for constructing PL, and the possible visual appearances of PL. We also identify how the present landscape of privacy certifications could be improved by PL.

Originality: We adopt a multidisciplinary approach to defining PL as well as give guidelines in the form of goals, characteristics, open problems, starting points, and a roadmap for creating the ideal PL.

Keywords: privacy labels, General Data Protection Regulation, usability, certification, automation, psychological models

1. Introduction

The right to privacy\(^1\) is something precious and frail (an integral value appearing in the Universal Declaration of Human Rights), which we need to take good care of in order not to lose it. In its current state, privacy does not always serve the people, but mainly a few very wealthy and influential actors prosper from its misuse. Loss of privacy has both micro implications, at a personal level (e.g., people being influenced to buy what they do not want or need (Matz et al., 2017), to vote for extremists (Isaak and Hanna, 2018), or to develop antisocial behavior), but also macro implications, at a societal level (e.g., a society living in fear of being watched by surveillance capitalists (Zuboff, 2019) or manipulated on social media (Starbird, 2019)).

Privacy is personal and contextual, having social and political ramifications, but most of the population does not see, or understand, even some of its basic implications. The lack of privacy literacy can be partly attributed to commercial entities that often, while profiting from handling data, work hard to keep privacy “out-of-sight [is out-of-mind]” e.g., telling people infamously “You have zero privacy anyway. Get over it.”\(^2\) (Solove, 2011). Privacy misapprehension by the population is also due to its complexity, having kept many brilliant minds preoccupied for at least a century, since photography as a new technology used by media became widespread (Brandeis and Warren, 1890). In the current digital society, privacy (Acquisti et al., 2007) has even stronger forces compounding its complexity, coming from, e.g., technological advances in miniaturizing hardware that enabled cheap privacy-invasive gadgets, powerful algorithms that can make inconceivable inferences (Schneier, 2015), or supercomputing in ‘invisible’ clouds (Borning et al., 2020).

---

\(^1\) We will be liberal in our use of the word “privacy”, including the concept of personal (digital) data protection as used by (GDPR, 2016); however, we are aware that “privacy” has different formulations and scopes in different European or international laws.

\(^2\) https://www.wired.com/1999/01/sun-on-privacy-get-over-it/
This paper analyzes how the concept of Privacy Labeling/Labels, which hereafter we refer to as PL, could contribute to resolving several of the challenges that privacy is currently facing. Given the many facets of privacy and its society-wide implications, it is important to adopt a multidisciplinary approach. This work started from a workshop in spring 2020 where experts from different fields of practice and research gathered to present and discuss their views on the topic of PL. We thus bring in the following perspectives: Business (relevant topics including, e.g., market potential, incentives, social responsibility, added value); Law (e.g., compliance, privacy policies, audit); Regulations (e.g., national, European, implementations, domain-specific standards); Usability and human factors (e.g., personas, easy to understand, completeness, contextual); Education (e.g., psychology of people, of SMEs (Small and Medium-sized Enterprises), of CEOs (Chief Executive Officers), nudging for good, mental heuristics); Technology (e.g., AI, reasoning, automation, dynamic labels, verification); Multidisciplinary (e.g., communication across fields, people, or companies, making synergies).

We elaborate on how to combine the seven different perspectives, the roles and priorities of each of these in relation to PL, and point to the state of affairs in the respective fields.

We define the concept of Privacy Labeling below, and throughout the paper we detail each of its elements.

A Privacy Label is a legally binding label containing information about the privacy that a product or service provides. The labels may be physical or digital. They are defined, and are visually presented, in a layered manner, where one can drill-down from a simple overview to more complex information, to allow the user to focus on different levels of detail, depending on the intended use. The labels measure privacy using graded-scales to make it easy to compare two labeled products with respect to privacy aspects relevant for a particular (type of) user.

Apart from defining PL and discussing it from all the seven relevant viewpoints, we propose along the way (schematically presented in Figure 1):

- goals (numbered as G.x and marked with a flag icon 🚩) intended to be achieved with the help of PL;
- characteristics that we think PL should have (numbered as C.x and marked with a diamond icon ◊);
- open problems (numbered as OP.x and marked with a scholar cap icon 🧕) that the community can address while striving for reaching any of the above;
- existing works that can function as good starting points (marked with a map-pointer icon 📍) for some of the above.
Following the practice of “eating our own dog food”, we mark the identified goals, characteristics, and open problems, with icons to make these important contributions of this paper more accessible (see in Section 6 our discussions on the use of icons in PL).

We believe that collecting all the characteristics would form an ideal PL that could be useful to attaining the goals that we have pointed out. To help the community work towards creating such a PL, we outline several open problems and also identify good starting points among the existing works, along with drawing, in the end of the paper, a general roadmap to follow while taking one (or more) of the three approaches that we propose for managing PL.

The structure of this paper is displayed in Figure 2. An example that we use throughout the paper is the PrivacyLabel.org project from the Netherlands.

2. Evaluating the need for Privacy Labels

The Recital 100 of GDPR encourages “the establishment of certifications mechanisms and data protection seals and marks [to allow] data subjects to quickly assess the level of data protection of relevant products and services” (GDPR, 2016). While Art. 42(1) encourages the implementation of certification and data protection seals for demonstrating compliance by accredited certification bodies, PL should go beyond and measure on a scale how well the privacy is respected and how easy is for a user to understand that (see also G.2). Compared, e.g., with energy consumption where one number is enough, maybe placed on a graded scale for visual comparison, and measured with well known instruments, privacy involves multiple concepts spanning from social to technological, most of which we do not know how to measure, definitely not in a universal way since they are relative to a person’s view on privacy influenced by this person’s level of knowledge about privacy.

G.1: One Goal is to build PL on/into existing certifications.

3https://www.privacylabel.org/ project from the Netherlands.
Examples of existing certifications include: Datenschutzgutesiegel, granted to systems and products by ULD (The Schleswig-Holstein Data Protection Authority)\(^4\), EuroPriSe\(^5\), Common Criteria\(^6\) (ISO/IEC 15408) including a Privacy Class meant for defining privacy functionality, focusing on aspects such as anonymity, pseudonymity, unlinkability, unobservability (Pfitzmann and Köhntopp, 2001). Some of these are partly required by law, e.g., ULD Datenschutzgutesiegel is used for public procurement in the Schleswig-Holstein German state.

For privacy certifications we also need evaluations. The challenge is that the focus of schemes such as the above is much on security testing and penetration testing. There is a need for more formal evaluation, verification, or testing of privacy requirements. Aspects of usability should also be included in the evaluation of privacy. Parallel this with security where the weakest link is often the end-user. Nowadays communication protocols are formally proven secure, but still security breaches occur because the user interfaces are not usable, leading end-users to doing mistakes, e.g., the security warnings for SSL certificates or other types of security warnings where the end-users have to make decisions without good guidance or usable instructions (Whitten and Tygar, 1999).

**G.2:** PL should offer a way to reach the usability goals of GDPR.

Usability and Human-Computer Interaction (HCI) techniques have mostly been developed for making technologies that are difficult to use, or made for highly specialized experts, more easy to understand and interact with, both for the expert user, but often also for novice users. Usability is even more important for privacy since privacy is a highly complex concept, related to complex technologies such as AI and Big Data, but which is especially addressed to the laypeople, not to experts, because privacy is a human right.

One starting point for evolving certification schemes from seals and trust marks towards privacy labels of the energy efficiency type, i.e., aiming for goal G.1, is by measuring the usability of privacy using HCI methods, thus covering also G.2. (Johansen and Fischer-Hübner, 2020) works in this direction by proposing a set of criteria thought to produce measurable evaluations of the effectiveness, efficiency, and satisfaction with which privacy goals of GDPR are reached.

Besides being adaptable to different contexts, the PL should also be trustworthy (hence G.3).

In 2013, the European Consumer Centres’ Network published a trust mark report “Can I trust the trust mark?” where it is brought to the attention the importance of establishing reliable trust and demanded a more uniform practice of European trust marks.

The research done in projects such as PRIME (Camenisch et al., 2011) and PrimeLife (Fischer-Hübner et al., 2011) has shown that the end-user is having a lack of trust especially in the case of PETs (Privacy Enhancing Technologies). Users often have difficulty believing PETs’ privacy protection claims mainly because these are counterruitive.

**G.3:** PL should be developed and applied uniformly so that it becomes an important trust factor.

Increasing customers’ trust can also be achieved by showing that the organization takes privacy seriously, by displaying privacy information that can be understood, instead of a long legal text. Privacy governance can in this way become a competitive asset, an unique selling proposition (Hoffman, 2014).

### 3. PL stakeholders and their needs

Privacy labels can have different purposes (e.g., for internal use, for showing compliance or only for fulfilling transparency requirements, for marketers to use for selling effectively/targeted) and be intended for different audiences (e.g., data subjects or the controllers). However, it is probably difficult to put everything in one label.

**C.1:** One Characteristic of PL is to be usable, for different purposes, by different types of stakeholders.  

\(^4\)https://www.datenschutzzentrum.de/guetesiegel/  
\(^6\)https://www.commoncriteriaportal.org/  
An Open Problem is how to make the concept of PL flexible enough to accommodate different purposes and audiences, and for each such PL instance how should it be designed in order to convey the intended information to the intended audience.

The rest of this section surveys the various audiences and purposes PL may have, starting with the “convenience users” being our primary target, and continuing to discuss the needs and expectations of businesses, lawyers, regulators and authorities, and programmers.

3.1. Convenience users

We define “convenience users” as those people that nowadays trade in their privacy for convenience, most often without knowing what they are trading in.

The convenience users are much of the time running on ‘autopilot’ when they are making judgments, e.g., when shopping online. This happens from multiple reasons, e.g., willpower depletion (Baumeister and Tierney, 2011), heuristic and intuitive thinking (Kahneman, 2011), or manipulations such as priming done through media channels and advertising (Cialdini, 2007; Thaler and Sunstein, 2009). Advertisers and commercial businesses use extensively behavioral psychology to influence, whereas the governments and authorities seem to assume that people are rational and as such they do not see the need to push any psychological buttons.

At the same time, we also know that people have different values and traits of personality (Fig. 3), and that different people may prefer different levels of privacy, that may also change with time and context (Westin, 1991; Gerber et al., 2018). These differences in people’s personality traits are susceptible to different kinds of influences (Acquisti, 2009).

People also have different cognitive styles. Some have an intuitive approach to making judgments and decisions about something, while others have a more analytical approach (Egelman and Peer, 2015). We also know that people’s judgments are very much influenced by their current emotional state, their affect, their moods (Kitkowska et al., 2020), e.g., car salespersons are known to try to get the buyer in a good mood, in order for the buyer to be less critical. Combining the personality and cognitive style where a person is not confident and is not prone to making analytical judgments, but instead has an intuitive cognitive style and also is in a good mood, is what might be the characteristic of the typical convenience user. In this case it might be necessary to push other types of buttons to slow the cognitive processing of this type of users down, so that they can really think about what they’re doing, instead of processing and making judgments intuitively and heuristically.

C.2: PL should be individualized by considering the psychology of personality, cognitive styles, and social influence (cf. Fig. 4).

One starting point is the combination of the elements from Fig. 4 that would result in a more nuanced distribution of privacy preferences and attitudes. The literature has identified several of such predominant

Figure 3: Personality traits, from Chap. 30 of (Howard, 2014).
profiles known as privacy personas (Morton and Sasse, 2014; Dupree et al., 2016). Rather than adopting an exclusivist and reductionist approach, the PL should be able to adapt to different privacy preferences or privacy personas.

3.2. Businesses

PL could provide a competitive advantage (Martin and Murphy, 2017), especially in markets such as Europe, where privacy protections are required by law. Even in privacy unregulated markets, a company that has visible and easily understandable statements of compliance will be perceived by the customers as one that takes privacy seriously. This has the potential to increase the customer’s trust in the company, which is seen in (Bachlechner et al., 2020) as essential for data-driven businesses.

However, against the possible benefits one has to weigh the possible costs. One factor to consider is the price of obtaining and maintaining a PL, e.g., how often re-certification is required or whether the PL will reduce or increase the use of other resources, e.g., compliance officers, additional IT staff, preparing more documentation, or legal council.

G.4: **PL should become something worth investing in, that brings clear benefits and offers a competitive advantage, outweighing the costs.**

Businesses tend to prioritize what is most requested by their customers. Recent research shows that privacy is a major blocker for the adoption of new technologies, e.g., (Barbosa et al., 2020) studied people’s considerations for adoption of smart home devices, and found that half of the 613 participants have named privacy or security concerns/risks as being a blocker for them acquiring an IoT device. Privacy was also ranked second, after ‘convenience’, as being considered when purchasing such devices.

Depending on the needs of the business, e.g., in which territory they plan to distribute a new product, the privacy related documentation is adjusted considering the costs and benefits. The level of detail of a privacy policy (and of the related PL, cf. C.4 & OP.6) depends on, e.g., what the local regulations require, but also on how complicated the processing of personal data of the respective company is, restricted by the resources available to the company (e.g., to pay lawyers or certification processes). Since privacy policies can vary widely, it is important to consider the flexibility of PL and create different modules that can adapt to more complicated needs.

C.3: **PL must be modular and flexible to accommodate the different needs of businesses in relation to local regulations and their customers’ demands.**

Together with stricter regulations such as GDPR and peer pressure, e.g., Google Play requiring all apps to have a privacy policy before being allowed on the app store, privacy has become a more important topic for businesses. Irrespective of whether it is a large multinational company or a start-up, every business that
Information and Computer Security
deals with digital data nowadays has to have a privacy policy.

C.4: *PL should be closely related to privacy policies,*\(^8\) which the law asks all businesses to have.

One starting point in the process of making a meaningful privacy policy could be to employ the support from a tool, such as the NL.PL, where one is guided into providing the information required for making a PL (and thus also a privacy policy).

A more general problem encompassing what has been said until now about privacy policies and businesses adopting PL is:

**OP.2: How can Privacy Labels reach the mass market?**

It can be said that privacy policies, being required by law for every digital product or service handling private information, have already reached the mass market. By attaching PL to privacy policies, these too would reach all the consumers under the condition that businesses would be willing to adopt PL. Therefore, in an unregulated market place, PL has to appeal also to businesses, to bring value to their products. Otherwise, regulations and legislation, along with incentives, can be used to attain a critical mass of businesses using PL. A tool such as NL.PL would also be useful in this respect, as it would make the creation of PL more practical and affordable, through being easy to use without needing legal expertise.

3.3. Lawyers

A common situation relevant for lawyers, is when a client is considering obtaining a PL. The primary choice is a PL that would also involve a certification by an independent certification body, e.g., if the client is established in the EEA (European Economic Area) then it will have to adhere to data protection legislation such as the GDPR and also local data protection requirements. However, PL could also be relevant for establishments outside the EEA, for example if they want to compete with businesses in the EEA. Since there are not many privacy certification options, it is also often that PL are desired rather as information conveying tools. Since privacy legislation (the same as certifications) can vary between different geographical regions (Kaminski, 2020), one can see PL as a harmonizing factor because of its international nature, managed by a global community. The type of client is important as well, with factors such as the business size, multinational, national, or an SME, combined with the nature of their commercial transactions. A PL can thus be relevant both for B2C (Business to Consumers) as well as B2B (Business to Business).

C.5: *PL must take into consideration both commercial and legal aspects, and their interdependencies.*

In terms of legal implications, having a PL does not relieve the company of its obligations to adhere to the data protection law. PL is only a modality to demonstrate compliance, as stipulated by GDPR in Art. 42(1). Therefore, besides maintaining documentation relevant for PL, the company still needs to implement technical and organizational measures, which should not be seen as something additional to having a PL, but as part of the requirements necessary for obtaining the PL.

C.6: *PL should reflect technical and organizational measures taken by the company.*

Certifications sometimes have additional requirements that are more onerous than the law. It can be more difficult to obtain the PL through a certification process than to just be compliant with data protection law. This could have potential additional costs but also potential benefits. Having a PL might have implications for other business areas where legal advise is usually needed, as in marketing where the lawyer has to assist in ensuring that the PL is not misleading or inaccurate, otherwise it can be judged as false marketing.

C.7: *PL should come with supporting guidelines for businesses so that they do not include false information unwittingly.*

---

*One can also consider other privacy related documents or certification processes.*
3.4. Programmers

Technology people are struggling to understand the legal terminology and how to implement a system so to conform with the statements appearing in the legislation and in the privacy policies made by their leadership. This “legal-text-to-code” gap is even larger than the well-known gap between software requirements (or specifications) and their implementation. Besides standard questions that programmers ask, such as “What does data minimization mean?”, one important problem that they face is how to match the “purpose” stated in a privacy policy (and presented by the PL) with the precise usage of the data during any execution of their software implementation. These are necessary questions when trying to enforce privacy or prove compliance with the GDPR or (maybe easier) to own privacy policies. It is already difficult for lawyers at an organizational level to deal with such questions, which become even more complicated to answer when trying to look at the software code.

OP.3: If PL use tools to translate/explain privacy policies to convenience users, we would like to investigate how can these same tools be useful to the programmers to understand how to implement the statements from privacy policies and law.

One starting point can be to try to use existing formal tools to analyze at least the more critical parts of the code by, e.g., doing code inspection. Since it is difficult to analyze code automatically, putting a human expert into the loop can be a more feasible first approach of doing semi-automated code evaluation and verification for privacy compliance. There exist several recent technological advances on automating particular aspects of privacy, e.g., on data-flow (Antignac et al., 2016); on data minimization (Antignac et al., 2017); on privacy by design (Hoepman, 2014); and in general on Privacy-Enhancing Technologies (PETs) (Danezis et al., 2015). However, it can take long for a research idea to reach the programmers, and even more so for PETs since more often than not, these prove too difficult for software development companies to implement or adopt in their software or DevOps tool-chains.

Since PL would be associated to privacy agreements (cf. C.4) and having one goal to explain concepts such as purpose of processing (see Section 4), they would help programmers to better understand those aspects from the privacy agreement that are relevant for the product they are building.

3.5. Regulators, Certification bodies, and Authorities

Certifications bodies as stakeholders can see the PL as a means to convey their certification results. The provisions in Art. 37-39 of GDPR strengthen the role of the certification bodies as a means for the companies to show compliance. PL should contribute to the further development and enhancement of the existing certification schemes (cf. G.1).

Data Protection Authorities (DPAs) tend to rely on detailed sources, such as privacy policies and technical documentation, in their audit work. Therefore, highly relevant for DPAs would be the deeper layers of the PL, where detailed information is offered (cf. Section 6.1). However, DPAs are also responsible with checking if the visual and the “surface” components of the PL are an accurate reflection of the privacy policies and actual practices. In this case, their auditing work could be simplified through the automation tools and process used to generate the PL (cf. Section 5).

Data-intensive technologies such as AI-based decision systems or management software have entered also in the many state institutions such as in policing or courts (Dressel and Farid, 2018). Privacy is maybe of a greater concern to such institutions than it is to companies. In state institutions, decisions on purchasing a piece of technology or service is done through a highly regulated and transparent process called procurements. Privacy would thus be part of the requirements mentioned in the procurement call. PL could also here be used to make it easier to evaluate the proposals. This is the same way of applying PL in any form of technology purchase decision, be it done by a governmental institution in a procurement process, or a company management person looking to acquire a new service, or a convenience user when looking to by a new IoT device.

G.5: In public/private procurement, PL could be an advantage or sometimes even a requirement.

Data Protection Officers (DPOs) are, according to Art. 37-39 of GDPR, acting as intermediaries between the supervisory authorities, data subjects, and the organization by which they have been appointed. The
role of DPOs is to facilitate compliance, and, besides certifications, are another instrument that can be adopted by companies to ensure accountability. In some cases, GDPR makes appointing a DPO mandatory, while for the rest of the organizations this is voluntary, in which case the organization may choose to use external DPOs. DPOs as stakeholders for PL would have commercial interests to foster self-evaluations, where they could provide companies with input.

4. PL as a means of education and behavior change

The New Chicago School model (Lessig, 1998) explains how there are several modalities of regulation for the behaviors of people, and we would also argue that it applies to businesses as well. PL are meant to help regulate the behavior of convenience users when making choices that might influence their privacy, as well as regulating the behavior of businesses that handle private data. Therefore, the multidisciplinary character of PL involves multiple stakeholders, besides the law and regulatory institutions, in driving privacy behavior changes.

OP.4: One open problem that PL could be useful for is to help change the behaviors and attitudes of people in regard to privacy.

The Prochaska model of stages of behavioral change (Prochaska and Velicer, 1997) identifies several stages of awareness and appropriate actions. One may not be aware at all that she needs to make behavioral changes, meaning that the action targeting this person is to raise her awareness, e.g., in regard to privacy aspects. Then the person moves into the contemplation phase when realizing that there is an important concern, e.g., privacy, which she needs to think about. Having learned and understood the problem, the person has to determine whether, and what, to do. This is a decision point where PL can help, e.g., when the person needs to take action when buying a digital product.

Many people are not even aware of their lack of privacy. We see the PL as a tool for raising awareness. We already have examples from the food industry where labels are used to raise awareness about the quality of the food. However, such markings can easily be used misleadingly for commercial purposes as well. Therefore, if not regulated, PL could be used in a suggestive and deceptive manner to induce subjective and irrational (i.e., inappropriate) behavior that is not in the user’s best interest.

G.6: PL aims first at raising awareness of privacy in the population.

A simple seal/label can be used to raise awareness. However, we would go beyond a mere seal. We envision a privacy label that displays information through which people can learn more about privacy related aspects, e.g., that location sharing is a privacy sensitive information. The label can thus be the point of entry, providing the information that can be used to further educate people.

Knowledge is power, and our PL can help increase the privacy knowledge of laypeople, thus empowering them to make well informed choices in the technological world and thus actively participate in the data economy.

G.7: PL as an educational tool has the goal to increase privacy literacy in the population.

According to the theory of planned behavior (Ajzen, 1991), behavioral change and judgments that one makes are conscious and are planned. This is a rational approach to behavioral change and decision making. However, it is known that people do not always make decisions rationally, e.g., when in time constraint or when one has insufficient or too complex information (Tversky and Kahneman, 1974; Acquisti et al., 2017). An alternative model is that of nudging (Thaler and Sunstein, 2009), which is an empirical approach to behavioral change exploiting the automatic, heuristic-based, intuitive thinking of (Kahneman, 2011).

Even if nudging for privacy is debatable because, e.g., it may restrict the individual’s autonomy (Renaud and Zimmermann, 2018) as it uses psychological mechanisms covertly, functioning unconsciously, nudging may be considered ethical as long as it is used in people’s best interest; presuming one knows what is actually in peoples’ best interest (Hausman and Welch, 2010). Having an ethical approach to nudging, and not use it for commercial purposes (Sunstein, 2017; Thaler, 2018), one can build a choice architecture that leads people to doing the right things and to carrying out the right activities that lead to the right decisions even
if they are not aware of what they are doing. However, there are many recent examples of consent forms that are GDPR compliant, but still nudge users to pick the privacy-intrusive choices, see e.g., cookie-banners (Matte et al., 2020; Sanchez-Rola et al., 2019). It is difficult to control by law or regulation the use of dark patterns (Waldman, 2020).

**OP.5:** How can PL be a privacy nudge instrument to use for helping people make more privacy-conscious decisions when choosing a product?

One starting point can be found among the existing works on using nudging for privacy purposes (Wang et al., 2014; Acquisti et al., 2017).

5. Automation and tools for creating PL

One way for having PL legally binding is to tie them to privacy policies (previously included in the Terms of Services, or ToS). Machine learning and formal reasoning methods can be used to build tools to help translate (more or less) automatically between privacy policies and PL.

**OP.6:** How can Privacy Labels and privacy policies be correlated?

One starting point for research and tools relevant for translating between privacy policies and privacy labels can be found in the area of logical/legal reasoning and controlled natural languages. One of the more advanced tools is the Contract Verifier\(^9\) (Camilleri and Schneider, 2017; Camilleri et al., 2018), using several different technologies and off-the-shelf tools (see architecture in Fig. 5), and taking input a normative document written in English. Using a standard natural language parser for English (e.g., the Stanford parser), a tabular view of the different clauses in the contract is generated. After having a tabular interpretation, everything is automatic. A formal model can be extracted and used to do queries, e.g.: what are all the obligations in the contract; whether there are obligations without deadlines; what is the data processing purpose; or which party the data is being shared with.

\(^9\)http://remu.grammaticalframework.org/contracts/verifier/
Constructing PL from privacy policies using AI is a two-step process: (i) use natural language processing (NLP) to identify privacy concepts in the privacy policy text and (ii) apply a set of rules to the identified concepts to generate the relevant labels.

One starting point is the award-winning tools from (Harkous et al., 2018). One of the few datasets that we consider to be useful for such purposes is the OPP-115 Corpus (Wilson et al., 2016), which contains 115 privacy policies that were manually annotated by law students and then checked for inconsistencies. Each privacy policy is segmented and each text segment is annotated with privacy concepts.

Having PL correlated to privacy policies using NLP can be used, e.g., to compare a company’s PL to their privacy policy to see whether they match or find discrepancies between the two. One could also convert the extracted privacy concepts to simplified natural language, thus making summaries of privacy policies.

6. Visual Appearance of PL

The appearance of PL is all about conveying information to the target group. We have much to learn from the fields of Information Design and Visualization (Mollerup, 2015; Ware, 2021; Cairo, 2013), but we need to expand beyond the content being presented, to include psychology so to reach the individualized PL from Fig. 4 and to make the looks of PL useful for the educational purposes mentioned in Section 4.

Privacy icons are important for conveying information on a first level of detail (Holtz et al., 2011). Privacy icons are useful for a “nutrition facts” style of PL, e.g., (Kelley et al., 2009), this approach being taken by all the privacy labels discussed below. However, an additional important part of PL would be a comparable view, in the style of energy consumption labels, involving graded scales, with the “nutrition facts” design and privacy icons appearing only beneath the grade.

OP.7: Evaluating the degree of data protection is complicated, but needs to be made measurable, at least for some of the privacy aspects, and fit into a graded-scale system.

6.1. The layered information of PL

We expand on the concept of “layered notices” as promoted by the (Article 29 Working Party, 2018) and include also graded scales, e.g., inspired by energy consumption labels. Our envisaged scenario is the following. First the user is presented with a privacy grade on a scale from A to F. The user can then click on the grade and be shown further minimum of information (following Art. 29 Working Party) and icons to explain the most important parts of the agreement, maybe color-graded to explain how these contributed to the overall privacy grade. Finally, the user can click on the simplified natural language to be referred to the section(s) in the privacy policy the statement was constructed from.

OP.8: How can a complex privacy policy be mapped into a hierarchical A-F scale, while remaining relevant for a particular context and individual user?

Even though it may be easier to grasp, a simple grade can be misleading if it is not given in a context (e.g., the type of app, the type of application domain, the role the user takes wrt. the application being evaluated). Moreover, the aggregation of the evaluations of the different aspects of a privacy policy is not easy since a policy may be more privacy-friendly in some aspects but not in others/all. Even more importantly, we want individualized privacy labels (as explained in Section 3.1 and pictured in Figure 4) because people have different privacy preferences (reflected in their privacy personas). This implies that a privacy label grade B for Alice may be perceived like an F for Bob, and hence the PL has to dynamically change based on the privacy person it is being coupled with (i.e., presented to). One set of measuring scales can come from criteria regarding the usability of privacy as defined in (Johansen and Fischer-Hübner, 2020).

G.8: The goal with “layered PL” is to give the user a bird’s-eye view about what privacy aspects are included in the privacy policy. Then if the user wants more information, she can drill down to a deeper level. At the bottom layer, one can find the whole privacy policy.
Another proposal of privacy labels focuses on conveying in a concise and precise way on a single label, privacy-relevant information specifically chosen for some target group (Kelley et al., 2009; Railean and Reinhardt, 2018). One other source of inspiration can be the layered approach used in cookies notices (Sanchez-Rola et al., 2019; Matte et al., 2020), which people might already be accustomed to.

7. Concluding with a Timeline for Finding PL

Before concluding we devise an action plan for finding PL.

**Phase I:** Attract companies and organizations to perform self-assessments, based on their own privacy policies, using a structuring tool such as the NL.PL.

**Phase II:** Bring in more of the automation and reasoning tools for creating privacy measurements, to identify the level of privacy protection to be shown by PL. Include psychological models to individualize PL, and give also a proper appearance for the intended purposes, one of these being educational. Aim more on generating the PL automatically from reliable documents like the privacy policies and from inputs such as those from a user’s privacy profile.

**Phase III:** Join forces with the authorities for introducing a regulatory framework to make the self-assessment mandatory and uniform. At the same time, build a community around PL, that could even include the authorities.

The self-evaluation is a logical first step from a GDPR perspective, because of the large emphasis on demonstrating compliance. The controllers need to show how they abide by the GDPR and how they implement the data protection principles. Further on, one can introduce checks and fines, including external audits and certifications, done by the authorities or the community.

Concluding remarks

When dealing with a complicated concept such as privacy, that is faced with multiple long-standing problems as discussed in Section 2, then to develop a solution can be a daunting task. In response, we propose an all-encompassing definition of Privacy Labeling as a possible start on the road towards a solution to many of the current privacy problems our society is struggling with.

When developing such a panoptic concept as the Privacy Labels proposed here, it is a good idea to find many discussion partners among the various stakeholder groups. Therefore, the ideas that we have presented have roots in our conversations with experts from the seven different fields that we considered relevant for privacy labeling. Our goal with bringing in all these views was to investigate the concept of privacy labeling (and its implications) from many different angles. No one single discussion partner has ‘the right answer’, but their collective opinions sum up to a result that is much more comprehensive and powerful than any one view could accomplish on its own.

References


Ware, C., 2021. Information visualization: perception for design. Morgan Kaufmann.


A Multidisciplinary Definition of Privacy Labels: The Story of Princess Privacy and the Seven Helpers


*Dept. of Informatics, University of Oslo
bBjørknes University College
cDept. of Mathematics and Computer Science, Karlstad University
dNorwegian University of Science and Technology
eDept. of Computer Science and Engineering, University of Gothenburg
fPrivacy Company
gUnabhängiges Landeszentrum für Datenschutz Schleswig-Holstein
hDept. of Technology Systems, University of Oslo

Abstract
Privacy is currently in distress and in need of rescue, much like princesses in the all-familiar fairytales. We employ storytelling and metaphors from fairytales to make reader-friendly and streamline our arguments about how a complex concept of Privacy Labeling (the ‘knight in shining armor’) can be a solution to the current state of Privacy (the ‘princess in distress’). We give a precise definition of Privacy Labeling (PL), painting a panoptic portrait from seven different perspectives (the ‘seven helpers’): Business, Legal, Regulatory, Usability and Human Factors, Educative, Technological, and Multidisciplinary. We describe a common vision, proposing several important ‘traits of character’ of PL as well as identifying ‘undeveloped potentialities’, i.e., open problems on which the community can focus. More specifically, this position paper identifies the stakeholders of the PL and their needs with regard to privacy, describing how PL should be and look like in order to address these needs. Throughout the paper, we highlight goals, characteristics, open problems, and starting points for creating, what we consider to be, the ideal PL. In the end we present three approaches to establish and manage PL, through: self-evaluations, certifications, or community endeavors. Based on these, we sketch a roadmap for future developments.

Keywords: privacy labels, General Data Protection Regulation, usability, certification, automation, psychological models

1. Introduction

The right to privacy is something precious and frail (an integral value appearing in the Universal Declaration of Human Rights1), which we need to take good care of in order not to lose it, much like princesses in fairytales. Just like European royalties, privacy is known to people only as a symbol, but does not have

---

1 The “right to privacy” emerged in the Universal Declaration of Human Rights, adopted in 1948, as one of the fundamental human rights. Shortly after, this right was reaffirmed in the European Convention on Human Rights (ECHR), drafted in 1950.

Preprint submitted to A Journal May 11, 2021
much power in the economy or society. In its current state it does not always serve the people, but mainly a few very wealthy and influential actors prosper from its misuse. Loss of privacy has both micro implications, at a personal level (e.g., people being influenced to buy what they do not want or need (Matz et al., 2017), to vote for extremists (Isaak and Hanna, 2018; Berghel, 2018; Stewart et al., 2019), or to develop antisocial behavior), but also macro implications, at a societal level (e.g., a society living in fear of being watched by surveillance capitalists (Zuboff, 2019) or manipulated on social media (Starbird, 2019; Grinberg et al., 2019)). Privacy is personal and contextual, having social and political ramifications, but most of the population does not see, or understand, even some of its basic implications. The lack of privacy literacy can be partly attributed to commercial entities that often, while profiting from handling data, work hard to keep privacy “out-of-sight [is out-of-mind]” – like a sleeping princess locked in a tower – e.g., telling people infamously “You have zero privacy anyway. Get over it.”2 (Solove, 2011). Privacy misapprehension by the population is also due to its complexity, having kept many brilliant minds preoccupied for at least a century, since photography as a new technology used by media became widespread (Brandeis and Warren, 1890). In the current digital society, privacy (Solove, 2004; Acquisti et al., 2007) has even stronger forces compounding its complexity, coming from, e.g., technological advances in miniaturizing hardware that enabled cheap privacy-invasive gadgets, powerful algorithms that can make inconceivable inferences (Schneier, 2015), or supercomputing in ‘invisible’ clouds (Borning et al., 2020); all too complex for laypeople to grasp. It is fair to say that against such rapidly changing technologies, a person alone, no matter how dedicated she may be, would find it impossible to protect her Princess Privacy.

This paper analyzes how the concept of Privacy Labeling/Labels, which hereafter we refer to as PL, could contribute to resolving several of the challenges that privacy is currently facing. Given the many facets of privacy and its society-wide implications, it is important to adopt a multidisciplinary approach. This work started from a workshop in spring 2020 where experts from different fields of practice and research gathered to present and discuss their views on the topic of PL. We thus bring in the following perspectives:

- Business (relevant topics including, e.g., market potential, incentives, social responsibility, added value);
- Law (e.g., compliance, privacy policies, audit);
- Regulations (e.g., national, European, implementations, domain-specific standards);
- Usability and human factors (e.g., personas, easy to understand, completeness, contextual);
- Education (e.g., psychology of people, of SMEs (Small and Medium-sized Enterprises), of CEOs (Chief Executive Officers), nudging for good, mental heuristics);
- Technology (e.g., AI, reasoning, automation, dynamic labels, verification);
- Multidisciplinary (e.g., communication across fields, people, or companies, making synergies).

We elaborate on how to combine the seven different perspectives, the roles and priorities of each of these in relation to PL, and point to the state of affairs in the respective fields.

Since it is rather intricate to provide a completely comprehensive picture of PL, we chose a storytelling style of discourse, and use the story of “Snow White and the Seven Dwarfs” as our parable. This inspired us to employ metaphors such as the ‘seven helpers’ as an analogy for our seven perspectives and ‘Princess Privacy’ as the one to be saved by the ‘Privacy Labeling Knight’. We give each helper a name, and we use it to mark parts of the text with the respective perspective it belongs to:

- **Bussy** – bringing in business arguments,
- **Lancey** – bringing in the legal perspective/argument,

---

2https://www.wired.com/1999/01/sun-on-privacy-get-over-it/
• Reggy – bringing in the regulatory perspective/argument,
• Upsy – bringing in the usability perspective/argument,
• Eddy – bringing in the educational perspective/argument,
• Techy – bringing in the technological perspective/argument,
• Multy – bringing in the perspective/argument of multidisciplinarity.

We define the concept of Privacy Labeling\(^3\) below, and throughout the paper we detail each of its elements.

A Privacy Label is a legally binding label containing information about the privacy that a product or service provides. The labels may be physical or digital. They are defined, and are visually presented, in a layered manner, where one can drill-down from a simple overview to more complex information, to allow the user to focus on different levels of detail, depending on the intended use. The labels measure privacy using graded-scales to make it easy to compare two labeled products with respect to privacy aspects relevant for a particular (type of) user.

More specifically, one can imagine PL as being similar to both nutrition facts labels and energy consumption labels. To make PL legally binding one can tie it to a privacy policy text, so that it cannot become a means of deceit in the hands of product advertisers, thus going beyond, but not against, laws and regulations such as GDPR.\(^4\) PL are promoting “privacy as an added value” to a digital product,\(^5,6\) allowing privacy conscious businesses to differentiate themselves from those market competitors that prefer to monetize on the big-data model at the expense of the privacy of the user\(^7\). As such, Privacy Labels should:

(i) be educational (“Oh, there’s a notion of privacy for this TV-set!”),
(ii) be an incentive and promote business differentiation (business slogans could sound like: “We care about your privacy. So should you!”),
(iii) be legally conscious yet
(iv) be usable for the layperson (“Hey son, what’s all this writing about privacy here?”), hence with sufficient detail as needed, yet visual and simple,
(v) be taken up into regulations and
(vi) supported by technologically innovative tools.

Including all these characteristics requires a multidisciplinary effort.

Apart from defining PL and discussing it from all the seven relevant viewpoints, we propose along the way: goals, characteristics, open problems, and starting points for research. Further contributions can be summarized as follows (and are schematically presented in Figure 1).

- We identify how the present landscape of privacy certifications (including privacy seals and marks (Rodrigues and Papakonstantinou, 2018)) could be improved by PL (see Section 2).
- We identify the stakeholders of PL, what are their needs and characteristics, as well as the relation between them (see Section 3).

\(^3\)A 90 seconds ‘elevator-pitch’ video where we present Privacy Labeling for a general public can be viewed online at: https://youtu.be/noE_vF2_GEs.

\(^4\)The European General Data Protection Regulation (GDPR) (GDPR, 2016).

\(^5\)https://www.nbcnews.com/tech/security/can-privacy-be-big-business-wave-startups-thinks-so-n1128626


\(^7\)ENISA. “The Value of Personal Online Data”. https://www.enisa.europa.eu/publications/info-notes/the-value-of-personal-online-data

3
Three important aspects of PL are presented: in Section 4, its educational power to change people’s knowledge of privacy; in Section 5, tools useful for constructing PL; in Section 6, the possible visual appearances of PL.

Three approaches to obtaining PL are presented in Section 7 and a roadmap for achieving PL is outlined in Section 8.

The above listed contributions constitute the major points of debate for the seven helpers. Imagine them sitting around the sleeping Princess Privacy discussing how to find a ‘Privacy Labeling Knight’ with the traits of character needed to awaken the princess. As the stories often depict it, not any knight will be right for the task. Therefore, throughout the paper we point out:

- goals (numbered as G.x and marked with a flag icon /flag_/) intended to be achieved with the help of PL;
- characteristics that we think PL should have (numbered as C.x and marked with a diamond icon /Diamond_/);
- open problems (numbered as OP.x and marked with a scholar cap icon /Cap_/) that the community can address while striving for reaching any of the above;
- existing works that can function as good starting points (marked with a map-pointer icon /Map_/) for some of the above.

Following the practice of “eating our own dog food”, we mark the identified goals, characteristics, and open problems, with icons to make these important contributions of this paper more accessible (see in Section 6 our discussions on the use of icons in PL).

We believe that collecting all the /Characteristics/ would form an ideal PL that could be useful to attaining the /Goals/goals that we have pointed out. To help the community work towards creating such a PL we outline several /Open Problems/open problems and also identify good /Starting Points/ among the existing works, along with drawing, in the end of the paper, a general roadmap to follow while taking one (or more) of the three approaches that we propose for managing PL. Since all of these are the results of the dialogue and agreements between the seven helpers’ different viewpoints, we expect the acceptance and usefulness of PL within the society to be considerable.

In the next section, the seven helpers examine why Princess Privacy is asleep, what is the cause of this present dark situation and how to make the future brighter for their princess by describing the impacts and benefits of PL detailed in the rest of the paper. The structure of this paper is displayed in Figure 2. An example that we use throughout the paper is the PrivacyLabel.org, which we hereafter refer to as NL.PL.

https://www.privacylabel.org/project from the Netherlands.
2. Sleeping Princess Privacy and the Privacy Labeling Knight

Multy: If Privacy Labeling is to be the ‘knight savior’ we need a multidisciplinary effort combining the positive forces from the seven domains mentioned previously. Compared, e.g., with energy consumption where one number is enough, maybe placed on a graded scale for visual comparison, and measured with well known instruments, privacy involves multiple concepts (thus we have Reggy and Lancey with us) spanning from social to technological, most of which we do not know how to measure (here Techy could help), definitely not in a universal way since they are relative to a person’s view on privacy (which Upsy can tell us more about) influenced by this person’s level of knowledge about privacy (which Eddy is preoccupied with). It is generally known that some big-tech companies are monetizing on this ignorance, but not all, usually not the SMEs, many of which would like to be able to promote their privacy consciousness as a value added to their products (isn’t that right, Bussy?). Our PL knight must go through a series of challenges to prove himself (these are what we mark as Goals) and will have to build up a set of skills (marked as Characteristics or Open Problems), before the seven helpers can deem PL worthy.

2.1. Evaluating the Situation of Princess Privacy

Multy: First we want to make clear one distinction, because people often confuse privacy with security (Techy: programmers do this quite often). Reggy: Although multiple standards and certifications currently do not make such clear distinctions (still looking mostly at security), security should only be a baseline, e.g., GDPR considers security as one of its several data protection principles (see Art. 5 I (f)). Privacy protection goals include the classical security protection goals confidentiality, integrity and availability (CIA), and in
addition also privacy goals such as transparency (Murmann and Fischer-Hübner, 2017), intervenability and unlinkability that go beyond CIA (Hansen et al., 2015). Upsy: Moreover, when looking at the attitudes of the users there are clear differences between security and privacy, due to individual differences (Egelman and Peer, 2015).

Bussy: If in security it is often said that the weakest link is the user, in privacy we see that the weakest link is the controller. Examples of privacy breaches for which the controller is responsible can be: the controller “tricks” the users into giving more data than the user is aware of, often through hiding information or by using privacy-invasive approaches known as “dark patterns” (Bösch et al., 2016; Mathur et al., 2019; Nouwens et al., 2020); lack of legal competence when drawing contracts with third parties; programming incompetence incurring leakage of data, e.g., usage of third party libraries; or not investing in measures for preventing security leakages (Palombo et al., 2020).

Reggy: The Recital 100 of GDPR encourages “the establishment of certifications mechanisms and data protection seals and marks [to allow] data subjects to quickly assess the level of data protection of relevant products and services” (GDPR, 2016). While Art. 42(1) encourages the implementation of certification and data protection seals for demonstrating compliance by accredited certification bodies, PL should go beyond and measure on a scale how well the privacy is respected and how easy is for a user to understand that (see also G.2).

G.1: One Goal is to build PL on/into existing certifications.

Reggy: Examples of existing certifications include: Datenschutzgutesiegel, granted to systems and products by ULD (The Schleswig-Holstein Data Protection Authority)9, EuroPriSe10, Common Criteria11 (ISO/IEC 15408) including a Privacy Class meant for defining privacy functionality, focusing on aspects such as anonymity (Elliot et al., 2018), pseudonymity, unlinkability (Madaan et al., 2018), unobservability (Pfitzmann and Köhntopp, 2001). Lancey: Some of these are partly required by law, e.g., ULD Datenschutzgutesiegel is used for public procurement in the Schleswig-Holstein German state, whereas Common Criteria are taken up in eIDAS (electronic IDentification, Authentication and trust Services) EU Regulation No 910/2014, and partly required for certain procurements in certain public sectors.

Reggy: For privacy certifications we also need evaluations. The challenge is that the focus of schemes such as the above is much on security testing and penetration testing. There is a need for more formal evaluation, verification, or testing of privacy requirements (as Techy can soon tell more about). Upsy: Aspects of usability should also be included in the evaluation of privacy. Parallel this with security where the weakest link is often the end-user. Nowadays communication protocols are formally proven secure, but still security breaches occur because the user interfaces are not usable, leading end-users to doing mistakes, e.g., the security warnings for SSL certificates or other types of security warnings where the end-users have to make decisions without good guidance or usable instructions, e.g., Whitten and Tygar already in 1999 tested Pretty Good Privacy and revealed several usability issues that lead to insecure decisions or that the encryption products/features were not used at all (Whitten and Tygar, 1999).

Upsy: In addition, privacy (like security) is usually only a secondary task for the users (Whitten and Tygar, 1999), e.g., when buying train tickets with a ticket-app the primary goal of the user is not to check how well the app protects her privacy, but to reach a certain destination. In addition, it is arduous for a regular person to keep track of all the electronic data that she is generating, given that many activities nowadays are happening online. It is even more difficult to know exactly which effect this data has on our privacy, because of the modern machine learning algorithms that can make inferences based of apparently non-private pieces of data (Rader et al., 2020; Acquisti et al., 2017).

Upsy: To overcome such user/usability related challenges, one has to make the privacy related measures usable. Lancey: The GDPR is a good place for finding examples of usability goals, e.g., “communication ... relating to processing [to be provided] to the data subject in a concise, transparent, intelligible and easily accessible form, using clear and plain language” (Art. 12); with 30 more such usable privacy goals

9https://www.datenschutzzentrum.de/guetesiegel/
11https://www.commoncriteriaportal.org/
identified in (Johansen and Fischer-Hübner, 2020). However, the usability goals appearing in GDPR are too general, given the inherent nature of the GDPR (and laws in general), which in this case allow for too much subjective interpretation by controllers for their own interest. One classical behavior is to make the privacy settings blend into the background, or even worse, the button for the privacy invasive option is highlighted by design; e.g., when emphasizing the ‘Accept’ button for the privacy policy, the users will give their consent without reading. This is the case with the privacy policies as well, which can containing information manipulated in a way that the user on average will not be well informed, though they formally meet the requirements of the GDPR (Karegar et al., 2020; McDonald and Cranor, 2008). Therefore:

**G.2:** PL should offer a way to reach the usability goals of GDPR.

*Upsy:* Usability and Human-Computer Interaction (HCI) techniques have mostly been developed for making technologies that are difficult to use, or made for highly specialized experts, more easy to understand and interact with, both for the expert user, but often also for new, less expert users. Usability is even more important for privacy since privacy is a highly complex concept, related to complex technologies such as AI and Big Data, but which is especially addressed to the laypeople, not to experts, because privacy is a human right.

*Lancey:* As shown by (Patrick and Kenny, 2003; Patrick et al., 2003), it can however be challenging to map legal requirements into HCI requirements. *Techy* agrees that for the programmers as well it can be difficult to implement the statements made in the privacy policies or regulations. Therefore, help is needed in bridging the gap between regulations or legal documents (such as privacy policies) and the software/technology that these talk about. Those needing support in this case being the lawyers, interaction designers, and the programmers.

*One starting point* for evolving certification schemes from seals and trust marks towards privacy labels of the energy efficiency type, i.e., aiming for goal G.1, is by measuring the usability of privacy using HCI methods, thus covering also G.2. (Johansen and Fischer-Hübner, 2020) works in this direction by proposing a set of criteria thought to produce measurable evaluations of the effectiveness, efficiency, and satisfaction with which privacy goals of GDPR are reached. This work extends the methodology of EuroPriSe certification scheme by adding, what is called, usable privacy criteria. Thus, the EuroPriSe certification assures that the GDPR legal ground is covered, including data protection principles and data subject rights, while the usable privacy criteria come on top, fine-graining the EuroPriSe evaluation with usability measurements showing how well the legislation is respected. All these are organized and visualized as a cube, called “Usable Privacy Cube”, composed of three variability axes containing: usable privacy criteria, rights of the data subjects, and privacy principles. This work has identified from the GDPR text 30 usability goals, which have been used as guidelines to define 23 usability criteria, each composed of several subcriteria designed to measure usability aspects (Johansen and Fischer-Hübner, 2019).

In ergonomics and HCI, the context of use is an important component of a usability evaluation. Consider the definition from the ISO standard 9241 (ISO9241, 2018): “The context of use comprises a combination of users, goals, tasks, resources, and technical, physical and social, cultural and organizational environments in which the system or service is used.” The context of use is translated into GDPR vocabulary as the *context of processing*. Recital (71) of GDPR states “In order to ensure fair and transparent processing [the controller should take] into account the specific circumstances and context in which the personal data are processed [...]”. An example of a context eliciting question is “What data is the product/system processing?”, which elicits information about type, volatility, accuracy, size/amount, persistence, value of the data. Creating guidelines for how to establish the context of processing is a necessary enhancement of the above work.

**G.3:** PL should reflect and communicate the context of processing.

### 2.2 The need for a Privacy Labeling knight

*Multy:* Besides being adaptable to different contexts, the PL knight should also be trustworthy (hence G.4) and economically inspiring (hence G.5).
Reggy: In 2013, the European Consumer Centres’ Network published a trust mark report “Can I trust the trust mark?” where it is brought to the attention the importance of establishing reliable trust and demanded a more uniform practice of European trust marks.

Upsy: The research done in projects such as PRIME (Camenisch et al., 2011) and PrimeLife (Fischer-Hübner et al., 2011) has shown that the end-user is having a lack of trust especially in the case of PETs (Privacy Enhancing Technologies) (Alaqra et al., 2018). They often have difficulty believing the claims made by the PETs that privacy can be really protected in that way, often because these are counterintuitive. Techy: PETs are not trusted or understood because they are based on cryptography and cryptographic schemes do things that are not easy to grasp. The user testing in PrimeLife showed that people had difficulty understanding and believing the concept of data anonymization via zero-knowledge proofs. There are also no good real-word analogies/metaphors that can be used to mediate these functionalities, which seem to be counterintuitive for users (Wästlund et al., 2009).

G.4: PL should be developed and applied uniformly so that it becomes an important trust factor. 

Upsy: Interviews with stakeholders involving privacy enhancing data analysis on encrypted data (homomorphically encrypted data) were done in the PAPAYA project. The scenario was that ECG (electrocardiography) data were sent to the cloud for data analysis. The ECG signals were encrypted, while the analysis was taking place only in encrypted form. In expert interviews, the more technically skilled users showed skepticism towards this form of analysis. The expert users had requirements to have assurance guarantees that data analysis on encrypted data really worked. When shown a privacy impact assessment (PIA) according to the tool from the French Data Protection Commission (CNIL) to increase trust, they also wanted to have complementary information about the PIA method and process, and qualification of the evaluators. This shows, for the case of expert users, the importance of privacy claims for establishing trust in PETS (Alaqra et al., 2020).

Bussy: Increasing customers’ trust can also be achieved by showing that the organization takes privacy seriously, by displaying privacy information that can be understood, instead of a long legal text. Privacy governance can in this way become a competitive asset, an unique selling proposition (Hoffman, 2014).

G.5: PL would facilitate the inclusion of people in the new data economy.

Bussy: There are many reasons (and controversies) for including people in the new data economy (Jentzsch et al., 2012; Acquisti et al., 2013; Spiekermann and Novotny, 2015; Acquisti et al., 2016; Li et al., 2017; Benndorf and Normann, 2018; Malgieri and Custers, 2018). To do so, one needs, besides trust, to consider how well the consumers are informed and how aware they are of the existing options. PL would achieve this by implementing, in a more accessible manner, the transparency principle of GDPR (Art. 5 I (a), 11, 12), which requires data controllers to inform their users about, among other, what data is processed, for which purposes, the legitimate requirements of the processing, or who are the recipients. Having access to such information, the consumers can make more informed choices. We wish to empower people to gain insight and control to make informed choices and comparisons. The consumer can then become part of the data economy not only as an asset, but also as a stakeholder that can influence the market.

3. PL stakeholders and their needs

Multy: Privacy labels can have different purposes (e.g., for internal use, for showing compliance or only for fulfilling transparency requirements, for marketers to use for selling effectively/targeted) and be intended for different audiences (e.g., data subjects or the controllers). However, it is probably difficult to put everything in one label.

C.1: One Characteristic of PL is to be usable, for different purposes, by different types of stakeholders.

---

13http://primelife.ercim.eu/
14PlAtform for PrivAcY preserving data Analytics https://www.papaya-project.eu/
As any knight who is evaluated on his traits of character, C.1 is only the first of many more characteristics to be argued for in the rest of this text.

**OP.1:** An Open Problem is how to make the concept of PL flexible enough to accommodate different purposes and audiences, and for each such PL instance how should it be designed in order to convey the intended information to the intended audience.

The rest of this section surveys the various audiences and purposes PL may have, starting with the “convenience users” being our primary target, and continuing to discuss the needs and expectations of businesses, lawyers, regulators and authorities, and programmers.

### 3.1. Convenience users

We define “convenience users” as those people that nowadays trade in their privacy for convenience, most often without knowing what they are trading in.

**Bussy:** The convenience users are much of the time running on ‘autopilot’ when they are making judgments, e.g., when shopping online. This happens from multiple reasons, e.g., willpower depletion (Baumeister and Tierney, 2011), heuristic and intuitive thinking (Kahneman, 2011), or manipulations such as priming done through media channels and advertising (Cialdini, 2007; Thaler and Sunstein, 2009; Harris et al., 2009). Advertisers and commercial businesses use extensively behavioral psychology to influence, whereas the governments and authorities seem to assume that people are rational and as such they do not see the need to push any psychological buttons. **Reggy:** The rational behavior assumption seems to be the case also with the current certification schemes, which are based on a rational model. They state that privacy is inherent in the product or the service as a sort of objective measurement, and they are made on the premise that one size fits all.

**Eddy:** Knowledge is power, and our PL knight can help increase the privacy knowledge of laypeople, thus empowering them to make well informed choices in the technological world and thus actively participate in the data economy. Convenience users would get in contact with the PL as a result of being interested in buying a digital product or using a new digital service that is collecting personal data (most of which do).

**G.6:** PL as an educational tool to increase privacy literacy in the general population.

**Eddy:** At the same time, we also know that people have different values and traits of personality (Fig. 3; see also the original Big Five model (Allport, 1937)), and that different people may prefer different levels of privacy, that may also change with time and context (Westin, 1991; Knijnenburg et al., 2013; Gerber et al., 2018). In addition, psychological studies show that people are not always able to make choices or judgments that are in their best interests, as e.g., with overly confident people that jump to conclusions without the necessary due diligence or due to anxiousness, which makes one avoid making decisions (John...

![Figure 3: Personality traits, from Chap. 30 of (Howard, 2014).](image_url)
et al., 2010a,b). These differences in people’s personality traits are susceptible to different kinds of influences (Acquisti, 2009; Warberg et al., 2019).

**Eddy:** People also have different cognitive styles. Some have an intuitive approach to making judgments and decisions about something, while others have a more analytical approach (Egelman and Peer, 2015). We also know that people’s judgments are very much influenced by their current emotional state, their affect, their moods (Kitkowska et al., 2020a), e.g., car salespersons are known to try to get the buyer in a good mood, in order for the buyer to be less critical. Whereas when in a bad mood or on defense, one not only becomes captious, but also more analytical (Peters et al., 2006). Combining the personality and cognitive style where a person is not confident and is not prone to making analytical judgments, but instead has an intuitive cognitive style and also is in a good mood, is what might be the characteristic of the typical convenience user. In this case it might be necessary to push other types of buttons to slow the cognitive processing of this type of users down, so that they can really think about what they’re doing, instead of processing and making judgments intuitively and heuristically.

**Bussy:** The concept of social influence that is a part of social psychology, describes how people may be influenced by different agents, in different ways, with different means, or for different purposes (Cialdini, 2007; Argo, 2020). A person buying an app that was advertised on the subway, is an example of social influence that does not always serve peoples’ best interests, but usually serves commercial interests. **Reggy:** Instead, governments or institutions (e.g., independent supervisory authorities), might be having a more ethical approach to influencing people to make decisions that are in their best interests. **Eddy:** If people are not aware and they run on cognitive autopilot, then even if the person is not capable of making an analytical judgment or decision, that serves his or her best interests, they can nevertheless be pushed in the right direction, if the ones that steer have the respective people’s best interests in mind.

**C.2:** PL should be individualized by considering the psychology of personality, cognitive styles, and social influence (cf. Fig. 4).

*One starting point* is the combination of the elements from Fig. 4 that would result in a more nuanced distribution of privacy preferences and attitudes. **Upsy:** The literature has identified several of such predominant profiles known as privacy personas (e.g., information controllers, security concerned, benefits seekers, crowd followers, and organizational assurance seekers) (Westin, 1967; Morton and Sasse, 2014; Woodruff et al., 2014; Dupree et al., 2016). Rather then adopting an exclusivist and reductionist approach, the PL should be able to adapt to different privacy preferences or privacy personas.

---

15A *persona* is a precise description of the user of a system and what she wishes to accomplish. Though a persona is not a real person, it is created based on synthesized characteristics and needs of real people, and it represents the profile of a typical user (Cooper, 2004).
3.2. Businesses

Bussy: PL could provide a competitive advantage (Martin and Murphy, 2017), especially in markets such as Europe, where privacy protections are required by law. Even in privacy unregulated markets, a company that has visible and easily understandable statements of compliance will be perceived by the customers as one that takes privacy seriously. This has the potential to increase the customer’s trust in the company, which is seen in (Bachlechner et al., 2020) as essential for data-driven businesses.

Bussy: However, against the possible benefits one has to weigh the possible costs. One factor to consider is the price of obtaining and maintaining a PL, e.g., how often re-certification is required or whether the PL will reduce or increase the use of other resources, e.g., compliance officers, additional IT staff, preparing more documentation, or legal council. Another important question would ask what is the scope of the label and what does it cover, coupled with an evaluation of how time consuming is to obtain and maintain a PL or, alternatively, how much time would it save on other business aspects such as marketing or customer retention. Since the benefits of a label also depend on its reputation, business are interested in, e.g., how widely known is the label, whether it is already recognizable, and if not, whether it will be so in the near future.

G.7: PL should become something worth investing in, that brings clear benefits and offers a competitive advantage, outweighing the costs.

Bussy: Businesses tend to prioritize what is most requested by their customers. Recent research shows that privacy is a major blocker for the adoption of new technologies, e.g., (Barbosa et al., 2020) studied people’s considerations for adoption of smart home devices, and found that half of the 613 participants have named privacy or security concerns/risks as being a blocker for them acquiring an IoT device. Privacy was also ranked second, after ‘convenience’, as being considered when purchasing such devices.

Lancey: Depending on the needs of the business, e.g., in which territory they plan to distribute a new product, the privacy related documentation is adjusted considering the costs and benefits. The level of detail of a privacy policy (and of the related PL, cf. C.4 & OP.9) depends on, e.g., what the local regulations require, but also on how complicated the processing of personal data of the respective company is, restricted by the resources available to the company (e.g., to pay lawyers or certification processes). Since privacy policies can vary widely, it is important to consider the flexibility of PL and create different modules that can adapt to more complicated needs.

C.3: PL must be modular and flexible to accommodate the different needs of businesses in relation to local regulations and their customers’ demands.

Lancey: Together with stricter regulations such as GDPR (Tikkinen-Piri et al., 2018) and peer pressure, e.g., Google Play requiring all apps to have a privacy policy before being allowed on the app store, privacy has become a more important topic for businesses. Irrespective of whether it is a large multinational company or a start-up, every business that deals with digital data nowadays has to have a privacy policy.

Multy: Previously, a common practice was to just copy and paste privacy policies, e.g., a new start-up wanting to provide something similar to an existing service, say like Github, they would copy the privacy policy from Github, maybe trying to read it themselves and maybe changing a few things. Lancey: However, nowadays standardized privacy policies is an obsolete practice, because complying with requirements like GDPR implies providing valuable information to the user. This information cannot be general, but has to explain what the company is actually doing with personal data. The lawyers use considerable time talking with the client (who is the controller), investigating, making sure that they completely understand their operations and what is necessary, in order to make a meaningful privacy policy.

C.4: PL should be closely related to privacy policies, which the law asks all businesses to have.  

---

16This practice has proliferated to the point that now one can find ToS generators and ToS templates as services online, see e.g.: www.termsandconditionsgenerator.net or getterms.io or privacyterms.io for generating various kinds of online agreements including privacy policies, and www.termservicetemplate.com or www.privacypolicies.com/blog/sample-terms-service-template for templates; and www.termsfeed.com/blog/terms-conditions-copyright-law for arguments and discussions around such services.

17One can also consider other privacy related documents or certification processes.
One starting point in the process of making a meaningful privacy policy could be to employ the support from a tool, such as the NL.PL (see Section 7.1 for more details), where one is guided into providing the information required for making a PL (and thus also a privacy policy). Similarly to what a lawyer would do, one still needs to do research ahead to extract the needed information to be used with such a tool.

Bussy: The NL.PL focuses on facilitating more transparency and clear communication by making the privacy statements easy to understand for customers. A tool like NL.PL can be used by everyone (both businesses and individuals) to have an overview of a privacy statement and be the starting point for both a privacy label and a fully compliant privacy policy. It is especially useful for SMEs, which often do not have the means to pay law firms to create the necessary legal documentation. The businesses can save costs by doing the needed preliminary work by themselves with the help of the tool. To help a privacy officer create a label for their own organization with ease, NL.PL follows a data flow organization model, which is provided with company specific privacy relevant details, e.g., about collected data, such as location and duration, or processing purpose (with predefined example to choose from and edit).

Multy: A more general problem encompassing what has been said until now about privacy policies and businesses adopting PL is OP.2.

OP.2: How can Privacy Labels reach the mass market?

Lancey: It can be said that privacy policies, being required by law for every digital product or service handling private information, have already reached the mass market. By attaching PL to privacy policies, these too would reach all the consumers under the condition that businesses would be willing to adopt PL. Bussy: Therefore, in an unregulated market place, PL has to appeal also to businesses, to bring value to their products. Reggy: Otherwise, regulations and legislation, along with incentives, can be used to attain a critical mass of businesses using PL, at least as a means of making their privacy policies more usable. Bussy: A tool such as NL.PL would also be useful in this respect, as it would make the creation of PL more practical and affordable, through being easy to use without needing legal expertise.

3.3. Lawyers

Lancey: A common situation relevant for lawyers, is when a client is considering obtaining a privacy label. The primary choice is a privacy label that would also involve a certification by an independent certification body, e.g., if the client is established in the EEA (European Economic Area) then it will have to adhere to data protection legislation such as the GDPR and also local data protection requirements. However, PL could also be relevant for establishments outside the EEA, for example if they want to compete with businesses in the EEA. Since there are not many privacy certification options, it is also often that PL are desired rather as information conveying tools. Since privacy legislation (the same as certifications) can vary between different geographical regions (Sullivan, 2019; Kaminski, 2020), one can see PL as a harmonizing factor because of its international nature, managed by a global community (as we detail in Section 7.3). The type of client is important as well, with factors such as the business size, multinational, national, or an SME, combined with the nature of their commercial transactions. A PL can thus be relevant both for B2C (Business to Consumers) as well as B2B (Business to Business).

C.5: PL must take into consideration both commercial and legal aspects, and their interdependencies. 

Lancey: In terms of legal implications, having a PL does not relieve the company of its obligations to adhere to the data protection law. PL is only a modality to demonstrate compliance, as stipulated by GDPR in Art. 42(1). Therefore, besides maintaining documentation relevant for PL, the company still needs to implement technical and organizational measures, which should not be seen as something additional to having a PL, but as part of the requirements necessary for obtaining the PL.

C.6: PL should reflect technical and organizational measures taken by the company.

Bussy: Certifications sometimes have additional requirements that are more onerous than the law. It can be more difficult to obtain the PL through a certification process than to just be compliant with data protection law. This could have potential additional costs but also potential benefits. Having a PL might have implications for other business areas where legal advise is usually needed, as in marketing where the
lawyer has to assist in ensuring that the PL is not misleading or inaccurate, otherwise it can be judged as false marketing.

C.7: PL should come with supporting guidelines for businesses so that they do not include false information unwittingly.

Too often in fairytales, malicious characters take the form of, or pretend to be, the good characters, like Snow White’s stepmother who disguises herself as an old peddler or a comb seller in her attempts to kill Snow White.

OP.3: We should develop a system to distinguish between a false and an authentic PL.

Reggy: Standardization is needed to make services comparable. We wish for a basic way to structure the privacy related aspects in a fixed and similar manner. Such a structure allows also for cross-comparison of labeled products and services. Lancey: Having an overview is always useful for a lawyer as well as for convenience users. It can be sometimes discrepancy in the needs a lawyer might have, e.g., the requirement of being transparent is not always appreciated for some lawyers, because they might want to have room to maneuver in case something goes wrong. However, PL aims to be a standardized and clear approach to presenting information towards customers and consumers. A lawyer or a consultant representing a company is required in this way to simply be transparent and demonstrate compliance.

3.4. Programmers

Lancey: Technology people are struggling to understand the legal terminology and how to implement a system so to conform with the statements appearing in the legislation and in the privacy policies made by their leadership. This “legal-text-to-code” gap is even larger than the well-known gap between software requirements (or specifications) and their implementation. Techy: Besides standard questions that programmers ask, such as “What does data minimization mean?”, one important problem that they face is how to match the “purpose” stated in a privacy policy (and presented by the PL) with the precise usage of the data during any execution of their software implementation. These are necessary questions when trying to enforce privacy or prove compliance with the GDPR or (maybe easier) to own privacy policies. It is already difficult for lawyers at an organizational level to deal with such questions, which become even more complicated to answer when trying to look at the software code.

OP.4: If PL use tools to translate/explain privacy policies to convenience users, we would like to investigate how can these same tools be useful to the programmers to understand how to implement the statements from privacy policies and law.

One starting point can be to try to use existing formal tools to analyze at least the more critical parts of the code by, e.g., doing code inspection. Since it is difficult to analyze code automatically, putting a human expert into the loop can be a more feasible first approach of doing semi-automated code evaluation and verification for privacy compliance. Techy: There exist several recent technological advances on automating particular aspects of privacy, e.g., on data-flow (Antignac et al., 2016); on data minimization (Antignac et al., 2017); on privacy by design (Langheinrich, 2001; Gürses et al., 2011; Hoepman, 2014; Romanou, 2018; Antignac et al., 2018; Schneider, 2018); and in general on Privacy-Enhancing Technologies (PETs) (Danezis et al., 2015). However, it can take long for a research idea to reach the programmers, and even more so for PETs since more often than not, these prove too difficult for software development companies to comprehend, let alone implement or adopt in their software or DevOps tool-chains.

One of the fastest spreading types of software is the AI/Deep learning based software (see the 2019 Turing award laureates excellent overview (LeCun et al., 2015)), with a considerable number of programmers actively involved. In his recent call for AI regulations, Etzioni urges regulators to focus on five critical areas “no killing, responsibility, transparency, privacy, and bias” (Etzioni, 2018). Privacy has earned a forth place on this list of concerns for AI software because AI is data-hungry and much of this data will presumably come from IoT systems close to humans – of course, for those AI applications that are interacting in some form with people and the society at large (e.g., in decision support or smart-* systems). Privacy labels could help in these regulating endeavors as well, this time not addressed only to the convenience users but more to the
businesses, e.g., for allowing well informed AI software purchases, as well as to AI engineers, e.g., to guide their choices of libraries and software components.

Privacy has long been an important concern for software developers, e.g., the (then) President of the ACM, David Patterson in (Patterson, 2005) put forward the “SPUR manifesto” which was placing (P)rivacy as one of the four main focus areas for software engineering, along with (S)ecurity, (U)sability, and (R)eliability. On the contrary, biases in software (chiefly in AI-based decision systems, as Etzioni describes) is a rather new concept for software developers.

Biases in AI systems (the fifth area of concern for Etzioni) normally come from improper use of training data, i.e., the system is trained on a data set that is not representative of the population/problem that it is applied to (or does predictions about). AI biases can be about gender, race, or other social aspects (Caliskan et al., 2017; Zou and Schiebinger, 2018; Silva and Kenney, 2019), but also about privacy. This last form of bias, which we call privacy biases, is largely not investigated because it is not seen as a machine bias, i.e., it does not appear from data or the software code. Privacy biases are human biases, in line with the traditional bias mechanisms studied in psychology (Gilovich et al., 2002; Tversky and Kahneman, 1974; Oliver, 2014; Wilson and Gilbert, 2003) – see also a nice account of how cognitive and behavioral biasing mechanisms (such as the anchoring heuristic or framing effect) influence privacy behaviors in (Acquisti et al., 2017, Sec.2.3). Quite a number of privacy biases could fall in the class that we would call I-have-nothing-to-hide, with a large collection of such privacy attitudes nicely presented in (Solove, 2011). A privacy bias that programmers often fall pray to can be called privacy=security, which we have already explained in the begging of Section 2.1. Recent results (Pedersen et al., 2020) have shown that human biases can be transferred from the programmer into the software that she is building.

Bias transference is thus an additional mechanism to the standard one studied in AI biases, through which human biases can manifest into the software that we build. Therefore, privacy biases are elevated to being a serious threat to the software that programmers develop as it can incorporate the privacy biases of their creators. Privacy biases have as a root cause the lack of adequate knowledge, either that the person is time constrained and cannot gather or infer the needed knowledge for the decision task at hand, or that simply the person is inexperienced for the new task. Programmers often find themselves in such uncertainty situations, e.g., when faced with incomplete specifications or vague requirements. This is even more so in the case of understanding privacy policies. As a result, programmers are mostly left to their own means and judgment when implementing privacy features or requirements; and any privacy bias or neglect can reflect on the users of the resulting software. This happens because of the transfer of the programmers privacy views and biases into the software artifact (Pedersen et al., 2020) when privacy aspects are not easy to comprehend.

**OP.5: How can PL prevent the transfer of the programmers’ privacy views and biases into a source of privacy problems in software?**

Since PL would be associated to privacy agreements (cf. C.4) and having one goal to explaining concepts such as purpose of processing (see Section 4), they would help programmers to better understand those aspects from the privacy agreement that are relevant for the product they are building.

3.5. Regulators, Certification bodies, and Authorities

Reggy: Certifications bodies as stakeholders can see the PL as a means to convey their certification results. The provisions in Art. 42/43 of GDPR strengthen the role of the certification bodies as a means for the companies to show compliance (Lachaud, 2018). PL should contribute to the further development and enhancement of the existing certification schemes (cf. G.1).

Data Protection Authorities (DPAs) tend to rely on detailed sources, such as privacy policies and technical documentation, in their audit work. Therefore, highly relevant for DPAs would be the deeper layers of the PL, where detailed information is offered (cf. Section 6.1). However, DPAs are also responsible with checking if the visual and the “surface” components of the PL are an accurate reflection of the privacy policies and actual practices. In this case, their auditing work could be simplified through the automation tools and process used to generate the PL (cf. Section 5). Furthermore, their work becomes universally valid if the
same tools, practices and methods are used across all services. Uniform practices is one of the goals (G.4) we set in this paper for PL.

DPAs are also part of the Data Protection Board where they can interact with privacy regulators on various aspects of the legislation and its applications. Privacy is a concern in various social/economical areas, such as health, with a major role to play in the future of AI regulations (Etzioni, 2018; Clarke, 2019). PL could be introduced as an essential aspect of such regulations since PL allows easy comparisons regarding privacy between AI systems.

Data-intensive technologies such as AI-based decision systems or management software have entered also in the many state institutions such as in policing (Brayne, 2017) or courts (Grgić-Hlača et al., 2019; Dressel and Farid, 2018; Malgieri, 2019). Privacy is maybe of a greater concern to such institutions than it is to companies. In state institutions, decisions on purchasing a piece of technology or service is done through a highly regulated and transparent process called procurements. Privacy would thus be part of the requirements mentioned in the procurement call. PL could also here be used to make it easier to evaluated the proposals. This is the same way of applying PL in any form of technology purchase decision, be that done by a convenience user when looking to by a new IoT device, or a company management person looking to acquire a new service, or a governmental institution in a procurement process.

G.8: In public/private procurement, PL could be an advantage or sometimes even a requirement.

Data Protection Officers (DPOs) are, according to Art. 37-39 of GDPR, acting as intermediaries between the supervisory authorities, data subjects, and the organization by which they have been appointed. The role of DPOs is to facilitate compliance, and, besides certifications, are another instrument that can be adopted by companies to ensure accountability. In some cases, GDPR makes appointing a DPO mandatory, while for the rest of the organizations this is voluntary, in which case the organization may choose to use external DPOs. This is already the case in countries such as Germany, where there is a large community of external data protection officials hired by companies. However, there are differences in the level of use of the external DPOs between the countries. In the Netherlands, for example, the companies chose to handle the compliance mostly by themselves, using external experts only for one or two days per month. DPOs as stakeholders for PL would have commercial interests to foster self-evaluations (expanded upon in Section 7.1), where they could provide companies with input.

4. PL as a means of education and behavior change

The New Chicago School model (Lessig, 1998) explains how there are several modalities of regulation for the behaviors of people, and we would also argue that it applies to businesses as well. PL are meant to help regulate the behavior of convenience users when making choices that might influence their privacy, as well as regulating the behavior of businesses that handle private data. Therefore, the multidisciplinary character of PL involves multiple stakeholders, besides the law and regulatory institutions, in driving privacy behavior changes.

OP.6: One open problem that PL could be useful for is to help change the behaviors and attitudes of people in regard to privacy.

Eddy: The Prochaska model of stages of behavioral change (Prochaska and Velicer, 1997), often used to change behavior of addicted people, identifies several stages of awareness and appropriate actions. One may not be aware at all that she needs to make behavioral changes, meaning that the action targeting this person is to raise her awareness, e.g., in regard to privacy aspects. Then the person moves into the contemplation phase when realizing that there is an important concern, e.g., privacy, which she needs to think about. Having learned and understood the problem, the person has to determine whether, and what, to do. This is a decision point where PL can help, e.g., when the person needs to take action when buying a digital product.

Multy: Many people are not even aware of their lack of privacy. We see the PL as a tool for raising awareness. We already have examples from the food industry where labels are used to raise awareness about
the quality of the food. For example, people may be accustomed to think that all foods are healthy, but by seeing the labels they realize that some foods are healthier than others. They might try to find out more about the meaning of the label and might start discussing it. Awareness towards specific characteristics of digital products are similarly triggered by displaying different labels. Privacy labels, when attached to, e.g., mobile apps, and are visible in an app-store or comparative table, they could be the starting point for people to realize that one product is different from another when it comes to privacy protection. Only then people might go and look for further information about the meaning of the label and its contents. However, such markings can easily be used misleadingly for commercial purposes as well. If it is not a standardized label with clearly established frames, but one that the businesses choose to give the product by themselves, the package of the product might emphasize some aspects and omit others, e.g., displaying that the food contains 30% less fat, but not saying that it contains 30% more sugar. Therefore, if not regulated, PL could be used in a suggestive and deceptive manner to induce subjective and irrational (i.e., inappropriate) behavior that is not in the user’s best interest.

Reygy: A simple seal/label can be used to raise awareness. However, we would go beyond a mere seal. We envision a privacy label that displays information through which people can learn more about privacy related aspects, e.g., that location sharing is a privacy sensitive information, or information about how much data the provider is collecting from their subjects and what kind of data is being collected and processed, and for what purposes. The label can thus be the point of entry, providing the information that can be used to further educate people.

G.9: PL aims first at raising awareness and then further increasing knowledge and understanding of privacy in the population.

Eddy: According to the theory of planned behavior (Ajzen, 1991), behavioral change and judgments that one makes are conscious and are planned. This is a rational approach to behavioral change and decision making. However, it is known that people do not always make decisions rationally, e.g., when in time constraint or when one has insufficient or too complex information, one will not be able to carry out this type of rational approach to solve a new problem (Tversky and Kahneman, 1974; Kahneman et al., 1991; Acquisti et al., 2017). An alternative model is that of nudging (Thaler and Sunstein, 2009), which is an empirical approach to behavioral change exploiting the automatic, heuristic-based, intuitive thinking of (Kahneman, 2011; Gilovich et al., 2002).

Reygy: Even if nudging for privacy is debatable because, e.g., it may restrict the individual’s autonomy (Renaud and Zimmermann, 2018; Jarovskey, 2018) as it uses psychological mechanisms covertly, functioning unconsciously, nudging may be considered ethical as long as it is used in people’s best interest; presuming one knows what is actually in peoples’ best interest (Hausman and Welch, 2010). Lancey: Having an ethical approach to nudging, and not use it for commercial purposes (Sunstein, 2017; Thaler, 2018; Caraban and Karapanos, 2020; Narayanan et al., 2020), one can build a choice architecture that leads people to doing the right things and to carrying out the right activities that lead to the right decisions even if they are not aware of what they are doing. We already have many examples of ethical nudges used in the traffic for the protection of the drivers and pedestrians, such as speed limit signs and speed bumps. Eddy: However, there are many recent examples of consent forms that are GDPR compliant, but still nudge users to pick the privacy-intrusive choices, see e.g., cookie-banners (Machuletz and Böhme, 2020; Matte et al., 2020; Sanchez-Rola et al., 2019), or emphasized buttons that nudge the users to select all cookies, while the possibility to not ‘Select all and continue’ have very little visibility and are ambiguous about which purpose they serve. One other famous and old example not related to consent forms is the “opt-in/opt-out” check-boxes (Bellman et al., 2001), e.g., for receiving newsletters or offers from the respective company after performing an online transaction such as registering for a service or shopping online. Bussy: In many cases, and especially for privacy, how the company sets the ‘default’ checked/unchecked is done to serve the interest of the company. Such practices are known as “dark patterns”, and often applied disrespecting privacy (Bösch et al., 2016;
Mathur et al., 2019; Nouwens et al., 2020); even though the same nudging can be used also for good purposes, e.g., to increase the number of organ donors (Johnson and Goldstein, 2003). Lancey: It is difficult to control by law or regulation the use of dark patterns (Waldman, 2020; Narayanan et al., 2020).

**OP.7:** How can PL be a privacy nudge instrument to use for helping people make more privacy-conscious decisions when choosing a product?

One needs to distinguish between a rational approach to influencing people and an empirical nudging approach. By considering the combination of personality psychology, social psychology and cognitive psychology, from Figure 4, one can influence people and raise their awareness towards preferences that are good for them, or one could simply nudge them into doing that, with or without them being aware of what they are doing. Nudging does not always need to be covertly. (Caraban et al., 2019) shows that 78% of the nudges presented in the HCI literature make their intentions and means transparent to the user, prompting them to make an reflective choice.

Nudging should be used predominantly for cases where it is known that people run on autopilot, and they need help with making the right decisions. How much should a person be autonomous, and how much should she be nudged into a direction, is a question of ethical considerations. However, when running on autopilot, it should not be expected that people would make rational judgments – indeed, privacy decisions are often not done rationally – and thus PL nudging should act in their best interest.

One starting point can be found among the existing works on using nudging for privacy purposes (Wang et al., 2014; Zhang and Xu, 2016; Acquisti et al., 2017). We need then to know who we are dealing with by considering the users’ specific cognitive and behavioral characteristics. As such, empirical data needs to be collected on the prevalence of cognitive styles in different situations and about dominating personality styles from different cultures and different regions, as well as gender, age, education; all of which are known to influence users’ experiences and shape attitudes related to privacy concerns (Kitkowska, 2018; Jarovsky, 2018; Kitkowska et al., 2020b). In one study (Murmann et al., 2019) from the Privacy&Us project19, done on users of mobile health services it is shown that the notification preferences of these users correlate with their privacy personas. Another study shows why the privacy of certain groups should be considered differently to those of the wider community. A test instrument could be created, where people are asked to answer a few questions, that will help with placing them in one of these domains. Furthermore, such instruments could also help raise the awareness of the users about who they are and what cognitive style they have. If we are able to raise people’s awareness about privacy issues, we may even be able to steer people ‘away’ from the ‘maladaptive’ use of mental heuristics in situations where heuristic thinking is less appropriate, and instead steer them into a more rational way of thinking, maybe even approximating a more consciously planned behavior (Ajzen, 1991). Consider, e.g., how PL could appear different to people with high curiosity personalities compared to someone that travels much and might be interested only in location aspects, e.g., whether location is shared and with whom.

**OP.8:** We need to understand how to make the same PL slightly different to best match the needs of different types of personalities or activities.

5. Automation and tools for creating PL

Lancey: One way for having PL legally binding is to tie them to privacy policies (previously included in the Terms of Services, or ToS20). Techy: Machine learning and formal reasoning methods can be used to build tools to help translate (more or less) automatically between ToS and PL.

**OP.9:** How can Privacy Labels and privacy policies be correlated?

Lancey: The adoption and use of such tools in law firms depends on how inclined these are towards new technologies. Law firms foremost have the client’s best interest in mind, and any tools that they adopt should serve that purpose.

19Marie Skłodowska-Curie Innovative Training Network Privacy&Us: privacyus.eu
20See a community effort on explaining ToS at https://tosdr.org/
One starting point for research and tools relevant for translating between privacy policies and privacy labels can be found in the area of logical/legal reasoning and controlled natural languages. **Techy:** One of the more advanced tools is the Contract Verifier\(^{21}\) (Camilleri and Schneider, 2017; Camilleri et al., 2018), using several different technologies and off-the-shelf tools (see architecture in Fig. 5), and taking input a contract, or any kind of normative document written in English. Using a standard natural language parser for English (e.g., the Stanford parser\(^{22}\)) it generates a tabular view of the different clauses in this contract. This tabular view can be edited manually because the parser sometimes cannot parse the whole text or identify all the aspects, since the parsing from natural language into a formal model is an undecidable property (i.e., it is impossible, in general, to write a program that can do this translation fully automatically). After having a tabular interpretation, everything is automatic. A formal model can be extracted and used to do queries, e.g.: what are all the obligations in the contract, or privacy policy, for a given party; whether there are obligations without deadlines; what is the data processing purpose; or which party the data is being shared with.

The Controlled Natural Language (CNL), placed in the red box in Fig. 5, is the part where the privacy policy’s natural language text is simplified into a CNL version that still looks like natural language, but is more structured and with a limited vocabulary. CNL is still readable, while also amenable to formal manipulation, which is essential for doing automated reasoning. One problem encountered when trying to formalize legal documents is that it is very difficult to do anything unless you have a very precise definition. When one looks into the normative text – any privacy policy written in natural language or the GDPR for the same matter – many questions arise because words such as ‘adequate’ or ‘efficient’ are used, and do not have a precise meaning for technology people. When wanting to do something technically, one needs to have precise meanings for these kinds of words. One approach can be to take the usability related definition of such words, where the Usable Privacy Cube of (Johansen and Fischer-Hübner, 2020; Johansen and Fischer-Hübner, 2019) identifies measurable criteria for such concepts related to privacy.

**Multy:** Such works on automatically generating PL combined with the psychological model from Fig. 4 in

---

\(^{21}\)Proof-of-concept: [http://remu.grammaticalframework.org/contracts/verifier/](http://remu.grammaticalframework.org/contracts/verifier/)

\(^{22}\)https://nlp.stanford.edu/software/lex-parser.shtml
Section 4 would fit well within the general agenda of Behavioral Computer Science (Pedersen et al., 2018) since for PL both models of human behavior and of computers are needed in combination.

Techy: Constructing PL from ToS using AI is a two-step process: (i) use natural language processing (NLP) to identify privacy concepts in the ToS text and (ii) apply a set of rules to the identified concepts to generate the relevant labels.

One starting point is the award-winning tools from (Harkous et al., 2018). One of the few datasets that we consider to be useful for such purposes is the OPP-115 Corpus (Wilson et al., 2016), which contains 115 privacy policies that were manually annotated by law students and then checked for inconsistencies. Each privacy policy is segmented and each text segment is annotated with privacy concepts.

The privacy concepts are split into a number of categories. Each text segment fits into one or more categories, each category having a fixed set of attributes and each attribute a fixed set of possible values. After determining that a text segment fits into a certain category, we know which attributes apply and we can determine which value each attribute has.

For example, the category “Third Party Sharing/Collection” has the attribute “Identifiability” which may have the values “identifiable”, “aggregated”, “anonymized”, etc. This is a two-stage machine learning problem: (i) first identify which category the text segment fits in, and then (ii) for each attribute determine which of the possible values are applicable. This approach is used by (Harkous et al., 2018).

Using NLP to extract data from privacy policies poses a number of challenges. One of the challenges is getting a data set that is big enough, as labeling privacy policies is challenging in itself and requires considerable time and effort. Hence the scarcity of data sets that annotate privacy concepts. Another challenge comes from the ambiguity of natural languages, which allow for sentences that cannot be automatically uniquely interpreted. Some ambiguities are even seen by some to be advantageous in legal language. In large legal text, like privacy policies or regulations, challenges also come from the complex document structure and potential for very long relations between sections of the text, e.g., one paragraph may state some processing purpose for the data, but later in the text there may be a list of exceptions to this statement.

Lancey: Having PL correlated to ToS using NLP can be used, e.g., to compare a company’s PL to their ToS to see whether they match or find discrepancies between the two. One could also convert the extracted privacy concepts to simplified natural language, thus making summaries of ToS.

The use of a set of rules that correlate between the privacy concepts annotations from the text and the elements of the constructed PL is new to standard machine learning models and can be used to explain the reasoning behind the PL’s creation, i.e., if we want to use AI for our PL, we want explainable AI (Samek et al., 2019; Hoffman et al., 2018; Hagras, 2018). A potential use case would be a system that sorts the labels from most negative to most positive. A company can then go through the labels and decide which they would like to improve, click on the label to see the rules and the sections of the privacy policy that contributed to it, and with some domain knowledge decide what actions to take to improve the situation.

6. The Looks and Appearance of PL

Even if in all the fairytales the good knight has pleasing looks, for us the appearance of PL is all about conveying information to the target group. Upso: We have much to learn from the fields of Information Design and Visualization (Tufte, 2001; Mollerup, 2015; Ware, 2021; Few, 2009; Cairo, 2013; Knaflic, 2015), but we need to expand beyond the content being presented, to include psychology so to reach the individualized PL from Fig. 4 and to make the looks of PL useful for the educational purposes mentioned in Section 4.

Multy: Privacy icons are important for conveying information on a first level of detail (Holtz et al., 2011; Efroni et al., 2019). Icons would be needed for each of the different privacy concepts included in the PL (Motti and Caine, 2016). Privacy icons are useful for a “nutrition facts” style of PL, e.g., (Kelley et al., 2009; Emami-Naeini et al., 2020), this approach being taken by all the privacy labels discussed below. However, a more important part of PL would be a comparable view, in the style of energy consummation labels, involving graded scales, with the “nutrition facts” design and privacy icons appearing only beneath this.

OP.10: Evaluating the degree of data protection is complicated, but needs to be made measurable, at least for some of the privacy aspects, and fit into a graded-scale system.
It is also good to consider the recent online privacy labeling projects that have been started by different organizations or individuals.\textsuperscript{23} A few examples include:

- the Privacy Label that won the Gold Jury Prize in the European Design Awards
  \url{https://europeandesign.org/submissions/privacy-label/}
- the IoT Security and Privacy Label developed recently at CMU
  \url{https://iotsecurityprivacy.org/labels}
- the Privacy Nutrition Labels patterns from
  \url{https://privacypatterns.org/patterns/Privacy-Labels}

\subsection{6.1. The layered information of PL}

Reggy: We expand on the concept of “layered notices” as promoted by the (Article 29 Working Party, 2018)\textsuperscript{24}. A layer should offer the data subject only the information needed to make the right decision at a certain moment and for a specific purpose. However, the layers in their cumulative totality should meet the requirements for compliance. The top layer of PL needs to convey prominently core policy information, especially the data processing purposes, who is the data controller and other core information that has to be made transparent according to the GDPR (Art. 13). Moreover, information on how far security and PETs are used for implementing privacy by design should be of interest and communicated. In addition, the top layer could also contain information on the processing that has the most impact on the data subjects and that enables them to understand for each specific processing purpose what consequences it would have for them, along with any other information that could ‘surprise them’.

PL would go further and include also graded scales, e.g., inspired by energy consumption labels. Our envisaged scenario is the following. First the user is presented with a privacy grade on a scale from A to F. The user can then click on the grade and be shown further minimum of information (following Art. 29 Working Party) and icons to explain the most important parts of the agreement, maybe color-graded to explain how these contributed, positively or negatively, to the overall privacy grade. Further, the user can click on each icon to see a simplified natural language explanation of what the icon means. Finally, the user can click on the simplified natural language to be referred to the section(s) in the privacy policy the statement was constructed from. The idea is not that the user should make a decision based on a simple grade, but rather that the user can select which products they are most interested in based on the grade, compare the more detailed information and use it to decide which product they want.

\textbf{OP.11: How can a complex privacy policy, addressing different dimensions (including purposes, data controller, data types, retention periods, etc.) be mapped into a hierarchical A-F scale, while remaining relevant for a particular context and individual user?}

Even though it may be easier to grasp, a simple letter can be misleading if it is not given in a context (e.g., the type of app, the type of application domain, the role the user takes wrt. the application being evaluated). Moreover, the aggregation of the evaluations of the different aspects of a privacy policy is not easy since a policy may be more privacy-friendly in some aspects but not in others/all. Even more importantly, we want individualized privacy labels (as explained in Section 3.1 and pictured in Figure 4) because people have different privacy preferences (reflected in their privacy personas). This implies that a privacy label grade B for Alice may be perceived like an F for Bob, and hence the PL has to dynamically change based on the privacy person it is being coupled with (i.e., presented to). One set of measuring scales can come from criteria regarding the usability of privacy as defined in (Johansen and Fischer-Hübner, 2020).

\textsuperscript{23}For the past few years the idea of Privacy Labels has caught also in the news circles, see e.g., the following opinion articles: \url{https://ksr.hkspublications.org/2017/07/10/mandatory-digital-privacy-labels-one-way-to-protect-consumer-data/} or \url{https://www.politico.com/agenda/story/2018/04/25/internet-privacy-label-000656/}.

With such a basic overview of their privacy policy the company can gain more transparency. Such a PL overview is more comprehensible for the consumer and facilitates an easier comparison of the way service providers process personal data.

G.10: The goal with “layered PL” is to give the user a bird’s-eye view about what privacy aspects are included in the privacy policy. Then if the user wants more information, she can drill down to a deeper level. At the bottom layer, one can find the whole privacy policy.

Reggy: Compliance seals such as the ones from ULD, EuroPriSe, or TrustArk, usually convey only the information about the issuer of the privacy seal (and the validity period). In addition there is a document online giving full details, e.g., describing the target of the evaluation and what has been evaluated. However, such a two-levels approach (minimal seal and full detailed document) does not fulfill our desires.

Another proposal of privacy labels focuses on conveying in a concise and precise way on a single label, privacy-relevant information specifically chosen for some target group (Kelley et al., 2009; Railean and Reinhardt, 2018; Emami-Naeini et al., 2020). Some elements can be clicked, to find the full details of the decision behind them. However, these only cover parts of a complete compliance document, aiming to simplify it into a small and easily understandable label. This may work for simple products or services but it would not scale up to complicated systems or ToS. One other source of inspiration can be the layered approach used in cookies notices (Sanchez-Rola et al., 2019; Matte et al., 2020), which people might already be accustomed to.

Similarly, in the case of the NL.PL the layered approach chooses a number of basic elements used to produce a visual separation of the data flow – this describes what are the sources of the data, how and for what purpose are they used in the processing activities, how long data are retained and when they are deleted. Other privacy aspects such as retention terms or security measures are also included (see Figure 6 for an overview). NL.PL allows to drill down towards more details, and is built in a structured and standardized way, so that the businesses have the same topics to cover when providing privacy-relevant information during the NL.PL process.

The standardized model makes it easier to detect differences between labels of different service providers (e.g., G.8). Take, for example, the case of how one can drill down for location information in the NL.PL. There is an icon for location saying that most data is processed outside the EU. One can click on a question mark for more information which says that the laws of other countries may apply, with some extra information and a link to learn more about what does location mean and what legal requirements apply. The basic information is about the location of the storage of data, if it is within the EU, ‘Yes’ or ‘No’. If it is not, you have extra information on what does it mean not being stored in the EU, what kind of other different laws may apply and then you can click through it to learn even more. This is the way the NL.PL label is built around all the different main requirements the controllers have to inform data subjects about. It is a layered approach, where what you see on the label is some part of the information and there is the question mark after each sentence, where one can click and have some basic additional information and then you can click further to learn, reaching a knowledge base where one finds more information about what it actually means.

7. Three Approaches to Managing PL

Multy: Arriving to a PL with the characteristics and the goals described so far require a process involving all the stakeholders described in Section 3. Introducing and developing PL requires research efforts on the open problems that we have identified, but the adoption of PL requires involvement of more than the research communities. For the adoption to be successful, PL needs to show not only that it can solve problems of all these stakeholders, but also that it can live and thrive after the initial starting phase. This requires good management of PL, which could involve three different important aspects, all fitting together.

---

25Known from 1997 to 2017 at TRUSTe: https://trustarc.com/blog/2017/06/06/truste-transforms-to-trustarc/
C.8: The privacy label can be a way to visually present and structure the privacy aspects detailed in a privacy policy, thus helping to implement the very important GDPR principle of transparency. This is typical of a self-evaluation approach, as taken by NL.PL, and detailed more in Section 7.1, but still tied to a legally binding document so that it cannot become a means of deceit.

C.9: The privacy label can be a means of establishing trust, usually created through an evaluation and audit process by a certification body, such as EuroPriSe, based on technical requirements that strive to identify whether existing legislations, such as the GDPR in Europe, are respected (see Section 7.2).

C.10: The privacy label can be a way to measure (the usability of) privacy on scales, allowing for comparisons. Measuring can be done either automatically (see Section 5) or with the help of a community (see Section 7.3).

7.1. Self-evaluation

One starting point for doing self-evaluations of privacy and producing a privacy label is the quite advanced web-platform of PrivacyLabel.org (NL.PL). This combines a visual design that uses icons, with succinct textual descriptions, into an accessible way of presenting how an organization manages privacy. This is a service for doing self-evaluations, and as such, the organization by itself will have to explain and include information about the privacy measures taken, be that technical, PETs, legal, procedural, etc. The self-evaluation can be done by the organization itself or perhaps with support from outside privacy experts. This is different from a trust mark or seal that imply an independent evaluation done by designated bodies, who are evaluating and then creating the label for the organization. NL.PL focuses on transparency, allowing an organization to show that they are taking privacy seriously and how they are doing that.

Reggy: There is a crucial difference between internal auditing, as above, and external certifications that provide a seal issued by certification bodies, as referred in the Art. 42/43 of GDPR. This type of label is usually not covering the entire organization’s privacy attitude. An external auditor first checks a specific
system and its processing activities for compliance and then provides the label. The NL.PL approach is to have a label that is in the style of nutrition facts labels, showing to consumers and customers more explanations about what the organization is doing, making more of a summary of a privacy statement, i.e., aiming to make the privacy statement comprehensible for the customer or the user of the service. In a privacy mark there is more focus on the technical audit, while in the NL.PL there are a number of icons representing main categories such as whether the data is stored within the EU or outside, what data is used for, how long the retention terms are. The NL.PL contains basic information answering to basic requirements that are mandatory according to GDPR. It is more about the information duties than the technical official audit.

**Multy:** Checking the validity of NL.PL can be done by either an external audit (see Section 7.2), an automated process (see Section 5), or by a community effort (see Section 7.3). NL.PL only focuses on the privacy statement replacement and transparency, and not on trust marks or seals, which are certificates connected to a specific processing activity, or a specific system. **Reggy:** A trust mark or privacy seal that involves an external audit is not a replacement for a privacy statement. The organization needs to have this anyway, and then NL.PL offers a form of the privacy statement that is more transparent and comprehensible for the consumer.

**Lancey:** Self-assessment can imply the assumption of honest controllers. However, one can think of methods to oversee that self-assessed PL do not become a means of deceit, e.g., by involving administrator fines, since being transparent is still enforced through Art. 12-13 of GDPR. **Bussy:** Transparency is also mandatory for the Common Market\(^\text{26}\). The \(\mathring{C}\) marking\(^\text{27}\) is an example of an existing self-assessment that is mandatory for more critical areas such as medical devices, which is a small fraction of the market. In this area it is mandatory to have the \(\mathring{C}\) marking even before getting the product on the market. This is an example of a self-assessment that does not involve an external expert looking over the manufacturing places. A well known example of self-assessment comes from the e-waste management area (Kirckpatrick, 2020) known as EPEAT\(^\text{28}\) from the Green Electronics Council\(^\text{29}\), where the participation is voluntary, yet over the years it has become quite adopted by manufacturers. The program provides labeling for electronic products that meet certain criteria across a range of 12 categories, covering materials and chemical usage, energy efficiency, recyclability, product lifespan, and product design.

**G.11:** We do not expect PL to be something mandatory in a first phase, in the sense that external entities are checking for compliance, but instead aiming first to have such self-declarations more widely spread.

### 7.2. Certification and Audit

**Reggy:** DPAs can accredit companies to do privacy certification (Art. 43 of GDPR); in Europe one of the most advanced is EuroPriSe which originated in 2001 from the German Schleswig-Holstein DPA. Currently, audit and certification usually focuses on a system (or more often on one component that is considered critical for the system’s security and privacy), and therefore it is important in a PL to properly identify the target of the evaluation. As companies tend to put labels on the package, they might misinform, e.g., the ULD seal ended up on the boxes of the whole product, even though it concerned only the activation part. They had though a footnote mentioning that only the online validation tool received such a privacy seal.

**Lancey:** Privacy agreements are important documents for providing information on data processing, as required by regulations all over the world.\(^\text{30}\) It is common for lawyers to work with businesses to prepare and present privacy policies for certification or compliance audit; in which case devising a corresponding PL could be seen both as a certification seal as well as an explanatory label. Depending on the company and the target of evaluation, one could end up with very long agreements that are complicated and not easy to

---

\(^{26}\)The European Single Market, Internal Market or Common Market is a single market which seeks to guarantee the free movement of goods, capital, services, and labor – the ‘four freedoms’ – within the European Union.

\(^{27}\)The \(\mathring{C}\) marking is a certification mark that indicates conformity with health, safety, and environmental protection standards for products sold within the European Economic Area.

\(^{28}\)Electronic Product Environmental Assessment Tool: [https://epeat.net/](https://epeat.net/)

\(^{29}\)[https://greenelectronicscouncil.org/epeat-criteria/](https://greenelectronicscouncil.org/epeat-criteria/)

\(^{30}\)GDPR in Europe, California Consumer Privacy Act, Personal Information Protection and Electronic Documents Act in Canada, Privacy Act 1988 in Australia (Yuvaraj, 2018).

23
understand, in which case one is interested in presenting the information in a layered manner, simplifying and organizing the information so it becomes easier to be read by the consumers.

C.11: For PL to be part of a certification scheme it needs a standardized way to present the information, harmonized to be suited for the multitude of actors that have to present rather diverse privacy aspects. 

This is already the case with privacy policies, some using legal terminology, some simplifying it to terms understandable by a more general reader. Depending on how a company wants to be perceived by their customers, they can work with their privacy policies to, e.g., simplify the language or highlight some of the important aspects, but also add videos, illustrations, symbols, or a combination of these as the PL. Sometimes it can be that a company (for some of their services) wants to be sure that the legal aspects are thoroughly covered, thus devising a longer, more complicated privacy policy. Other business strategies are more concerned with giving a good impression, in terms of having easily readable policies, with simpler text that is more easily understandable by the end users.

7.3. Crowd-sourced and Community driven

There are numerous community efforts, from the widespread open-source software developments (OSM) or the Wikimedia projects, to the more recent and relevant LeDA\textsuperscript{31} or ToS;DR\textsuperscript{32}. In the same spirit and management style, we envision a Community Coordinated Privacy Labeling, or CoCoPL (see also the more general socio-technical framework called CoCoAI (Sivesind, 2021)). CoCoPL would include as part of the community both lawyers, e.g., from ToS;DR, as well as developers, e.g., from OSM projects, but also members from all the stakeholders identified in Section 3. One example is to involve the laypeople in a crowd-sourcing effort of annotating privacy policies to help the AI-based automation tools of Section 5. Another example can be to involve the DPAs as more trusted members of the community, though a trust-model is first needed, with such communities usually employing meritocracy. Internet communities have for a long time organized themselves in forums to evaluate businesses and products. Companies are well aware of this, and often misinformation becomes a problem that forums (or tech-magazines) have to deal with. CoCoPL would do a similar activity of evaluating the privacy practices of businesses and products, to the benefit of the community and everyone else as already mentioned.

Crowd-sourcing the data annotation means that we can have continued expansion of the data sets. Engaging interest groups that have experience with privacy policies, such as ToS;DR, would help fine-grain the privacy concepts used in the annotation models and how these would refine the constructed PL. Another benefit of crowd-sourcing data labeling is easy adaptation to changing standards and trends.

8. Concluding with a Timeline for Finding PL

Before concluding we devise an action plan for finding PL.

Phase I: Attract companies and organizations to perform self-assessments, based on their own privacy policies, using a structuring tool such as the NL.PL.

Phase II: Bring in more of the automation and reasoning tools for creating privacy measurements, to identify the level of privacy protection to be shown by PL. Include psychological models to individualize PL, and give also a proper appearance for the intended purposes, one of these being educational. Aim more on generating the PL automatically from reliable documents like the ToS and from inputs such as those from a user’s privacy profile.

Phase III: Join forces with the authorities for introducing a regulatory framework to make the self-assessment mandatory and uniform. At the same time, build a community around PL, that could even include the authorities as one (rather important) member in the community.

\textsuperscript{31}Legal Design Alliance https://www.legaldesignalliance.org

\textsuperscript{32}“Term of Service; Didn’t Read” community effort on explaining ToS https://tosdr.org
The self-evaluation is a logical first step from a GDPR perspective, because of the large emphasis on demonstrating compliance. The controllers need to show how they abide by the GDPR and how they implement the data protection principles. Further on, one can introduce checks and fines, including external audits and certifications, done by the authorities or the community.

Concluding remarks

When dealing with a complicated concept such as privacy, that is faced with multiple long-standing problems as discussed in Section 2, then to develop a solution can be a daunting task. In response, we propose an all-encompassing definition of Privacy Labeling as a possible start on the road towards a solution to many of the current privacy problems our society is struggling with.

When developing such a panoptic concept as the Privacy Labels proposed here, it is a good idea to find many discussion partners among the various stakeholder groups. Therefore, the ideas that we have presented have roots in our conversations with experts from the seven different fields that we considered relevant for privacy labeling. Our goal with bringing all these views was to investigate the concept of privacy labeling (and its implications) from many different angles. No one single discussion partner has ‘the right answer’, but their collective opinions sum up to a result that is much more comprehensive and powerful than any one view could accomplish on its own.

References

Sullivan, C., 2019. EU GDPR or APEC CBPR? A comparative analysis of the approach of the EU and APEC to cross
doi:10.1016/j.clcr.2019.05.004.
Tilkinen-Piri, C., Rohunen, A., Markkula, J., 2018. EU general data protection regulation: Changes and implications for
2556288.2557413.
Ware, C., 2021. Information visualization: perception for design. 4 ed., Morgan Kaufmann.
Symposium.
Wilson, S., Schaub, F., Dara, A.A., Liu, F., Cherdiral, S., Giovanni Leon, P., Schaarup Andersen, M., Zimmerk, S., Sathyen-
Privacy Policy Corpus, in: 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),
Advances in Experimental Social
Woodruff, A., Birn, V., Consolvo, S., Brandimarte, L., Acquisti, A., 2014. Would a privacy fundamentalist sell their dna
for $1000... if nothing bad happened as a result? the westin categories, behavioral intentions, and consequences, in: 10th
Zou, J., Schiebinger, L., 2018. AI can be sexist and racist – it’s time to make it fair. Nature 559, 324–326. doi:10.1038/d41586-
018-05707-8.
Books.
Paper II

Making GDPR Usable: A Model to Support Usability Evaluations of Privacy

Johanna Johansen, Simone Fischer-Hübner


DOI: https://dx.doi.org/10.1007/978-3-030-42504-3_18

A long version (starting at page 112) supporting this publication was published in May 2020 on the ArXiv: arXiv:1908.03503
Making GDPR Usable: A Model to Support Usability Evaluations of Privacy

Johanna Johansen\textsuperscript{1} and Simone Fischer-Hübner\textsuperscript{2}

\textsuperscript{1} Department of Informatics, University of Oslo, Oslo, Norway
\texttt{johanna@johansenresearch.info}
\textsuperscript{2} Department of Mathematics and Computer Science, Karlstad University, Karlstad, Sweden
\texttt{simone.fischer-huebner@kau.se}

Abstract. We introduce a new model for evaluating privacy that builds on the criteria proposed by the EuroPriSe certification scheme by adding usability criteria. Our model is visually represented through a cube, called Usable Privacy Cube (or UP Cube), where each of its three axes of variability captures, respectively: rights of the data subjects, privacy principles, and usable privacy criteria. We slightly reorganize the criteria of EuroPriSe to fit with the UP Cube model, i.e., we show how EuroPriSe can be viewed as a combination of only rights and principles, forming the two axes at the basis of our UP Cube. In this way we also want to bring out two perspectives on privacy: that of the data subjects and, respectively, that of the controllers/processors. We define usable privacy criteria based on usability goals that we have extracted from the whole text of the General Data Protection Regulation. The criteria are designed to produce measurements of the level of usability with which the goals are reached. Precisely, we measure effectiveness, efficiency, and satisfaction, considering both the objective and the perceived usability outcomes, producing measures of accuracy and completeness, of resource utilization (e.g., time, effort, financial), and measures resulting from satisfaction scales. In the long run, the UP Cube is meant to be the model behind a new certification methodology capable of evaluating the usability of privacy, to the benefit of common users. For industries, considering also the usability of privacy would allow for greater business differentiation, beyond GDPR compliance.

Keywords: Usable privacy · Human-Computer Interaction · Usability goals · Usable privacy criteria · Privacy certification · GDPR

A long version of this paper is available as [17].

We would like to thank the anonymous reviewers for helping improve the paper. The first author was partially supported by the project IoTSec – Security in IoT for Smart Grids, with nr. 248113. Thanks go also to Josef Noll for introducing me to the topic of privacy evaluations and labeling.
1 Introduction

The complexity of the privacy concept as such and of digital data and technology, make it difficult for one to evaluate the privacy properties of a specific piece of technology (e.g., web service, Internet of Things (IoT) product, or communication device). The difficulty is not only for average people, but also for regulators to check compliance, and for developers to be able to provide privacy-aware digital services/products/systems.\(^1\) Indeed, there are multiple concepts involved in digital privacy, like data sharing (which for normal business practices nowadays can form a highly intricate network of relationships), ownership and control of data, accountability or transparency (both towards the regulators as well as the users). Many of the privacy concepts are even a challenge by themselves, when it comes to their evaluation, since they are difficult to measure or to present/explain.

For explaining the intricacies of privacy, besides research articles and books \([13]\), there are several legislative texts adopted in different jurisdictions. The General Data Protection Regulation (GDPR)\(^2\) in Europe makes a good effort in clarifying many aspects of data privacy, providing the legislative support to enforce better data protection practices on anyone (within its jurisdiction) collecting and processing personal data. However, these regulations only specify the requirements on the data controllers in the form of basic principles, and the rights of the data subjects, but do not make any strict claims about the extent to which a controller (or processor) should go about implementing these requirements so that they are beneficial for the user, and to what degree.

As such, one motivation for usability evaluations of privacy is the fact that usability goals of GDPR, s.a. "... any information ... and communication ... relating to processing [to be provided] to the data subject in a concise, transparent, intelligible and easily accessible form, using clear and plain language, ..." (Article 12 (1) of GDPR), are left open to the subjective interpretation of both evaluators and controllers. The provisions of GDPR regarding usability are too general and high-level to be suitable for a certification process \([18]\). To remedy this, we propose a set of criteria thought to produce measurable evaluations of the usability with which privacy goals of data protection are reached.

For evaluating privacy we take as starting point the methodology developed by EuroPriSe \([3]\) that has as purpose to evaluate compliance with GDPR. We are guided by the EuroPriSe criteria when eliciting, what we call, principles and rights, which form the two variability axes at the basis of our model, i.e., which principles are followed and which rights are respected. However, EuroPriSe does not consider usability, which is the main focus of our work here. As such, one contribution of this paper is to show how to add usability aspects to the existing evaluation criteria of EuroPriSe.

Unlike EuroPriSe (and other existing certification schemes) that provides a seal showing compliance with data protection regulations (or industry stan-

\(^1\) Note that system/product/service are used interchangeably throughout the paper.

\(^2\) GDPR – General Data Protection Regulation from European Union \([1]\).
standards), our evaluation measures on a scale how well data protection obligations are respected and how easy it is for a user to understand that. The measurements can be presented to the user in different ways, e.g., using “traffic light” scales, showing which level of usability has been reached by the privacy of a certain technological product. A “traffic light” presentation of privacy is recommended by [2, Chapter 6(235)] as a way to “foster competition” and “show good practice on privacy policies”.

Traditionally, usability is a quality related to the use of a product. In our case, we are not interested in the usability of a product per se, but only in those aspects of a product that concern privacy. Our conceptualization of usable privacy is based on the definition of usability as presented in the ISO 9241-11:2018 [4], which we adapt to include privacy as follows:

**Usable privacy** refers to the extent to which a product or a service protects the privacy of the users in an efficient, effective and satisfactory way by taking into consideration the particular characteristics of the users, goals, tasks, resources, and the technical, physical, social, cultural, and organizational environments in which the product/service is used.

Our long term goal is to create a methodology to support service providers to make the privacy of their products more usable. The Usable Privacy Cube (UP Cube) described in Sect. 3 and the usable privacy criteria introduced in Sect. 6 are the first building blocks of the methodology we are aiming for. They are meant as tools, for both usability engineering experts and certification bodies, to evaluate if a product was designed to respect and protect the privacy of its users in an usable way. Once privacy measures and privacy enhancing technologies are integrated into the design of a product, it still remains to find out if (and how much or to what extent) those measures empower and respect the rights of their particular user as intended. In Human-Computer Interaction (HCI) this is determined based on user testing and usability evaluations. The criteria we propose presume the use of such established HCI methods for usability evaluations (e.g., [12]).

The legislation does not directly refer to usability goals and context of use as known in the ergonomics/human factors or human-centered design. However, requirements as the one in the Recital (39) of GDPR asking for the information addressed to the data subject to be “easily accessible and easy to understand” are categorized in this paper as usability goals, for which we create usable privacy criteria meant to measure effectiveness, efficiency and satisfaction – as usability outcomes – with regard to privacy aspects (we henceforth call these Usable Privacy criteria, and abbreviated it as UP criteria).

After a short digression into Related Work in Sect. 2, we introduce in Sect. 3 the UP Cube model, which is the main contribution of this work. We then continue to detail the UP Cube in the rest of the paper. Section 4 presents the EuroPriSe in the new light of the UP Cube, forming the two axes of criteria at its basis. The third vertical axis of the UP Cube, a genuine contribution of this paper, is formed of the UP criteria detailed in Sect. 6. To the best of our
knowledge, there is not other work that extends privacy certification schemes with usability criteria. Section 5 presents usable privacy goals that the criteria are meant to measure. The UP Cube naturally captures Interactions between all the axes, which we talk about in Sect. 7. We conclude in Sect. 8, presenting also some avenues for further work.

2 Putting the Work into Context

Usable Privacy and Security. The present work can be placed in the research field called usable privacy and security, with seminal works s.a. [6,10,14,27] and conference series s.a. the Symposium On Usable Privacy and Security (SOUPS). We consider that research on privacy requires, even more than security, an interdisciplinary approach (encompassing the expertise coming from research fields such as Psychology, Law or Human-Computer Interaction). As [5] points out, privacy has its meaning rooted in larger cultural and social practices and has political, ethical as well as personal connotations.

Regarding the relation between security and privacy, in this paper we consider security as one integral aspect of privacy, where privacy implies security but not the other way around. We consider such a clarification necessary, as we have seen a tendency in the general public to equalize the meanings of the two terms in favor of security. In computer science, privacy research has been closely intertwined with security research, reflected e.g. in the contents and the structure of the book [11]. However, in this paper, we favor the term “usable privacy”, as it includes by default security, which is in accordance with the data protection legislation, where security (integrity and confidentiality) is specified as one of the several principles to abide by in order to assure the privacy of users’ data.

Human-Computer Interaction. Having the goal to evaluate the usability of privacy in technological systems and products, makes our work part of the larger HCI research on privacy [5,19,20,23]. Following the classifications made by Iachello and Hong in their review [16], we approach privacy from a “data protection” perspective by extracting usability related goals from the GDPR. A similar approach is taken in [23], which translates legislative clauses of the Directive 95/46/EC (now replaced by GDPR) into interaction implications and interface specifications.

For evaluating how well a product meets privacy requirements, context of use variables s.a. user capabilities, tasks, the field where the technology is going to be deployed (e.g., healthcare, industrial facilities), should be defined. We thus adopt the ergonomic approach from ISO 9241-11:2018 where usability is always considered in a specified context of use, since the usability to be applied to a certain technology can be significantly different for varied combinations of users, goals, tasks and their respective contexts.
3 The Usable Privacy Cube Model

We devise a model for organizing the criteria to use in privacy evaluations and measurements, and represent it as a cube with three axes of variability (see Fig. 1), which we call the Usable Privacy Cube (UP Cube). The two axes found at the base of the UP Cube are composed of the existing EuroPriSe criteria, which we slightly reorganize in the Sect. 4 to fit in one of the two categories: data protection principles or rights of the data subjects.

We want to emphasize two perspectives on privacy that the UP Cube represents (hence our restructuring of the EuroPriSe criteria): the perspective of the controllers and of the data subjects. The controllers are thus given an overview of the principles that they are obliged to follow, whereas the data subjects are offered an overview of their rights.

The UP Cube allows to visualize interactions between the axes, made easier by our separation of the criteria into the three categories. Each such intersection has its specifics and could be studied in itself; we identify a few exemplary points of intersection between the axes in Sect. 7.

Example 1. The intersection between the transparency principle and the right to be informed is identified in Article 12 of GDPR. The controllers are obliged to provide the data subject information that should be concise, transparent, intelligible and in easily accessible form, using clear and plain language.

The third vertical axis of the cube is composed of our UP criteria, presented in Sect. 6. The UP criteria are determined based on usable privacy goals and are evaluated considering the context of use by following the guidelines in the ISO 9241-11:2018 standard. Interactions exist also with this third axis.
Example 2. For the case presented in Example 1, in order to establish how easily accessible or clear the information is, we must measure the level of efficiency, effectiveness and satisfaction in a specific context of use. Efficiency implies measuring the time and effort spent by a specific user for finding the information needed and for understanding it. Effectiveness measures the completeness with which a goal was achieved. In this case we would like to know how much of the needed information was the specific user able to access and understand. At the same time, what a certain type of user perceives as intelligible information, might be perceived by another as difficult to comprehend. Establishing the perceived characteristics of information is an activity categorized under the satisfaction usability outcome.

The UP Cube also brings the idea of orderings on each axis, hence the arrows. Such orderings are important for several reasons, e.g., UP criteria can be ordered based on “how little effort is required to evaluate it compared to how much overall evaluation outcome it entails” or “covers most technologies”. Usual for certification methods is to use a decision tree order to capture the impact of each criterion (e.g., choosing the most discriminating first), thus which to prioritize in the evaluation.

Judging from practice, one is inclined to think that an ordering is not always possible to find as some principles are equally important, therefore the orders are not necessarily strict. Moreover, one can even see one principle as more important than another only in some industry or context, whereas in a different industry the same two principles would be ordered the other way, therefore one may think that the orders are only partial (i.e., not total). However, in a specific cube (i.e., used in a specific methodology by a specific authority for privacy usability evaluations in a specific industry and context) there must always be an ordering in which the criteria should be applied. One can always generate a strict and total order from a partial order by just taking a random decision on ordering two criteria when no reasonable order exists. For example, one can any time pick as default order the one arising from the textual placement of the criteria in the data protection legislation texts (maybe considering content from articles as more general than content from recitals), or in the EuroPriSe (or the regulator/company) catalogs. What is certain is that each use case or industry has its specific requirements from which a meaningful ordering would be created.

Forming a specific UP Cube, i.e., deciding on the precise details of each criteria on the three axes and the orderings, is to some degree dependent on the specific context of use for the respective product to be evaluated. Therefore, one can think of infinitely many cubes, one for each different context. The criteria will not be different between the cubes, but their scope, depth, and evaluation might be different, depending on the context.

4 EuroPriSe

EuroPriSe originated from the Schleswig-Holstein Data Protection Seal, which was led by the Schleswig-Holstein Data Protection Authority (DPA) from
Table 1. Overview of the the EuroPriSe criteria categorized to fit into our UP Cube model, i.e., as the two axes with Principles and Rights, as well as Context of use.

<table>
<thead>
<tr>
<th>EuroPriSe Criteria: We list the names of (sub)sections as appearing in the EuroPriSe document [3], which has two parts, the second being subdivided into four sets of criteria, whereas the first contains preliminary issues, from where only section C is relevant for us</th>
<th>Principles</th>
<th>Rights</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>C. Target of Evaluation (ToE)</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>1.1.1 Processing Operations; Purpose(s)</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>1.1.2 Processed Personal Data</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>1.1.3 Controller</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>1.1.4 Transnational Operations</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>1.2.1 Data Protection by Design and by Default</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.2.2 Transparency</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.1 Legal Basis for the Processing of Personal Data</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.2 General Requirements</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.3.1 Data Collection (Information Duties)</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>2.3.2 Internal Data Disclosure</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>2.3.3 Disclosure of Data to Third Parties</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>2.3.4 Erasure of Data after Cessation of Requirement</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.4.1 Processing of Data by Joint Controllers</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.4.2 Processing of Data by a Processor</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.4.3 Transfer to the Third Countries</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.4.4 Automated Individual Decisions</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.4.5 Processing of Personal Data Relating to Children</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>2.5 Compliance with General Data Protection Principles</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Set 3: Technical-Organisational Measures</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Set 4: Data Subjects’ Rights</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

c.a. 2001 until the end of 2013, when it was transferred to a company, EuroPriSe GmbH. The scheme has a history of eighteen years [15] and is one of the oldest privacy and data protection seals based on a law, i.e., the State Data Protection Act of the German federal State Schleswig-Holstein. The role of the seal is to help the vendors of IT products and services to comply with the data protec-
We have chosen EuroPriSe as the basis for our UP Cube because of its long history, its continuous improvement, strong list of well-developed criteria, being led in the past by a DPA, and being based on the European data protection legislation. EuroPriSe also integrates with widely acknowledged IT security certification methods s.a. ISO 27000 and the The Standard Data Protection Model.

The way the criteria are formulated, as questions, also fits with the form of our usable privacy evaluation criteria. In addition, the existing EuroPriSe evaluation, which is at the basis of our model, assures that the GDPR legal grounds are covered, including data protection principles and duties and data subject rights. The UP criteria evaluations come on top, fine-graining the EuroPriSe evaluation with usability measurements, showing how well the legislation is respected.

Another feature that is relevant for our user-centered approach is that the EuroPriSe criteria catalog has been updated to include the data protection by default paradigm, promoting built-in data protection and privacy-friendly default settings. Moreover, EuroPriSe takes into account the technical, organizational and legal framework within which the product or service is operated and asks for considering the requirements of all the parties involved in the system, aiming at strengthening the position of the data subjects. Our work shares with EuroPriSe its high-level goal of making transparent for the general public how companies are managing data protection in their products and services.

In order to build on EuroPriSe, we first look into how its methodology fits with our UP Cube model. We show how EuroPriSe criteria can be redistributed into one of the two axes at the basis, i.e., as either rights of the data subjects or as privacy principles, or otherwise as a context of use criterion. Table 1 gives an overview of this redistribution. The distinction between principles and rights is inspired by the structure in [13], where principles and rights represent the core of this handbook. One purpose of the principles, mentioned in [13], is to serve as the starting point when interpreting the more detailed provisions in the subsequent articles of data protection law. The law also requires that these principles should correspond to the rights presented in the articles 12 to 22. This correspondence can be visualized through the intersection between the respective rights and principles axes of the UP Cube.

5 Usable Privacy Goals

We identify usable privacy goals (henceforth called Usable Privacy goals, and abbreviated as UP goals) that appear in the GDPR text. These guide the work...
in Sect. 6 where we present the UP criteria meant to measure to what extent these goals are being achieved. We give here only some examples of goals, numbered as in the long version [17], where the full list of 30 UP goals can be found. The goals are listed in the order they appear in the legislation. The words emphasized in each goal relate to usability. The chosen words are those that can be interpreted differently based on the context they are used in, and can result in objective and perceived measurements when evaluated in usability tests. These words also capture goals that can be achieved up to certain degrees, and thus can be translated into a level in an evaluation scale. In addition to the GDPR, there are more specific data protection laws, such as the proposed ePrivacy Regulation that have implication for usability, from where one could eventually extract additional usability goals.

**UPG.3** Consent should be given by a clear affirmative act establishing a freely given, specific, informed and unambiguous indication of the data subject’s agreement to the processing of personal data relating to him or her. [Recital (32) of GDPR]

**UPG.8** Make the natural persons aware of how to exercise their rights in relation to processing of personal data. [Recital (39) of GDPR]

**UPG.18** Any information addressed to the public or to the data subject to be concise, easily accessible and easy to understand. [Article 12 (1) and Recital (58) of GDPR]

**UPG.19** Any information addressed to the public or to the data subject to use clear and plain language. [Article 12 (1) and Recital (58) of GDPR]

**UPG.21** Provide information of the intended processing in an easily visible, intelligible and clearly legible manner. [Article 12 (7) and Recital (60) of GDPR]

**UPG.24** Allow the data subjects to quickly assess the level of data protection of relevant products and services. [Recital (100) linking to Article 42 of GDPR]

### 6 Usable Privacy Criteria

The proposed criteria are always measurable, which makes the results of a privacy evaluation easier to present visually through the use of a privacy labeling scheme. The use of privacy labels will then fulfill the goal UPG.24. This goal has a special significance from a usability point of view as it reduces considerably the effort spent by the data subject for evaluating privacy, which for most users is not the primary task [5] and it gets in the way of buying or using a product or service.

For generic goals like [17, UPG.1] that regards protection of personal data in general, we formulate a criterion that considers usability as follows:
What is the level of the *usability* of the personal data protection/privacy that the product or service ensures?

For being able to establish a level of how usable the privacy protection is, the evaluation needs to produce *measurable* outcomes. The structure that we follow is the one proposed in the ISO 9241-11:2018 where the measures consider both the objective and the perceived outcomes of usability (the UP criteria are labeled accordingly). The measurements will produce *counts* or *frequencies* (e.g., how many errors the user does when probed to do certain privacy related tasks) and *continuous data* (e.g., how much time does the user spend on completing a task related to privacy). The evaluation based on the UP criteria proposed below will produce three *main categories of measures*:

1. measures of accuracy and completeness,
2. resource utilization (time, effort, financial, and material resources), and
3. measures resulting from satisfaction scales.

The score for a main UP criterion is established based on evaluations of more specific UP criteria, called subcriteria. In order to reach a *high level* of “control of their own personal data” (Recital (7) of GDPR) the scores from evaluations of the subcriteria should also be high. The resources used to achieve a criterion, i.e., *time, effort, financial, and material* (which we abbreviate TEFM), should be measured to be able to determine the efficiency with which a specific criterion was reached. In addition, the results from the evaluations should show the level of perception that the data subjects have about their data being protected. The data subjects need to be highly satisfied with the offered privacy protection. The “high satisfaction” level is defined based on the user satisfaction evaluation of the respective subcriteria, and is also later important for the adoption of privacy technologies.

The UP criteria are categorized based on their area of application from the GDPR text. Figure 2 gives an overview of the number of criteria in each category.

A high-level UP criterion, like UPC.2, is labeled with the goal that it is related to, UPG.18. We then classify each UP subcriterion (e.g., from UPC.2.1 to
A Model to Support Usability Evaluations of Privacy

We try to be exhaustive in our UP subcriteria and to give enough questions to cover all major aspects that need to be measured to achieve the respective goal that the high-level UP criterion relates to. The UP subcriteria are labeled with sublabels representing various specific measures of usability for the above three general categories, e.g.: [Effectiveness:Completeness].

6.1 List of UP Criteria

We give few examples of the usable privacy criteria, while the full list of all 24 UP criteria can be found in the long version [17]. Since our criteria are modular (i.e., each high-level criterion is thought independent of the other) and can be ordered based on their importance for different application cases, they could be introduced gradually and selectively. It can be that certification bodies (like EuroPriSe) would start to include our UP criteria in their future test catalogs on an article-basis, e.g., a good candidate is Article 12 of GDPR (referring to rights that intersect with the transparency principle) as it contains five UP goals.

UPC.2 Is any information and communication addressed to the public or to the data subjects related to the processing of personal data concise, easily accessible and easy to understand? [UPG.18] [Type of criteria: Information and communication addressed to the public or to the data subjects]

How much [Time/Effort/Financial/Material resources] do the data subjects need to invest in order to [UPC.2.1 access, UPC.2.2 read through, UPC.2.3 understand] the information? [Efficiency:Time used, Human effort expanded, financial resources expanded, materials expanded] [Measure: Objective]

How much of the information were the data subjects able to [UPC.2.4 access, UPC.2.5 understand, UPC.2.6 read through]? [Measure: Objective] [Effectiveness: Completeness]

UPC.2.7 To what degree the data subjects perceive the information as concise? [Satisfaction: Cognitive responses] [Measure: Perceived]

To what degree the data subjects perceive the information as easy to [UPC.2.8 access, UPC.2.9 understand]? [Satisfaction: Cognitive responses] [Measure: Perceived]

Remark 1. The subcriteria in UPC.2 refer to cognition and understanding, while the subcriteria in UPC.3 refer to visual aspects of the information presented.

Remark 2. In different HCI works one can find different formulations that could seem related to how we formulate the subcriteria, e.g.: “Can the data subjects make sense of the information at all?”; “What is the extent to which the data subjects make sense of the information?”. However, we intend to measure the proportion of the information that is made sense of. Therefore we use formulations that give a statistically measurable outcome, such as “How much?”, “What is the percentage?”, “What is the degree?”. 
**UPC.3** Is the information about the intended processing provided in an easily visible, intelligible and clearly legible manner? [UPG.21] *Type: Info*

How much TEFM do the data subjects need to invest in order to [UPC.3.1 see/locate, and UPC.3.2 distinguish] the information? [Ey: Time used, Human effort expanded, Financial resources expanded, Materials expanded]

How well were the data subjects able to [UPC.3.3 visually locate and UPC.3.4 distinguish] the information? [Es: Accuracy]

How much of the information were the data subjects able to [UPC.3.5 visually locate and UPC.3.6 distinguish]? [Es: Completeness]

To what degree the data subjects perceive the information as [UPC.3.7 easily visible, UPC.3.8 intelligible, and UPC.3.9 clearly legible]? [S: Cognitive responses]

**Remark 3.** Poor visibility can affect the perception of trust, as information that has low visibility can appear to be hidden with a purpose. Poor legibility can reflect sloppiness in the way the content is produced, which again can give an impression of lack of professionalism. Poor visibility and legibility affects the satisfaction of the data subjects and it can cause physical discomfort (e.g., to the eyes, by having to read a text written in a very small font).

**UPC.4** Is any information and communication addressed to the public or to the data subjects related to the processing of personal data using clear and plain language? [UPG.19] *Type: Info*

What is the level of [UPC.4.1 clearness and UPC.4.2 plainness] of the language? [Es: Accuracy]

**UPC.4.3** What is the percentage of the data subjects that understand the language? [Es: Completeness]

What is the percentage of the language considered [UPC.4.4 plain and UPC.4.5 clear]? [Es: Completeness]

How [UPC.4.6 clear and UPC.4.7 plain] do the data subjects perceive the language to be? [S: Cognitive responses]

Several usability goals are found in the consent related provisions. These provisions are evaluated in detail in the EuroPriSe sections 2.1.1.1 Processing on the Basis of Consent and 2.1.1.2 Processing on the Basis of a Contract. The criteria we generate here are meant to complement the ones in the EuroPriSe through bringing in usability concerns. Marc Langheinrich presents several of the problems with how consent can be misused [21]. One of these is the “take it or leave it” dualism where the person does not have a real choice and thus getting consent comes very close to blackmailing. This problem has been ameliorated in the GDPR law by asking the controllers to allow for separate consent for different data processing operations. A usability evaluation could help further by revealing how the data subjects perceive the consenting act, as well as whether
the data subjects consider consent a real choice and if the options to consent to some of the processing operations only, are satisfactory.

**UPC.8** Is consent given by a clear affirmative act establishing a freely given, specific, informed and unambiguous indication of the data subjects’ agreement to the processing of personal data relating to them? [UPG.3]/[Type: Consent]

**UPC.8.1** How much of the consent text do the data subjects understand? [Es:Completeness]

**UPC.8.2** How much of the implications of consenting do the data subjects understand? [Es:Completeness]

To what degree the data subjects perceive the agreement to be freely given, informed, and unambiguous? [S:Cognitive responses]

### 7 Interactions Between the Three Axes

Characteristic to the data legislation text is that it always refers to how principles and rights intersect and depend on each other. In this section, we give examples of such references found in the recitals of GDPR, relevant for some of the identified usability goals. The recitals, though not legally biding, are meant to provide more details to the GDPR’s articles. The lawfulness, fairness, and transparency of processing principles, and the right to be informed appear to be closely interrelated, having also the highest occurrence of usability goals.

1. The UP criterion [17, UPC.1] refers to the control the data subjects have over their data. The criterion can be related to the right to data portability, through the Recital (68), where due to the aim of strengthening the control of the data subject, the “data subject should also be allowed to receive personal data concerning him or her, which he or she has provided to a controller in a structured, commonly used, machine-readable and interoperable format, and to transmit it to another controller ...”. It can also be linked to data security principle through the provision in the Recital (75) where the “risk to the rights and freedoms of natural persons” can result in data subjects being deprived of their rights and freedoms or prevented from exercising control over their personal data. The “risk to the rights and freedoms of natural persons” is also mentioned by the [13, pp. 131, 134] in the context of data security principle.

2. The UP criteria UPC.2 and UPC.4 are related to the transparency of processing principle, which is referred to directly in the Recital (58), where the respective goals are extracted from – “The principle of transparency requires that any information and communication related to the processing of those personal data ...” – and principles of lawfulness and fairness, which are also directly referred to in the Recital (39) – “Any processing of the personal data should be lawful and fair”.

107
3. The UPG.8 goal relates to the fairness and transparency of processing principles, and is placed under these respective categories, also by the [13, pp. 117, 120].

4. The goal UPG.19 is mentioned in the context of the transparency principle, in the Recital (39), where the information to be given to the data subject relates to the purpose of processing. This connects the principle of transparency with the principle of data minimization.

5. The UP criterion [17, UPC.22] about “the personal data [being] adequate, relevant and limited to what is necessary for the purposes for which they are processed”, is based on the [17, UPG.10], extracted from the Recital (39) of GDPR. This criterion is mentioned in Recital (39) as one of the requirements for complying with the transparency principle, while also referring to the purpose of processing. This connects the present criterion also with the principle of data minimization and in addition with the data protection by design principle. The link between the last two principles can also be seen in EuroPriSe criteria catalog, where data minimization is the focus of the [3, 1.2.1 Data Protection by Design and by Default, p.18] section.

8 Conclusion and Further Work

The benefits of the UP Cube model are multiple: (i) emphasizing both the perspectives of data subjects and of controllers; (ii) representing visually on the three variability axes the existing rights and principles criteria from EuroPriSe, together with our new UP criteria; (iii) visualizing intersections between the three axes; (iv) allowing ordering of the criteria on each axis.

The theory behind our usability evaluation of privacy is based on the well established standards ISO/IEC 29100:2011 and ISO 9241-11:2018. We worked directly with the GDPR text, guided by [13], which also inspired our structuring of the EuroPriSe criteria into rights and principles. Our HCI and usability perspective on privacy is influenced by the seminal works [5,6,10,14,19,20,23].

To build the UP Cube we have:

- identified from the GDPR text 30 UP goals,
- created 24 UP criteria, each with measurable subcriteria, and
- restructured the criteria of EuroPriSe, laid as the basis of the UP Cube.

**Further Work.** The UP Cube is meant as the groundwork for building a certification methodology, extending EuroPriSe to evaluate the usability of privacy. The proposed UP criteria are designed to produce measurable evaluations, useful for generating privacy labels in order to guide stakeholders when choosing technological products, by representing and visualizing the different levels of privacy. To achieve this larger goal, one needs to investigate which existing HCI methods for usability testing should be used for each of the UP criteria, and in what way.

One example of such a usability method for measuring the perceived usability of a system is the System Usability Scale (SUS) [8], a ten-item attitude Likert scale questionnaire. The standard [4, Annex B: Usability measurements]
also gives examples of methods that produce measurements relevant for our UP criteria, s.a. observing the user behavior to identify the actual usability problems, or asking the users to carry out tasks in a real or simulated context of use and measuring the outcomes. The experts can also run heuristic evaluations following design principles, theories and standards from the design and cognitive fields. More concrete examples of HCI methods and how these could be used for privacy and security solutions can be found in [19].

Which methods are appropriate to use, the number of test persons, and other test related concerns, depend on contextual factors, s.a. the type of technology, users and industry. Defining the required context is what our model offers support for. However, more work (e.g., providing guidelines and examples) is needed on how the context of use can be established.

HCI practices conduct user studies throughout the whole lifecycle of a product. These studies are run by the company itself, with the help of HCI (User Experience or Interaction Design) experts. For certification, the accredited data protection assessors would be using the results provided by the company to answer the UP criteria questions. In the cases of not enough or not reliable results, the assessors can recommend/require further testing. It would be valuable to have guidelines, e.g., in the form of a check-list, to help the assessors with establishing if the results from the company are reliable and sufficient. Recommendations for the businesses are useful as well, to guide how to conduct privacy related user testing, so that the results would be reliable later for certification.

With the same goal of achieving a complete methodology that can be taken in use by the accreditation bodies, building on the present model, one could create a visual representation of the evaluation, i.e., a translation of the measurements of usability of privacy provided by the UP criteria into a visually appealing privacy label. This should serve as a vertically graded scale to differentiate a customer product from another. According to ISO 9241-11:2018, “where usability is higher then expected, the system, product or service can have a competitive advantage (e.g. customer retention, or customers who are willing to pay a premium)”. The visuals will be thought to come in addition to the GDPR compliance seal and reflect the usability of the privacy implemented. The purpose will be the same as for the methodology, to help the businesses that have already achieved GDPR compliance to further differentiate themselves on the market. From the point of view of the user of the product, the visual scale would offer support for choosing the service or product that best respects her privacy expectations.

To further validate our UP Cube model and for exemplification, we are applying the UP criteria to three use cases taken from pilots done in an ongoing European project called Secure COnnected Trustable Things (SCOTT): (i) Assisted Living and Community Care System, (ii) Air Quality Monitoring for healthy indoor environments, and (iii) Diabetes App. These are examples of IoT systems [24–26] for which our model is especially relevant, as the privacy protection is even more variable and context-dependent. IoT technologies, due to their nature (i.e., ubiquity, invisibility, and continuous sensing) [21], are able to generate granular and intimate data about people and everything or everyone in their surroundings, by that reducing privacy to zero.
References


Making GDPR Usable: A Model to Support Usability Evaluations of Privacy*

Johanna Johansen\textsuperscript{1}\textsuperscript{⋆}\textsuperscript{1}[0000–0003–4908–9045] and Simone Fischer-Hübner\textsuperscript{2}

\textsuperscript{1} Department of Informatics, University of Oslo. 
johanna@johansenresearch.info
\textsuperscript{2} Department of Mathematics and Computer Science, Karlstad University. 
simone.fischer-huebner@kau.se

Abstract. We introduce a new model for evaluating privacy that builds on the criteria proposed by the EuroPriSe certification scheme by adding usability criteria. Our model is visually represented through a cube, called Usable Privacy Cube (or UP Cube), where each of its three axes of variability captures, respectively: rights of the data subjects, privacy principles, and \textit{usable privacy criteria}. We slightly reorganize the criteria of EuroPriSe to fit with the UP Cube model, i.e., we show how EuroPriSe can be viewed as a combination of only \textit{rights} and \textit{principles}, forming the two axes at the basis of our UP Cube. In this way we also want to bring out two perspectives on privacy: that of the data subjects and, respectively, that of the controllers/processors. We define usable privacy criteria based on usability goals that we have extracted from the whole text of the General Data Protection Regulation. The criteria are designed to produce measurements of the level of usability with which the goals are reached. Precisely, we measure effectiveness, efficiency, and satisfaction, considering both the objective and the perceived usability outcomes, producing measures of accuracy and completeness, of resource utilization (e.g., time, effort, financial), and measures resulting from satisfaction scales. In the long run, the UP Cube is meant to be the model behind a new certification methodology capable of evaluating the \textit{usability of privacy}, to the benefit of common users. For industries, considering also the usability of privacy would allow for greater business differentiation, beyond GDPR compliance.

Keywords: usable privacy \cdot Human-Computer Interaction \cdot usability goals \cdot usable privacy criteria \cdot privacy certification \cdot GDPR.


\textsuperscript{⋆⋆} The first author was partially supported by the project IoTSec – Security in IoT for Smart Grids, with nr. 248113. Thanks go also to Josef Noll for introducing me to the topic of privacy evaluations and labeling.
1 Introduction

1.1 Motivations

Privacy is a human right and an essential prerequisite for protecting fundamental human values, such as dignity and autonomy. However, the goal of protecting the privacy of the individuals and public has proved to be challenging to achieve. Currently, people have difficulty in understanding how their privacy is affected by the current practices in the technological world and are often unaware of its value and the implications of its loss. With our work we intend to build on the recent developments in the data protection law and support the business actors to invest into adopting privacy protecting measures. We offer a way to quantify the level of data protection and its usability in technological products and services. Displaying the achieved level of privacy protection, can be used by businesses to compete and differentiate themselves on the market. In addition, the measurements we produce can be translated into visual labels that can inform the users, at their respective level of understanding and interest, about how well a respective product respects their privacy. In this section we give an overview of the current state of privacy protection in technological products and society at large. This is to argue how our work could benefit the society and why the present context is propitious for such work.

With the boom of electronic commerce and the pervasive IoT (Internet of Things) and web services, people are producing enormous amounts of electronic data that are collected by various actors under very imbalanced privacy agreements (i.e., signing Terms of Services on a “take-it-or-leave-it” manner). The reasons are multiple; for one, it is difficult (if not impossible) for a normal person to know exactly what data her online behavior (or IoT device) produces. Then, much of this data is of private nature, but it is difficult to understand which and in what situation. Lastly, there are numerous and powerful algorithms (for search, machine-learning, etc.) that can make new (some unthinkable) inferences out of seemingly non-private pieces of data; and even more dangerous when multiple data sources are combined.

Digital privacy has eluded people up to the point that many have given up the hope to have both privacy and access to digital services; recognized even in the General Data Protection Regulation (GDPR)\(^3\) Recital 9 as “...widespread public perception that there are significant risks to the protection of natural persons, in particular with regard to online privacy”. One of the reasons, however, is that privacy is a complex concept. It has different personal and contextual connotations and larger social, ethical, legal and political implications that are difficult to grasp by laypeople [6]. In the presence of technology, digital privacy becomes even more complicated to understand. For example, there is a constant tension between how much privacy a person can have (i.e., how much of her activity she can keep only to herself) and how much authorities are allowed to see in order to ensure safety in the society (e.g., how many and where should

\(^3\) GDPR – General Data Protection Regulation from European Union [2].
public surveillance cameras be placed?). Another uncertainty surrounding privacy comes from the difficulty modern people have in separating their public and private lives (e.g., use of social media both with coworkers and friends; should (or not) data from browsing the Internet be shared with Internet companies or search engines, and when or for what functionality in return?).

When designing for usability, we also need to consider that the main goal of a person when using a piece of technology or a technological system is to fulfill the task or need that the product was intended for, e.g., to buy train tickets from a ticket machine. The primary goal of the train ticket buyer is not to check how well the machine protects her privacy, but to reach a certain destination. The buyer might even be in a hurry, so that it catches the train that leaves in 10 minutes. There will be no time left to read the privacy terms. And even if the buyer has time to read them, in case she does not agree with the terms, she still has to take the train to reach her destination. There might be a bus alternative, but the privacy terms of the bus company might be even more invasive [22].

The complexity of the privacy concept as such and of digital data and technology, make it difficult for one to evaluate the privacy properties of a specific piece of technology (e.g., web service, IoT product, or communication device). The difficulty is not only for average people, but also for regulators to check compliance, and for developers to be able to provide privacy-aware digital services/products/systems. Indeed, there are multiple concepts involved in digital privacy, like data sharing (which for normal business practices nowadays can form a highly intricate network of relationships), ownership and control of data, accountability or transparency (both towards the regulators as well as the users). Many of the privacy concepts are even a challenge by themselves, when it comes to their evaluation, since they are difficult to measure or to present/explain.

To solve the intricacies of privacy around the use of personal data in technologies, collaboration between specialists from different fields – such as computer science, interaction design, human ergonomics, law, and cognitive science – is needed. When a common understanding has been reached this should be translated into a form that can be grasped by regular people with limited time and level of expertise in fields such as technology or law. The work of [25] is an example of such initiative meant to bridge the computer science and legal approaches to privacy. (See more such examples in Section 2.)

People having difficulties in evaluating the implications of their choices and behavior in respect to privacy can be taken advantage of by the companies that have business interests in harvesting their data. Moreover, powerful and influential business actors, by using media channels, seek to further misinform people about the value of privacy and encourage them to give up their rights related to privacy and personal data protection.

One type of misconception about privacy that businesses try to spread is that “only the wrongdoers have something to hide and that the people showing
different faces in different situations lack integrity” [30, Chapter 10. Privacy]. Some renown comments that have shaped the public opinion in this negative way are:

- “You already have zero-privacy anyway, get over it.” (former Sun Microsystems CEO Scott McNealy, 1999) [24, – source of the citation]
- “If you have something that you don’t want anyone to know, maybe you shouldn’t be doing it in the first place.” (Google’s CEO Eric Schmidt, 2009)
- “Privacy is no longer a social norm.” [30, – source of the citation]
- “You have one identity. The days of you having a different image for your work friends or co-workers and for the other people you know are probably coming to an end pretty quickly. Having two identities for yourself is an example of a lack of integrity.” (Facebook’s Mark Zuckerberg, 2010) [30, – source of the citation]

The companies might be interested in subverting people’s privacy so that they can sell them more of their services and products. The information people produce is used mostly for advertising [30, Chapter 4. The Business of Surveillance]. As most of our modern life is conducted online, we do not have much choice, as we cannot simply stop using mobile phones, read or shop online or use e-mail. The solution is to motivate businesses to compete on being ethical and on protecting their users’ privacy. Some ways for businesses to achieve this could be: displaying privacy policies shortly and clearly in product descriptions, adopting labels of the energy consumption type, adopting privacy enhancing technologies (PETs), or/and becoming certified by privacy certifications schemes [26].

One example that consumer communities can do is to include privacy features in technology reviews (e.g., reviews for smartphones), as one main evaluation criterion, at the same level with, e.g., battery life-time. However, people need to be educated to ask for “privacy-as-a-feature”. The task of educating the public cannot be left in the hands of media, which might be controlled by businesses, or leave the users to learn/study privacy by themselves – “if people need to go to great length to protect their privacy, they won’t” [24] because privacy is a too complex matter for a layperson to be able to handle it alone. People need support and guidance in this respect. This can be done through making privacy-related information usable for the general public and motivate the businesses to adopt such an approach to data privacy.

1.2 Making privacy usable

For explaining the intricacies of privacy, besides research articles and books [16], there are several legislative texts adopted in different jurisdictions. GDPR in Europe makes a good effort in clarifying many aspects of data privacy, providing the legislative support to enforce better data protection practices on anyone (within its jurisdiction) collecting and processing personal data.6 However, these

6 The two concepts “privacy” and “data protection” are used interchangeably throughout this paper. The right to privacy, as emerged in 1984 in the Universal Declaration
A Model to Support Usability Evaluations of Privacy

regulations only specify the requirements on the data controllers (i.e., the organizations providing the service) in the form of basic principles, and the rights of the data subjects, but do not make any strict claims about the extent to which a controller (or processor) should go about implementing these requirements so that they are beneficial for the user, and to what degree.

As such, one motivation for usability evaluations of privacy is the fact that usability goals of GDPR, s.a. “… any information … and communication … relating to processing [to be provided] to the data subject in a concise, transparent, intelligible and easily accessible form, using clear and plain language, …” (Article 12 (1) of GDPR), are left open to the subjective interpretation of both evaluators and controllers. The provisions of GDPR regarding usability are too general and high-level to be suitable for a certification process [21]. To remedy this, we propose a set of criteria thought to produce measurable evaluations of the usability with which privacy goals of data protection are reached.

For evaluating privacy we take as starting point the methodology developed by EuroPriSe [4] that has as purpose to evaluate compliance with GDPR. We are guided by the EuroPriSe criteria when eliciting, what we call, principles and rights, which form the two variability axes at the basis of our model, i.e., which principles are followed and which rights are respected. However, EuroPriSe does not consider usability, which is the main focus of our work here. As such, one contribution of this paper is to show how to add usability aspects to the existing evaluation criteria of EuroPriSe.

Certification schemes (s.a. EuroPriSe) provide a seal showing compliance with data protection regulations and industry standards. In addition to such a certification, our evaluation measures on a scale how well data protection obligations are respected and how easy it is for a user to understand that. These measurements can be presented to the user in different ways, e.g., using “traffic light” scales, showing which level of usability has been reached by the privacy of a certain technological product. A “traffic light” presentation of privacy is recommended by [3, Chapter 6(235)] as a way to “foster competition” and “show good practice on privacy policies”.

The marketing motivation for the providers to adopt such a methodology is that the users tend to choose the product that answers best to their specific and real needs for privacy. Only the lack of alternatives in the market today explains why the data subjects still accept detrimental privacy conditions that would rather fit the interest and attitudes of other type of stakeholders (controllers of Human Rights, is closely related to the nowadays right to protection of personal data. Though the two rights are related, protecting similar values, the right to personal data is though broader as it applies to processing of all kinds of personal data, beyond data related to privacy.[16]. In addition, we address only the privacy issues concerning digital technology.

7 At the moment of writing the paper, EuroPriSe’s criteria catalog has not been approved pursuant to Article 42(5) GDPR and EuroPriSe GmbH has not been accredited as a certification body pursuant to Article 43 GDPR yet. EuroPriSe is dedicated to receiving the approval of its certification criteria and the accreditation as a certification body in accordance with Art. 42 f. GDPR asap.

117
and processors) [30]. The metaphor of the “dancing bear” of Alan Cooper [11] illustrates well this situation.

Creating alternative products or features is a way for the businesses to differentiate themselves. Now a new way of differentiation is the level of privacy protection offered, beyond the minimum required by GDPR, as well as how usable this is. Certification schemes such as EuroPriSe will give a product the seal of GDPR compliance [10], but as the GDPR compliance is mandatory, all businesses will seek to conform. Beyond this, our methodology would facilitate the differentiation between services by considering usability aspects when implementing measures for protecting privacy. Usability is known as a market differentiator, e.g., the ISO 9241-11:2018 standard [5] asserts that designing for usability helps with marketing of a product and with offering the user better customized choices.

Traditionally, usability is a quality related to the use of a product. In our case, we are not interested in the usability of a product per se, but only in those aspects of a product that concern privacy. Our conceptualization of usable privacy is based on the definition of usability as presented in the ISO 9241-11:2018, which we adapt to include privacy as follows:

**Usable privacy** refers to the extent to which a product or a service protects the privacy of the users in an efficient, effective and satisfactory way by taking into consideration the particular characteristics of the users, goals, tasks, resources, and the technical, physical, social, cultural, and organizational environments in which the product/service is used.

Our long term goal is to create a methodology to support service providers to make the privacy of their products more usable. The Usable Privacy Cube (UP Cube) described in Section 3 and the usable privacy criteria introduced in Section 6 are the first building blocks of the methodology we are aiming for. They are meant as tools, for both usability engineering experts and certification bodies, to evaluate if a product was designed to respect and protect the privacy of its users in an usable way. Once privacy measures and privacy enhancing technologies are integrated into the design of a product, it still remains to find out if (and how much or to what extent) those measures empower and respect the rights of their particular user as intended. In Human-Computer Interaction (HCI) this is determined based on user testing and usability evaluations. The criteria we propose presume the use of such established HCI methods for usability evaluations (e.g., [14]).

Furthermore, the usable privacy criteria that we propose for evaluating how efficient or effective a product is in protecting privacy require the technology providers to take into consideration the context of use, including the characteristics and needs of different types of users. We adopt the definition of context of use proposed by ISO 9241-11:2018:

“[context of use] comprises a combination of users, goals, tasks, resources, and the technical, physical and social, cultural and organizational environments in which a system, product or service is used.”
The legislation does not directly refer to usability goals and context of use as known in the ergonomics/human factors or human-centered design. However, requirements as the one in the Recital (39) of GDPR asking for the information addressed to the data subject to be “easily accessible and easy to understand” are categorized in this paper as usability goals, for which we create usable privacy criteria meant to measure effectiveness, efficiency and satisfaction – as usability outcomes – with regard to privacy aspects (we henceforth call these *Usable Privacy criteria*, and abbreviated it as UP criteria).

After a short digression into related work in Section 2, we introduce in Section 3 the UP Cube model, which is the main contribution of this work. We then continue to detail the UP Cube in the rest of the paper. Section 4 presents the EuroPriSe in the new light of the UP Cube, forming the two axes of criteria at its basis. The third vertical axis of the UP Cube, a genuine contribution of this paper, is formed of the UP criteria detailed in Section 6. To the best of our knowledge, there is not other work that extends privacy certification schemes with usability criteria. Section 5 presents usable privacy goals that the criteria are meant to measure. The UP Cube naturally captures Interactions between all the axes, which we talk about in Section 7. We conclude in Section 8, presenting also some avenues for further work.

## 2 Putting the work into context

*Usable privacy and security.* The present work can be placed in the research field called *usable privacy and security*, with seminal works s.a. [7,17,34,12] and conference series s.a. the Symposium On Usable Privacy and Security (SOUPS). We consider that research on privacy requires, even more than security, an interdisciplinary approach (encompassing the expertise coming from research fields such as Psychology, Law or Human-Computer Interaction). As [6] points out, privacy has its meaning rooted in larger cultural and social practices and has political, ethical as well as personal connotations.

There have been considerable efforts towards including specialist from different areas of research on issues related to privacy. Examples of such efforts are the constitution of The Privacy & Us Innovative Training Network (ITN)\(^8\) or the organization of the IFIP Summer School on Privacy and Identity Management\(^9\).

Other examples of cross-disciplinary research efforts come from the automation of privacy agreements (or Terms of Services – ToS) where the goal is to presented ToS in an accessible way to the general user. Notable contributions in

---

\(^8\) This project has received funding from the European Unions Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 675730, within the Marie Skłodowska-Curie Innovative Training Networks (ITN-ETN) framework; https://privacyus.eu/

\(^9\) https://www.ifip-summerschool.org/
this regard are the endeavors of the LeDA network\textsuperscript{10}, The Usable Privacy Policy Project\textsuperscript{11} or the CLAUDETTE project\textsuperscript{12}.

Regarding the relation between security and privacy, in this paper we consider security as one integral aspect of privacy, where privacy implies security but not the other way around. We consider such a clarification necessary, as we have seen a tendency in the general public to equalize the meanings of the two terms in favor of security. In computer science, privacy research has been closely intertwined with security research, reflected e.g. in the contents and the structure of the book [13]. However, in this paper, we favor the term “usable privacy”, as it includes by default security, which is in accordance with the data protection legislation, where security (integrity and confidentiality) is specified as one of the several principles to abide by in order to assure the privacy of users’ data.

\textit{Human-Computer Interaction.} Having the goal to evaluate the usability of privacy in technological systems and products, makes our work part of the larger HCI research on privacy [6,23,22,28]. Following the classifications made by Iachello and Hong in their review [19], we approach privacy from a “data protection” perspective by extracting usability related goals from the GDPR. A similar approach is taken in [27,28], which translates legislative clauses of the Directive 95/46/EC (now replaced by GDPR) into interaction implications and interface specifications. Similarly, [24] develops principles for guiding system design based on fair information practices found in the US Privacy Act of 1974 and the EU Directive 95/46/EC. The model we propose integrates well with a user-centered design where HCI methods are applied to elicit requirements based on understanding the users, their needs and the context of use.

For evaluating how well a product meets privacy requirements, context of use variables s.a. user capabilities, tasks, the field where the technology is going to be deployed (e.g., healthcare, industrial facilities), should be defined. Preferably these definitions should be established in the requirements phase of a product’s lifecycle, but definitely these would be defined and considered when running the privacy evaluation based on the UP criteria that we propose here. We thus adopt the ergonomic approach from ISO 9241-11:2018 where \textit{usability is always considered in a specified context of use}, since the usability to be applied to a certain technology can be significantly different for varied combinations of users, goals, tasks and their respective contexts.

\textsuperscript{10}The Legal Design Alliance (LeDA) is formed of lawyers, designers, technologists, academics, and other professionals who are committed to making the legal system more human-centered and effective, through the use of design. https://www.legaldesignalliance.org/
\textsuperscript{11}The Usable Privacy Policy Project, https://usableprivacy.org/. Visit also https://explore.usableprivacy.org/ to navigate privacy policy annotations extracted by both humans and machine learning techniques.
\textsuperscript{12}CLAUDETTE (automated CLAUse DETectEr), http://claudette.eui.eu/about/index.html. See also their tool http://www.claudette.eu/gdpr/
ISO/IEC29100:2011 gives a good example of how the context of use is decisive for establishing if a certain type of information can be used to identify a natural person [1, 4.4.2 Other distinguishing characteristics, p.7]:

“The last name of a person is insufficient to identify a person at a global scale, but might be enough to identify that person at a company level.”

We also make a distinction between user experience goals and usability goals, focusing in this paper on the latter [29]. User experience goals are concerned with how users, as individuals, are perceiving a product. As the nature of our work is to find criteria that can be generalized to groups or types of people, to be measurable, and part of an evaluation, usability goals will be more appropriate for such a function. This difference is clearly stressed by the ISO 9241-11:2018 standard, which states that usability typically deals with goals sheared by a user group, while user experience has more emphasis on individual goals.

3 The Usable Privacy Cube model

We devise a model for organizing the criteria to use in privacy evaluations and measurements, and represent it as a cube with three axes of variability (see Fig. 1), which we call the Usable Privacy Cube (UP Cube). The two axes found at the base of the UP Cube are composed of the existing EuroPriSe criteria, which we slightly reorganize in the Section 4 to fit in one of the two categories: data protection principles or rights of the data subjects.

We want to emphasize two perspectives on privacy that the UP Cube represents (hence our restructuring of the EuroPriSe criteria): the perspective of the controllers and of the data subjects. The controllers are thus given an overview of the principles that they are obliged to follow, whereas the data subjects are offered an overview of their rights.

Fig. 1. A generic version of the cube with the three axes of variability: data protection principles, the rights of the data subjects, and usable privacy criteria.
The UP Cube allows to visualize interactions between the axes, made easier by our separation of the criteria into the three categories. Each such intersection has its specifics and could be studied in itself; we identify a few exemplary points of intersection between the axes in Section 7.

**Example 1.** The intersection between the transparency principle and the right to be informed is identified in Article 12 of GDPR. The controllers are obliged to provide the data subject information that should be concise, transparent, intelligible and in easily accessible form, using clear and plain language.

The third vertical axis of the cube is composed of our UP criteria, presented in Section 6. The UP criteria are determined based on usable privacy goals and are evaluated considering the context of use by following the guidelines in the ISO 9241-11:2018 standard. Each of these criteria has several subcriteria intended for measuring the usability level by using different methods and respective tools from HCI, depending on what the criterion asks to be measured [29]. Interactions exist also with this third axis.

**Example 2.** For the case presented in Example 1, in order to establish how easily accessible or clear the information is, we must measure the level of efficiency, effectiveness and satisfaction in a specific context of use. Efficiency implies measuring the time and effort spent by a specific user for finding the information needed and for understanding it. Effectiveness measures the completeness with which a goal was achieved. In this case we would like to know how much of the needed information was the specific user able to access and understand. At the same time, what a certain type of user perceives as intelligible information, might be perceived by another as difficult to comprehend. Establishing the perceived characteristics of information is an activity categorized under the satisfaction usability outcome.

The UP Cube also brings the idea of *orderings* on each axis, hence the arrows. Such orderings are important for several reasons, e.g., UP criteria can be ordered based on “how little effort is required to evaluate it compared to how much overall evaluation outcome it entails” or “covers most technologies”. Usual for certification methods is to use a decision tree order to capture the impact of each criterion (e.g., choosing the most discriminating first), thus which to prioritize in the evaluation.

Judging from practice, one is inclined to think that an ordering is not always possible to find as some principles are equally important, therefore the orders are not necessarily strict. Moreover, one can even see one principle as more important than another only in some industry or context, whereas in a different industry the same two principles would be ordered the other way, therefore one may think that the orders are only partial (i.e., not total). However, in a specific cube (i.e., used in a specific methodology by a specific authority for privacy usability evaluations in a specific industry and context) there must always be an ordering in which the criteria should be applied. One can always generate a strict and total order from a partial order by just taking a random decision.
on ordering two criteria when no reasonable order exists. For example, one can any time pick as default order the one arising from the textual placement of the criteria in the data protection legislation texts (maybe considering content from articles as more general than content from recitals), or in the EuroPriSe (or the regulator/company) catalogs. What is certain is that each use case or industry has its specific requirements from which a meaningful ordering would be created.

Forming a specific UP Cube, i.e., deciding on the precise details of each criteria on the three axes and the orderings, is to some degree dependent on the specific context of use for the respective product to be evaluated. Therefore, one can think of infinitely many cubes, one for each different context. The criteria will not be different between the cubes, but their scope, depth, and evaluation might be different, depending on the context.

The context of use as such is not mentioned in GDPR, but the context of processing is brought up often [2, e.g., the recitals 43, 47, 71, 74, 76]. The context of processing, as defined in the legislation, overlaps and has similar purpose with the notion of context of use defined in the ISO standard. However, unlike usability engineering/HCI where the context is a general concern, the data protection law requires the context to be considered only in certain special situations, e.g., when evaluating the risks to the rights and freedoms of natural persons. We go beyond this and consider for each proposed criterion the context and the group of users that the evaluation aims at.

4 EuroPriSe

EuroPriSe originated from the Schleswig-Holstein Data Protection Seal, which was led by the Schleswig-Holstein Data Protection Authority (DPA) from ca. 2001 until the end of 2013, when it was transferred to the EuroPriSe GmbH company. The scheme has a history of eighteen years [18] and is one of the oldest privacy and data protection seals based on a law, i.e., the State Data Protection Act of the German federal State Schleswig-Holstein. The role of the seal is to help the vendors of IT products and services to comply with the data protection requirements derived from the applicable law in Europe [8,10,26]. EuroPriSe, in collaboration with Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein (ULD), received support from EU to establish a trans-European privacy seal. EuroPriSe is now intended to provide EU-wide privacy certifications that assure compliance with European data protection law. In addition, the EuroPriSe criteria are already updated to consider the fairly new GDPR.

We have chosen EuroPriSe as the basis for our UP Cube because of its long history, its continuous improvement, strong list of well-developed criteria, being led in the past by a DPA, and being based on the European data protection legislation. EuroPriSe also integrates with widely acknowledged IT security certification methods s.a. ISO 27000 and the The Standard Data Protection Model.[13]

[13] Following the requirement for a consistency mechanism set out in the Article 63 of GDPR, the work of the certifications bodies and DPAs in Germany is coordinated and made consistent through “The Standard Data Protection Model” (https://www.}
The way the criteria are formulated, as questions, also fits with the form of our usable privacy evaluation criteria. In addition, the existing EuroPriSe evaluation, which is at the basis of our model, assures that the GDPR legal grounds are covered, including data protection principles and duties and data subject rights. The UP criteria evaluations come on top, fine-graining the EuroPriSe evaluation with usability measurements, showing how well the legislation is respected.

Another feature that is relevant for our user-centered approach is that the EuroPriSe criteria catalog has been updated to include the data protection by default paradigm, promoting built-in data protection and privacy-friendly default settings. Moreover, EuroPriSe takes into account the technical, organizational and legal framework within which the product or service is operated and asks for considering the requirements of all the parties involved in the system, aiming at strengthening the position of the data subjects. Our work shares with EuroPriSe its high-level goal of making transparent for the general public how companies are managing data protection in their products and services.

4.1 The EuroPriSe criteria in the UP Cube structure

In order to build on EuroPriSe, we first look into how its methodology fits with our UP Cube model. We show how EuroPriSe criteria can be redistributed into one of the two axes at the basis, i.e., as either rights of the data subjects or as privacy principles, or otherwise as a context of use criterion. Table 1 gives an overview of this redistribution. The distinction between principles and rights is inspired by the structure in [16], where principles and rights represent the core of this handbook. One purpose of the principles, mentioned in [16], is to serve as the starting point when interpreting the more detailed provisions in the subsequent articles of data protection law. The law also requires that these principles should correspond to the rights presented in the articles 12 to 22. This correspondence can be visualized through the intersection between the respective rights and principles axes of the UP Cube.

In the following subsections we detail the Table 1.

4.1.1 Data protection principles. EuroPriSe has a dedicated subset of criteria dealing with data protection principles, i.e., “2.5 Compliance with General Data Protection Principles”. In addition, several other criteria sets from EuroPriSe can be related to data protection principles, as detailed below.

– The second part of the section “C. Target of Evaluation (ToE)”, called “Regulatory Analysis”, as well as the “Purpose(s)” part of the Subset 1.1.1, refer to the principle of purpose limitation. This principle requires that the purpose of data processing must be defined before processing is started.

[datenschutz-mv.de/datenschutz/datenschutzmodell/](datenschutz-mv.de/datenschutz/datenschutzmodell/), issued by the Conference of the Independent Data Protection Authorities of the Bund and the Länder (Germany) on 9-10 November 2016. This document is a good reference for methods and guidance for implementing the data protection principles.
EuroPriSe Criteria: We list the names of (sub)sections as appearing in the EuroPriSe document [4], which has two parts, the second being subdivided into four sets of criteria, whereas the first contains preliminary issues, from where only section C is relevant for us.

<table>
<thead>
<tr>
<th>C. Target of Evaluation (ToE)</th>
<th>✓</th>
<th>✓</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1.1 Processing Operations; Purpose(s)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>1.1.2 Processed Personal Data</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>1.1.3 Controller</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>1.1.4 Transnational Operations</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>1.2.1 Data Protection by Design and by Default</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>1.2.2 Transparency</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>2.1 Legal Basis for the Processing of Personal Data</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>2.2 General Requirements</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>2.3.1 Data Collection (Information Duties)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>2.3.2 Internal Data Disclosure</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>2.3.3 Disclosure of Data to Third Parties</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>2.3.4 Erasure of Data after Cessation of Requirement</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>2.4.1 Processing of Data by Joint Controllers</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>2.4.2 Processing of Data by a Processor</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>2.4.3 Transfer to the Third Countries</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>2.4.4 Automated Individual Decisions</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>2.4.5 Processing of Personal Data Relating to Children</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>2.5 Compliance with General Data Protection Principles</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Set 3: Technical-Organisational Measures</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Set 4: Data Subjects’ Rights</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 1. Overview of the the EuroPriSe criteria categorized to fit into our UP Cube model, i.e., as the two axes with Principles and Rights, as well as Context of use.

- The “1.2.1 Data Protection by Design and by Default” refers to the data minimization principle. This is pointed out in a note introducing the subset.
- The criterion “How long are the data retained? Is this no longer than necessary for the purposes concerned?” of the “1.2.1.1 Data protection by Design” refers to the storage limitation principle.
- “1.2.2 Transparency” relates to the transparency of processing principle.
- “2.1 Legal Basis for the Processing of Personal Data” refers to the principle of lawfulness. Consent of the data subject or another legitimate ground (provided in the data protection legislation) are required as legal basis for processing personal data. This subset expounds as well on aspects that we categorize as belonging to the context of use.
- The following sets of criteria are related to the accountability principle:
  - The subset “2.2.1 Record of Processing Activities” details how controllers can facilitate compliance with the accountability requirement through recording processing activities and making them available to the super-
visory authority upon request (this is also explained in [16, 3.7. The accountability principle]).

- The subset “2.2.2 Designation of a Data Protection Officer” details how controllers can facilitate compliance with the accountability requirement through designating a data protection officer who is involved in all issues relating to personal data protection (this is also explained in [16, 3.7. The accountability principle]).

- The subset “2.2.4 Data Protection Impact Assessment” details how controllers can facilitate compliance with the accountability requirement through undertaking data protection impact assessments for types of processing likely to result in high risks to the rights and freedoms of natural persons (this is also explained in [16, 3.7. The accountability principle]).

- The subset “2.2.5 Prior consultation” details how compliance is promoted through prior consultation of the relevant supervisory authority if the impact assessment indicates that processing presents risks that cannot be mitigated (this is also explained in [16, 4.3. Rules on accountability and promoting compliance]).

- The following sets are related to the data security principle:
  - The subset “2.2.6 Notification of a personal Data Breach” details how the controller is required to notify the competent supervisory authority without undue delay, in cases where a personal data breach (with risks for rights and freedoms of individuals) takes place. The data subject being concerned needs to be informed as well (this is also explained in [16, 3.6. The data security principle]).

4.1.2 The rights of the data subjects. EuroPriSe has a dedicated subset of criteria dealing with data subjects’ rights: “Set 4: Data Subjects’ Rights”. In addition, several other criteria sets from EuroPriSe can be related to rights of the data subjects, as detailed below.

- “2.3.1 Data Collection (Information Duties)” refers to the right to be informed, following [2, Articles 12, 13 and 14].
- “2.3.4 Erasure of Data after Cessation of Requirement” refers to the right to erasure, following [2, Article 17].
- “2.4.4 Automated Individual Decisions” refers to the rights related to automated individual decision-making, following [2, Article 22].

4.1.3 Mixed and context of use

- “2.3.2 Internal Data Disclosure” and “2.3.3 Disclosure of Data to Third Parties” refer to a mixture of rights and principles. These subsets have pointers to the GDPR articles they are based on, which are indicators for which rights
and principles the sets can be linked to. Article 5(b) refers to the principle of purpose limitation, (c) to the principle of data limitation, (f) to the data security principle, article 6 to the lawfulness of processing principle, while articles 13 and 14 to the right to be informed.

- Section “C. Target of Evaluation” and other subsections from 1.1 and from 2.4 are seen as relevant for defining the context of use.
- Criteria relevant for defining stakeholder groups are found in “2.2.7 Processing under the Authority of the Controller or Processor”, “2.4.1 Processing of data by Joint Controllers”, and “2.4.2 Processing of Data by a Processor”, and regard obligations of the controllers, processors, joint controllers, and third parties. These groups are also mentioned in [16, 2.3. Users of personal data] as users of the data. The criteria set “2.2.3 Designation of the Representative in the EU” defines representatives within the EU; these are also mentioned in [16, Data protection terminology], in the context of defining the controllers group. When the controller is established outside the EU it needs to appoint a representative within the EU territory.
However, some of the criteria included in these sets could also be related to principles or rights, e.g., the criterion “Does the processor adhere to an approved code of conduct or an approved certification mechanism?” (“2.4.2 Processing of data by a Processor”) belongs to the accountability principle, which states that the controllers must be able to demonstrate compliance with data protection provisions.

5 Usable Privacy Goals

We identify usable privacy goals (henceforth called Usable Privacy goals, and abbreviated as UP goals) that appear in the GDPR text. These guide the work in Section 6 where we present the UP criteria meant to measure to what extent these goals are being achieved. The goals are listed in the order they appear in the legislation. The words emphasized in each goal relate to usability. The chosen words are those that can be interpreted differently based on the context they are used in, and can result in objective and perceived measurements when evaluated in usability tests. These words also capture goals that can be achieved up to certain degrees, and thus can be translated into a level in an evaluation scale. A list of the Recitals and Articles from GDPR where the goals were extracted from can be found in full in the Annex A. In addition to the GDPR, there are more specific data protection laws, such as the proposed ePrivacy Regulation, that have implications for usability, from where one could eventually extract additional usability goals.

**UPG.1** Ensuring a high level of protection of personal data. [Recital (6) of GDPR]

**UPG.2** Natural persons should have control of their own personal data. [Recital (7) of GDPR]
UPG.3 Consent should be given by a clear affirmative act establishing a freely given, specific, informed and unambiguous indication of the data subject’s agreement to the processing of personal data relating to him or her. [Recital (32) of GDPR]

UPG.4 If the data subject’s consent is to be given following a request by electronic means, the request must be clear, concise and not unnecessarily disruptive to the use of the service for which it is provided. [Recital (32) of GDPR]

UPG.5 Any information and communication related to the processing of personal data should be easily accessible and easy to understand. [Recital (39) of GDPR]

UPG.6 Any information and communication related to the processing of personal data should use clear and plain language. [Recital (39) of GDPR]

UPG.7 Make the natural persons aware of risks, rules, safeguards and rights in relation to the processing of personal data. [Recital (39) of GDPR]

UPG.8 Make the natural persons aware of how to exercise their rights in relation to processing of personal data. [Recital (39) of GDPR]

UPG.9 The specific purposes for which personal data are processed should be explicit. [Recital (39) of GDPR]

UPG.10 The personal data should be adequate, relevant and limited to what is necessary for the purposes for which they are processed. [Recital (39) of GDPR]

UPG.11 Personal data should be processed only if the purpose of the processing could not reasonably be fulfilled by other means. [Recital (39) of GDPR]

UPG.12 In the context of a written declaration on another matter, safeguards should ensure that the data subject is aware of the fact that and the extent to which consent is given. [Recital (42) of GDPR]

UPG.13 A declaration of consent pre-formulated by the controller should be provided in an intelligible and easily accessible form, using clear and plain language and it should not contain unfair terms. [Recital (42) of GDPR]

UPG.14 The data subject should have genuine and free choice in giving the consent. [Recital (42) of GDPR]

UPG.15 The data subject should be able to refuse or withdraw consent without detriment. [Recital (42) of GDPR]
UPG.16 Carefully assess the existence of a legitimate interest of a controller taking into consideration the reasonable expectations of data subjects based on their relationship with the controller. [Recital (47) of GDPR]

UPG.17 Assess if the interests and fundamental rights of the data subject could override the interest of the controller where personal data are processed in circumstances where data subjects do not reasonably expect further processing. [Recital (47) of GDPR]

UPG.18 Any information addressed to the public or to the data subject should be concise, easily accessible and easy to understand. [Article 12 (1) and Recital (58) of GDPR]

UPG.19 Any information addressed to the public or to the data subject should use clear and plain language. [Article 12 (1) and Recital (58) of GDPR]

UPG.20 Any information addressed to the public or to the data subject should use, when appropriate, visualization. [Recital (58) of GDPR]

UPG.21 Provide information of the intended processing in an easily visible, intelligible and clearly legible manner. [Article 12 (7) and Recital (60) of GDPR]

UPG.22 Provide a meaningful overview of the intended processing. [Article 12 (7) and Recital (60) of GDPR]

UPG.23 A data subject should have the right of access to personal data which have been collected concerning him or her, and should exercise that right easily and at reasonable intervals, in order to be aware of, and verify, the lawfulness of the processing. [Recital (63) of GDPR]

UPG.24 Allow the data subjects to quickly assess the level of data protection of relevant products and services. [Recital (100) linking to Article 42 of GDPR]

UPG.25 The request for consent should be presented in a manner which is clearly distinguishable from the other matters, in an intelligible and easily accessible form, using clear and plain language. [Article 7 (2) of GDPR]

UPG.26 It should be as easy to withdraw as to give consent. [Article 7 (3) of GDPR]

UPG.27 Facilitate the exercise of the data subjects rights under Articles 15 to 22 – right of access, right to rectification, right to erasure, right to restriction of processing, right to data portability, right to object and automated individual decision-making. [Article 12 (2) of GDPR]

UPG.28 The data subject should obtain from the controller meaningful information about the logic involved, as well as the significance and the envisaged
consequences of automated decision-making, including profiling to which s/he is object to. [Article 15 (1) (h) of GDPR]

**UPG.29** The right to object should be explicitly brought to the attention of the data subject and should be presented clearly and separately from any other information, at the latest at the time of the first communication with the data subject. [Article 21 (4) of GDPR]

**UPG.30** The data subject should have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her. [Article 22 (1) of GDPR]

### 6 Usable Privacy Criteria

The criteria presented in this section can be used in an evaluation process for establishing the level of effectiveness, efficiency, and satisfaction with which the goals from Section 5 are reached, wrt. a specific context of use.

The proposed criteria are always measurable, which makes the results of a privacy evaluation easier to present visually through the use of a privacy labeling scheme. The use of privacy labels will then fulfill the goal UPG.24. This goal has a special significance from a usability point of view as it reduces considerably the effort spent by the data subject for evaluating privacy, which for most users is not the primary task [6] and it gets in the way of buying or using a product or service.

Evaluating privacy and compliance with GDPR is done by certification bodies, providing seals and marks with the purpose of enhancing consumer trust and promoting transparency and compliance with the data protection regulations. Prior to GDPR, the lack of legal constraints, disconnection from official regulatory oversight, and lack of effective enforcement has resulted in inaccurate, false, or outdated privacy certificates. This made the existing certification schemes to lose their trustworthiness with the users (e.g., see in [15] criticisms of TRUSTe – now known as TrustArc¹⁴). Though the certification is still voluntary, GDPR endorses and facilitates a certification mechanism as a means to demonstrate compliance with data protection provisions. In addition, the existence of a certificate makes the process of choosing processors easier for the controllers too, especially so since GDPR establishes responsibility and liability for any processing carried out on the controller’s behalf [21].

### 6.1 General considerations for UP Criteria

Some of the goals have a more general purpose, and their achieved levels can be decided based on measurements coming from more specific criteria. In the case of the UPG.2, in order to establish if the data subjects have reached a high

¹⁴ [https://en.wikipedia.org/wiki/TrustArc#Criticism_and_Controversies](https://en.wikipedia.org/wiki/TrustArc#Criticism_and_Controversies)
level of “control of their own personal data”, the scores from evaluations of e.g., UPG.14 should also be high.

Another example of a general goal is UPG.11, which related to the data minimization principle. This goal is of special importance for our model since when there is no processing we give automatically the highest score on the evaluation scale. The criterion in the section “1.1.2.1 Personal data” of EuroPriSe: “Are any personal data processed when the product or service is used?” can be used to establish if personal data is being processed. This can be complemented by another EuroPriSe criterion, found in the section “1.2.1.1 Data Protection by design” (p. 18): “Is it possible to carry out the processing without the use of identifiable data all together?”. This criterion has the function to encourage the companies – if possible – to not process identifiable data at all.

The UPG.1 is another general goal, dealing with the protection of personal data in general. We use this goal to exemplify how a criterion should be formulated to consider usability:

What is the level of the usability of the personal data protection / privacy that the product or service ensures?

For being able to establish a level of how usable the privacy protection is, the evaluation needs to produce measurable outcomes. The structure that we follow is the one proposed in the ISO 9241-11:2018 where the measures consider both the objective and the perceived outcomes of usability (the UP criteria are labeled accordingly). The measurements will produce counts or frequencies (e.g., how many errors the user does when probed to do certain privacy related tasks) and continuous data (e.g., how much time does the user spend on completing a task related to privacy). The evaluation based on the UP criteria proposed below will produce three main categories of measures:

1. measures of accuracy and completeness,
2. resource utilization (time, effort, financial, and material resources), and
3. measures resulting from satisfaction scales.

In our structuring of the UP criteria we first give a high-level criterion, numbered e.g. UPC.1, which is closely related – which can also be seen from the wording – to one of the goals that we identified in Section 5. The score for a main UP criterion is established based on evaluations of more specific UP criteria, called subcriteria. The resources used to achieve a criterion, i.e., time, effort, financial, and material (which we abbreviate TEFM), should be measured to be able to determine the efficiency with which a specific criterion was reached. In addition, the results from the evaluations should show the level of perception that the data subjects have about their data being protected. The data subjects need to be highly satisfied with the offered privacy protection. The “high satisfaction” level is defined based on the user satisfaction evaluation of the respective subcriteria.

6.1.1 Notation and organization principles. We use here as example the first UP criterion, numbered UPC.1, to explain the organization of the UP criteria and the various notations that we use in the rest of this section.
Fig. 2. An overview of the distribution of usable privacy criteria in each category.

We categorize the UP criteria based on their area of application from the GDPR text. Figure 2 gives an overview of the number of criteria in each category.

1. Consent (lawful grounds for processing data principle).
2. Information and communication addressed to the public or to the data subject (transparency principle),
3. Rights of the data subjects (rights in general),
4. Purpose of processing,
5. Legitimate interest of either the processor or the data subject (lawful grounds for processing data principle).

A high-level UP criterion, like UPC.1, is labeled with the goal that it is related to, e.g.:

[Based on goal UPG.2].

Each UP criterion also is categorized using a label, e.g.:

>Type of criteria: generic.

For subsequent UP criteria we use a short version of these labels, which should be self-explanatory, e.g., we label UPC.2 by the goal that it aims at [UPG.18].

For each high-level UP criterion we identify several specific UP subcriteria, numbered accordingly, e.g., UPC.2.4.

Considerations regarding the classification of UP subcriteria. We then classify each UP subcriterion (e.g., from UPC.2.1 to UPC.2.9) into either effectiveness, efficiency, or satisfaction, and label it accordingly; e.g.:

a short version of the label */Effectiveness* would be */Es*;

similarly */Ef* and respectively */S*.

For the first UP criterion UPC.1, we give objective and perceived subcriteria. This is the structure we recommend to be used in a real evaluation. However, with
the intention of reducing the complexity of the present paper, for the subsequent UP criteria, we only label the various subcriteria with the respective labels and sublabels of effectiveness, efficiency, and satisfaction, e.g.:

UPC.1.1 is labeled with [Effectiveness] and [Measure:Objective].

We try to be exhaustive in our UP subcriteria and to give enough questions to cover all major aspects that need to be measured to achieve the respective goal that the high-level UP criterion relates to. The UP subcriteria are labeled with sublabels representing various specific measures of usability for the above three general categories, e.g.:

[Effectiveness:Completeness] or
[Satisfaction:Cognitive responses].

Considerations regarding the context of use. The specific context of use needs to be considered for each of our questions. To avoid repetition, we only give one example of how the questions should be formulated so that they relate to the context. This formulation applies to all the questions we propose in this section. For the example of UPC.1 one would read it:

– Without context
  What is the level of control the data subjects have over their data?
– With context
  What is the level of control the specified type of data subjects have over their data in the specified context of use?

6.2 List of Usable Privacy criteria

Since our criteria are modular (i.e., each high-level criterion is thought independent of the other) and can be ordered based on their importance for different application cases, they could be introduced gradually and selectively. It can be that certification bodies (like EuroPriSe) would start to include our UP criteria in their future test catalogs on an article-basis, e.g., a good candidate is Article 12 of GDPR (referring to rights that intersect with the transparency principle) as it contains five UP goals.

UPC.1 What is the level of control that the data subjects have over their data? [Based on goal UPG.2][Type of criteria: generic]

UPC.1.1 To what degree are the data subjects in control of the personal data? [Effectiveness][Measure:Objective]

UPC.1.2 What is the data subjects’ perceived level of control? [Effectiveness][Measure:Perceived]
UPC.1.3 How much [Time / Effort / Financial / Material resources] do the data subjects need to invest in order to have control over the processed data? [Efficiency:Time used, Human effort expanded, Financial resources expanded, Materials expanded][Measure:Objective]

UPC.1.4 How much [Time / Effort / Financial / Material resources] do the data subjects perceive that they need to invest in order to have control over the processed data? [Efficiency:Time used, Human effort expanded, Financial resources expanded, Materials expanded][Measure:Perceived]

UPC.1.5 How frequent do the data subjects use the data controlling tools put to their disposition? [Satisfaction][Measure:Objective]

UPC.1.6 What is the level of satisfaction of the data subjects with the achieved level of control? [Satisfaction][Measure:Perceived]

6.2.1 UP criteria related to information and communication

UPC.2 Is any information and communication addressed to the public or to the data subjects related to the processing of personal data concise, easily accessible and easy to understand? [UPG.18][Type of criteria: Information and communication addressed to the public or to the data subjects]

How much TEFM do the data subjects need to invest in order to [UPC.2.1 access, UPC.2.2 read through, UPC.2.3 understand] the information? [Efficiency:TEFM][O]

How much of the information were the data subjects able to [UPC.2.4 access, UPC.2.5 understand, UPC.2.6 read through]? [O][Es:Completeness]

UPC.2.7 To what degree the data subjects perceive the information as concise? [Satisfaction:Cognitive responses][P]

To what degree the data subjects perceive the information as easy to
[UPC.2.8 access, UPC.2.9 understand]? [Satisfaction:Cognitive responses][P]

Remark 1. The subcriteria in UPC.2 refer to cognition and understanding, while the subcriteria in UPC.3 refer to visual aspects of the information presented.

Remark 2. In different HCI works one can find different formulations that could seem related to how we formulate the subcriteria, e.g.: “Can the data subjects make sense of the information at all?”; “What is the extent to which the data subjects make sense of the information?”. However, we intend to measure the proportion of the information that is made sense of. Therefore we use formulations that give a statistically measurable outcome, such as “How much?”, “What is the percentage?”, “What is the degree?”. 
**UPC.3** Is the information about the intended processing provided in an easily visible, intelligible and clearly legible manner? [UPG.21][Type: Info]

How much TEFM do the data subjects need to invest in order to **[UPC.3.1]** see/locate, and **[UPC.3.2]** distinguish the information? [Ey:TEFM]

How well were the data subjects able to **[UPC.3.3]** visually locate and **[UPC.3.4]** distinguish the information? [Es:Accuracy]

How much of the information were the data subjects able to [ **[UPC.3.5]** visually locate and **[UPC.3.6]** distinguish]? [Es:Completeness]

To what degree the data subjects perceive the information as **[UPC.3.7]** easily visible, **[UPC.3.8]** intelligible, and **[UPC.3.9]** clearly legible? [S:Cognitive responses]

**Remark 3.** Poor visibility can affect the perception of trust, as information that has low visibility can appear to be hidden with a purpose. Poor legibility can reflect sloppiness in the way the content is produced, which again can give an impression of lack of professionalism. Poor visibility and legibility affects the satisfaction of the data subjects and it can cause physical discomfort (e.g., to the eyes, by having to read a text written in a very small font).

**UPC.4** Is any information and communication addressed to the public or to the data subjects related to the processing of personal data using clear and plain language? [UPG.19][Type: Info]

What is the level of **[UPC.4.1]** clearness and **[UPC.4.2]** plainness of the language? [Es:Accuracy]

**UPC.4.3** What is the percentage of the data subjects that understand the language? [Es:Completeness]

What is the portion of the language considered **[UPC.4.4]** plain and **[UPC.4.5]** clear]? [Es:Completeness]

How **[UPC.4.6]** clear and **[UPC.4.7]** plain do the data subjects perceive the language to be? [S:Cognitive responses]

**UPC.5** Is the information and communication addressed to the public or to the data subjects using, when appropriate, visualization? [UPG.20][Type: Info]

**UPC.5.1** How much of the data subjects' expended TEFM is reduced by the use of visualization? [Ey:TEFM]

**UPC.5.2** How well is the information understood when visualization is used, in comparison to when only text is used? [Es:Accuracy]

**UPC.5.3** How many data subjects consider the use of visualization in the evaluated context of use as appropriate? [Es:Accuracy][P]
UPC.5.4 What is the percentage of data subjects that understand the information better, when visualization is used? [Es:Completeness]

UPC.5.5 To what degree is the understanding of the information improved by the use of visualization? [Es:Completeness]

UPC.5.6 What is the level of satisfaction of the data subjects when visualization is used? [S:Cognitive responses]

UPC.5.7 How appropriate do the data subjects perceive the use of visualization to be for the evaluated context of use? [S:Cognitive responses]

Remark 4. Some of the subcriteria in UPC.5 mention the “understanding of information” in relation with visualization. However, measurements of other aspects, besides cognitive effort, such as how visualization improves the rate of finding and accessing information, should be evaluated here as well.

UPC.6 Are the data subjects provided a meaningful overview of the intended processing? [UPG.22][Type: Info]

UPC.6.1 How much of the data subjects’ expended TEFM is reduced by using the provided overview? [Ey:TEFM]

What is the percentage of the data subjects that [ UPC.6.2 use and UPC.6.3 understand the content better due to] the provided overview? [Es:Accuracy]

UPC.6.4 What is the degree of improvement that the overview brings to the understanding of the content by data subjects? [Es:Completeness]

UPC.6.5 What is the percentage of data subjects able to express the correct and intended meaning of the provided overview, when probed? [Es:Completeness]

UPC.6.6 How meaningful do the data subjects perceive the provided overview? [S:Cognitive responses]

UPC.7 Have the data subjects obtained from the controller meaningful information about <<the logic involved, as well as the significance and the envisaged consequences of automated decision-making, including profiling to which they are object to>>? [UPG.28][Type: Info]

Remark 5. To avoid repetition, in the subordinated subcriteria, we write the above text between angle brackets in the following short form: <<LOGIC>>.

UPC.7.1 How much TEFM do the data subjects need to invest in order to understand the information about <<LOGIC>>? [Ey:TEFM]
UPC.7.2 What is the percentage of data subjects being able to express the correct and intended meaning of the information provided – in respect to \( \text{LOGIC} \) – when probed? [Es:Accuracy]

To what degree do the provided information | UPC.7.3 affect the choices and actions and UPC.7.4 improve the understanding of the data subjects in respect to \( \text{LOGIC} \)? [Es:Accuracy] resp. [Es:Completeness]

UPC.7.5 How much of the provided information – in respect to \( \text{LOGIC} \) – is understood by the data subjects? [Es:Completeness]

UPC.7.6 How meaningful do the data subjects perceive the provided information in respect to \( \text{LOGIC} \)? [S:Cognitive responses]

6.2.2 UP criteria related to consent

Several UP goals are found in the consent related provisions. These provisions are evaluated in detail in the EuroPriSe sections 2.1.1.1 Processing on the Basis of Consent and 2.1.1.2 Processing on the Basis of a Contract. The criteria we generate here are meant to complement the ones in the EuroPriSe through bringing in usability concerns. Marc Langheinrich presents several of the problems with how consent can be misused [24]. One of these is the “take it or leave it” dualism where the person does not have a real choice and thus getting consent comes very closed to blackmailing. This problem has been ameliorated in the GDPR law by asking the controllers to allow for separate consent for different data processing operations. A usability evaluation could help further by revealing how the data subjects perceive the consenting act, as well as whether the data subjects consider consent a real choice and if the options to consent to some of the processing operations only, are satisfactory.

UPC.8 Is consent given by a clear affirmative act establishing a freely given, specific, informed and unambiguous indication of the data subjects’ agreement to the processing of personal data relating to them? [UPG.3][Type: Consent]

UPC.8.1 How much of the consent text do the data subjects understand? [Es:Completeness]

UPC.8.2 How much of the implications of consenting do the data subjects understand? [Es:Completeness]

To what degree do the data subjects perceive the agreement to be | UPC.8.3 freely given, UPC.8.4 informed, and UPC.8.5 unambiguous]? [S:Cognitive responses]

UPC.9 Are the consents of the data subjects given by following a request by electronic means? If yes, is the request clear, concise and not unnecessarily disruptive to the use of the service for which it is provided? [UPG.4][Type: Consent]
UPC.9.1 How much TEFM do the data subjects need to invest in order to understand the request? [Ey:TEFM]

UPC.9.2 How much of the request do the data subjects understand? [Es:Completeness]

UPC.9.3 How much of the TEFM needed to fulfill the tasks that the data subjects are currently doing, is being wasted by attending to the request? [Ey:TEFM]

UPC.9.4 To what degree do the data subjects perceive the request to be unnecessarily disruptive? [S:Cognitive responses]

UPC.10 In the context of a written declaration on another matter, are safeguards ensured so that the data subjects are aware of the fact that, and the extent to which, consent is given? [UPG.12][Type: Consent]

UPC.10.1 What is the percentage of the data subjects being able to show that they are aware of the fact that, and the extent to which, consent is given, when probed? [Es:Accuracy]

UPC.10.2 Is the level of awareness showed matching the intended degree of awareness? [Es:Completeness]

UPC.10.3 How sufficient are the safeguards taken to ensure that the data subjects are being aware of the fact that, and the extent to which, consent is given? [Es:Completeness]

UPC.10.4 How much TEFM do the data subjects need to invest in order to become aware of the fact that, and the extent to which, consent is given? [Ey:TEFM]

To what degree do the data subjects perceive [UPC.10.5 themselves as being aware, and UPC.10.6 that enough safeguards have been taken to help them become aware] of the fact that, and the extent to which, consent is given? [S:Cognitive responses]

UPC.11 Is the declaration of consent, pre-formulated by the controller, provided in an intelligible and easily accessible form, using clear and plain language, and not containing unfair terms? [UPG.13][Type: Consent]

How much TEFM do the data subjects need to invest in order to [UPC.11.1 access, UPC.11.2 read, and UPC.11.3 understand] the declaration of consent? [O][Ey:TEFM]

How is the TEFM spent in relation to the TEFM expected by [UPC.11.4 the controllers, or UPC.11.5 the data subjects]? Are the differences reasonable?
To what degree do the data subjects perceive [UPC.11.6] the terms as unfair, [UPC.11.7] the language of the declaration of consent as clear and plain, and [UPC.11.8] the declaration of consent as being intelligible and having an easily accessible form? [P][S:Cognitive responses]

Remark 6. The criterion UPC.11 is similar to the criteria UPC.2 and UPC.4, only that it refers to the declaration of consent (or terms of services), and thus we expect that besides the above subcriteria one would also employ subcriteria analogous to those in UPC.2.x and UPC.4.x.

UPC.12 Is the request for consent presented in a manner clearly distinguishable from the other matters? [UPG.25][Type: Consent]

What is the percentage of the data subjects able to [UPC.12.1] understand that their consent is requested, and [UPC.12.2] clearly distinguish the request for consent from the other matters] when probed? [P][Es:Accuracy]

UPC.12.3 How much TEFM do the data subjects need to invest in order to distinguish the request for consent from the other matters? [Ey:TEFM]

UPC.12.4 To what degree do the data subjects perceive the request for consent as clearly distinguishable from other matters? [P][S:Cognitive responses]

Remark 7. The criterion UPC.12 is to some extent similar to the criterion UPC.3 only that it talks about distinguishability of the declaration of consent (or terms of services), and thus one can expect more subcriteria similar to those from UPC.3 to be useful.

UPC.13 Do the data subjects have free and genuine choice in giving consent? [UPG.14][Type: Consent]

UPC.13.1 To what degree do the data subjects perceive the choice of consenting as free and genuine? [P][S:Cognitive responses]

UPC.13.2 Are the data subjects being offered any alternatives in case of not being able/not wanting to consent? [O][Es:Completeness]

UPC.14 Are the data subjects being able to refuse or withdraw consent without detriment? [UPG.15][Type: Consent]

UPC.14.1 How much TEFM losses are there for the data subjects, related to withdrawing the consent? [Ey:TEFM]

UPC.14.2 When evaluating the overall consequences for the data subjects, in case of withdrawing the consent, what is the degree of impact on the data subjects? [O][Es:Accuracy]
UPC.14.3 To what degree do the data subjects perceive that it is detrimental for them to refuse or withdraw consent? [P][S:Cognitive responses]

UPC.15 Is it as easy to withdraw consent as it is to give consent? [UPG.26]
[Type: Consent]

UPC.15.1 How much TEFM do the data subjects spend to withdraw consent? Compare this to the TEFM needed to give consent (i.e., sum up results from UPC.9.1, UPC.10.4, UPC.11.1-UPC.11.3, and UPC.12.3). [Ey:TEFM]

UPC.15.2 Do the data subjects perceive withdrawing of the consent similarly easy to giving consent? [P][S:Cognitive responses]

6.2.3 UP criteria related to data subject rights

UPC.16 Are the rights of the data subjects, under Articles 15 to 22, (i.e., right of access, right to rectification, right to erasure, right to restriction of processing, right to data portability, right to object, and rights related to automated individual decision-making) facilitated? [UPG.27][Type: Rights]

UPC.16.1 How much TEFM do the data subjects spend in order to exercise their rights? [Ey:TEFM]

UPC.16.2 How many of the rights under Articles 15 to 22 are facilitated; and to what degree? [Es:Completeness]

UPC.16.3 To what degree do the data subjects perceive that their rights are facilitated? [P][S:Cognitive responses]

UPC.16.4 What is the percentage of the data subjects that are able to exercise their rights with ease, when probed? [Es:Accuracy]

UPC.17 Are the data subjects being aware of how to exercise their rights in relation to processing of personal data? [UPG.8][Type: Rights]

UPC.17.1 To what degree do the data subjects feel that they are aware of how to exercise their rights in relation to processing of personal data? [P][S:Cognitive responses]

UPC.17.2 What is the percentage of data subjects being able to explain which are the ways they could use to exercise their rights in relation to processing of personal data, when probed? [Es:Accuracy]
UPC.18 Do the data subjects have the right of access to personal data that has been collected concerning them, and can they exercise this right easily and at reasonable intervals, in order to be aware of, and verify, the lawfulness of the processing? [UPG.23][Type: Rights]

UPC.18.1 How much TEFM do the data subjects spend in order to access the personal data that has been collected concerning them? [Ey:TEFM]

To what degree do the data subjects perceive [UPC.18.2 accessing the personal data as easy, UPC.18.3 the intervals they are given access to the data as reasonable, and UPC.18.4 themselves as being aware of the lawfulness of the processing]? [P][S:Cognitive responses]

What is the percentage of the data subjects [UPC.18.5 being able to access the personal data as easy as intended, UPC.18.6 found to be aware of the lawfulness of the processing, UPC.18.7 that can verify the lawfulness of the processing], when probed? [Es:Accuracy]

UPC.18.8 How much of the personal data concerning them are the data subjects being able to access? [O][Es:Completeness]

UPC.19 Is the right to object explicitly brought to the attention of the data subjects and presented clearly and separately from any other information, at the latest at the time of the first communication with the data subjects? [UPG.29][Type: Rights]

UPC.19.1 How much TEFM do the data subjects spend to find the information related to the right to object? [Ey:TEFM]

What is the percentage of the data subjects being able to [UPC.19.2 separate the right to object from any other information, and UPC.19.3 exercise their right to object] – when probed? [Es:Accuracy]

To what degree do the data subjects perceive [UPC.19.4 the right to object as clearly presented, and UPC.19.5 the way the right to object has been brought to their attention as explicit]? [O][Es:Completeness]

UPC.20 Are the data subjects being aware of risks, rules, safeguards and rights in relation to the processing of their personal data? [UPG.7][Type: Rights]

UPC.20.1 How much TEFM do the data subjects spend in order to become aware of the risks, rules, safeguards and rights in relation to the processing of their personal data? [Ey:TEFM]

UPC.20.2 How accurately can the data subjects remember which are the risks, rules, safeguards and rights in relation to the processing of their personal data, when probed? [Ey:Cognitive responses]
UPC.20.3 How many of the risks, rules, safeguards and rights in relation to the processing of their personal data are the data subjects being able to remember? [Es:Completeness]

UPC.20.4 To what degree do the data subjects feel that they are aware of the risks, rules, safeguards and rights to their personal data? [P][S:Cognitive responses]

UPC.20.5 What is the percentage of the data subjects being able to understand the implications of the risks, rules, safeguards and rights to their personal data? [P][Es:Accuracy]

6.2.4 UP criteria related to the purpose of processing

UPC.21 Is the specific purpose for which personal data are processed explicit? [UPG.9][Type: Purpose]

UPC.21.1 How accurately can the data subjects remember the purpose? [Ey:Cognitive responses]

UPC.21.2 How many of the purposes can the data subjects remember correctly when several purposes are given? [Es:Completeness]

UPC.21.3 What is the percentage of the data subjects being able to show that they know what is the purpose for which personal data are processed? [P][Es:Accuracy]

UPC.22 Is the personal data adequate, relevant and limited to what is necessary for the purposes for which they are processed? [UPG.10][Type: Purpose]

UPC.22.1 To what degree do the data subjects feel that the processing of their personal data are adequate, relevant and limited to what is necessary for the given purposes? [P][S:Cognitive responses]

UPC.22.2 How many aspects do the data subjects point out to be inadequate, irrelevant and less or more than necessary? [Es:Completeness]

6.2.5 UP criteria related to the legitimate interest of either the processors or the data subjects

UPC.23 Is the existence of a legitimate interest of a controller carefully assessed, taking into consideration the reasonable expectations of the data subjects based on their relationship with the controller? [UPG.16][Type: Legitimate]

UPC.23.1 How much TEFM do the data subjects spend to assess the legitimate interest of the controller? [Ey:TEFM]
**UPC.23.2** To what degree do the data subjects perceive the legitimate interest of the controller as carefully assessed? [P][S:Cognitive responses]

**UPC.24** Has it been assessed whether the interests and fundamental rights of the data subjects could override the interest of the controllers where personal data are processed in circumstances where the data subjects do not reasonably expect further processing? [UPG.17][Type: Legitimate]

**UPC.24.1** How much is known by the controllers about which are the circumstances where the data subjects do not reasonably expect further processing? How much of these knowledge have been confirmed by the data subjects? [Es:Completeness]

**UPC.24.2** Do the data subjects and their controllers have a mutual agreement on what is considered to be reasonable further processing? [Es:Accuracy]

### 7 Interactions between the three axes

Characteristic to the data legislation text is that it always refers to how principles and rights intersect and depend on each other. In this section, we give examples of such references found in the recitals of GDPR, relevant for some of the identified usability goals. The recitals, though not legally binding, are meant to provide more details to the GDPR’s articles. The lawfulness, fairness, and transparency of processing principles, and the right to be informed appear to be closely interrelated, having also the highest occurrence of usability goals.

1. The UP criterion UPC.1 refers to the control the data subjects have over their data. The criterion can be related to the right to data portability, through the Recital (68), where due to the aim of strengthening the control of the data subject, the “data subject should also be allowed to receive personal data concerning him or her, which he or she has provided to a controller in a structured, commonly used, machine-readable and interoperable format, and to transmit it to another controller ...”. The same UP criterion can also be linked to data security principle through the provision in the Recital (75) where the “risk to the rights and freedoms of natural persons” can result in data subjects being deprived of their rights and freedoms or prevented from exercising control over their personal data. The “risk to the rights and freedoms of natural persons” is also mentioned by the [16, pp. 131, 134] in the context of data security principle.

2. The UP criteria UPC.2 and UPC.4 are related to the transparency of processing principle, which is referred to directly in the Recital (58), where the respective goals are extracted from – “The principle of transparency requires that any information and communication related to the processing of those personal data ...” – as well as to the principles of lawfulness and fairness,
which are also directly referred to in the Recital (39) – “Any processing of
the personal data should be lawful and fair”.

3. The goals UPG.7 and UPG.8 relate to the fairness and transparency of
processing principles, and are placed under these respective categories also
by the [16, pp. 117, 120].

4. The goals UPG.9 and UPG.19 are mentioned in the context of the trans-
parency principle, in the Recital (39), where the information to be given
to the data subject relates to the purpose of processing. This connects the principle of transparency with the principle of purpose limitation in the case of UPG.9 and the principle of data minimization in the case of UPG.19.

5. The UP criterion UPC.22 is based on the goal UPG.10 extracted from the
the Recital (39) of GDPR. This criterion is mentioned in Recital (39) as one
of the requirements for complying with the transparency principle, while also
referring to the purpose of processing. This connects the present criterion
also with the principle of data minimization and in addition with the data
protection by design principle. The link between the last two principles can
also be seen in the EuroPriSe criteria catalog, where data minimization is
the focus of the [4, 1.2.1 Data Protection by Design and by Default, p.18].

8 Conclusion and Further Work

The benefits of the UP Cube model are multiple: (i) emphasizing both the per-
spectives of data subjects and of controllers; (ii) representing visually on the
three variability axes the existing rights and principles criteria from EuroPriSe,
together with our new UP criteria; (iii) visualizing intersections between the
three axes; (iv) allowing ordering of the criteria on each axis.

The theory behind our usability evaluation of privacy is based on the well
established standards ISO/IEC29100:2011 and ISO 9241-11:2018. We worked
directly with the GDPR text, guided by [16], which also inspired our structur-
ing of the EuroPriSe criteria into rights and principles. Our HCI and usability
perspective on privacy is influenced by the seminal works [7,17,6,23,22,28,12].

To build the UP Cube we have:

– identified from the GDPR text 30 UP goals,
– created 24 UP criteria, each with measurable subcriteria, and
– restructured the criteria of EuroPriSe, laid as the basis of the UP Cube.

Further Work. The UP Cube is meant as the groundwork for building a certi-
fication methodology, extending EuroPriSe to evaluate the usability of privacy.
The proposed UP criteria are designed to produce measurable evaluations, useful
for generating privacy labels in order to guide stakeholders when choosing tech-
nological products, by representing and visualizing the different levels of privacy.
To achieve this larger goal, one needs to investigate which existing HCI methods
for usability testing should be used for each of the UP criteria, and in what way.
One example of such a usability method for measuring the perceived usability of a system is the System Usability Scale (SUS) [9], a ten-item attitude Likert scale questionnaire. The standard [5, Annex B: Usability measurements] also gives examples of methods that produce measurements relevant for our UP criteria, s.a. observing the user behavior to identify the actual usability problems, or asking the users to carry out tasks in a real or simulated context of use and measuring the outcomes. The experts can also run heuristic evaluations following design principles, theories and standards from the design and cognitive fields. More concrete examples of HCI methods and how these could be used for privacy and security solutions can be found in [22].

Which methods are appropriate to use, the number of test persons, and other test related concerns, depend on contextual factors, s.a. the type of technology, users and industry. Defining the required context is what our model offers support for. However, more work (e.g., providing guidelines and examples) is needed on how the context of use can be established.

HCI practices conduct user studies throughout the whole lifecycle of a product. These studies are run by the company itself, with the help of HCI (User Experience or Interaction Design) experts. For certification, the accredited data protection assessors would be using the results provided by the company to answer the UP criteria questions. In the cases of not enough or not reliable results, the assessors can recommend/require further testing. It would be valuable to have guidelines, e.g., in the form of a check-list, to help the assessors with establishing if the results from the company are reliable and sufficient. Recommendations for the businesses are useful as well, to guide how to conduct privacy related user testing, so that the results would be reliable later for certification.

With the same goal of achieving a complete methodology that can be taken in use by the accreditation bodies, building on the present model, one could create a visual representation of the evaluation, i.e., a translation of the measurements of usability of privacy provided by the UP criteria into a visually appealing privacy label. This should serve as a vertically graded scale to differentiate a customer product from another. According to ISO 9241-11:2018, “where usability is higher than expected, the system, product or service can have a competitive advantage (e.g. customer retention, or customers who are willing to pay a premium)”. The visuals will be thought to come in addition to the GDPR compliance seal and reflect the usability of the privacy implemented. The purpose will be the same as for the methodology, to help the businesses that have already achieved GDPR compliance to further differentiate themselves on the market. From the point of view of the user of the product, the visual scale would offer support for choosing the service or product that best respects her privacy expectations.

To further validate our UP Cube model and for exemplification, we are applying the UP criteria to three use cases taken from pilots done in an ongoing European project called Secure COmmected Trustable Things (SCOTT): (i) Assisted Living and Community Care System, (ii) Air Quality Monitoring for healthy indoor environments, and (iii) Diabetes App. These are examples of IoT systems [33,32,31] for which our model is especially relevant, as the privacy
protection is even more variable and context-dependent. IoT technologies, due to their nature (i.e., ubiquity, invisibility, and continuous sensing) [24], are able to generate granular and intimate data about people and everything or everyone in their surroundings, by that reducing privacy to zero.

Acknowledgements

We would like to thank the anonymous reviewers for helping improve the paper, as well as the members of the PriSec group at Karlstad University, for useful discussions and help with the present research while doing a secondment there.

References

9 Annexes

9.1 Annex A: A list of the Recitals and Articles of GDPR, in full text, from which the usability goals have been extracted.

(6) Rapid technological developments and globalisation have brought new challenges for the protection of personal data. The scale of the collection and sharing of personal data has increased significantly. Technology allows both private companies and public authorities to make use of personal data on an unprecedented scale in order to pursue their activities. Natural persons increasingly make personal information available publicly and globally. Technology has transformed both the economy and social life, and should further facilitate the free flow of personal data within the Union and the transfer to third countries and international organisations, while ensuring a high level of the protection of personal data.

(7) Those developments require a strong and more coherent data protection framework in the Union, backed by strong enforcement, given the importance of creating the trust that will allow the digital economy to develop across the internal market. Natural persons should have control of their own personal data. Legal and practical certainty for natural persons, economic operators and public authorities should be enhanced.

(32) Consent should be given by a clear affirmative act establishing a freely given, specific, informed and unambiguous indication of the data subject’s agreement to the processing of personal data relating to him or her, such as by a written statement, including by electronic means, or an oral statement. This could include ticking a box when visiting an internet website, choosing technical settings for information society services or another statement or conduct which clearly indicates in this context the data subject’s acceptance of the proposed processing of his or her personal data. Silence, pre-ticked boxes or inactivity should not therefore constitute consent. Consent should cover all processing activities carried out for the same purpose or purposes. When the processing has multiple purposes, consent should be given for all of them. If the data subject’s consent is to be given following a request by electronic means, the request must be clear, concise and not unnecessarily disruptive to the use of the service for which it is provided.

(39) Any processing of personal data should be lawful and fair. It should be transparent to natural persons that personal data concerning them are collected, used, consulted or otherwise processed and to what extent the personal data are or will be processed. The principle of transparency requires that any information and communication relating to the processing of those personal data be easily accessible and easy to understand, and that clear and plain language be used. That principle concerns, in particular, information to the data subjects on the identity of the controller and the purposes of the processing and further information to ensure fair and transparent processing in respect of the natural persons concerned and their right to obtain confirmation and communication of personal data concerning them which are being processed. Natural persons should
be made aware of risks, rules, safeguards and rights in relation to the processing of personal data and how to exercise their rights in relation to such processing. In particular, the specific purposes for which personal data are processed should be explicit and legitimate and determined at the time of the collection of the personal data. The personal data should be adequate, relevant and limited to what is necessary for the purposes for which they are processed. This requires, in particular, ensuring that the period for which the personal data are stored is limited to a strict minimum. Personal data should be processed only if the purpose of the processing could not reasonably be fulfilled by other means. In order to ensure that the personal data are not kept longer than necessary, time limits should be established by the controller for erasure or for a periodic review. Every reasonable step should be taken to ensure that personal data which are inaccurate are rectified or deleted. Personal data should be processed in a manner that ensures appropriate security and confidentiality of the personal data, including for preventing unauthorised access to or use of personal data and the equipment used for the processing.

(42) Where processing is based on the data subject’s consent, the controller should be able to demonstrate that the data subject has given consent to the processing operation. In particular in the context of a written declaration on another matter, safeguards should ensure that the data subject is aware of the fact that and the extent to which consent is given. In accordance with Council Directive 93/13/EEC (1) a declaration of consent pre-formulated by the controller should be provided in an intelligible and easily accessible form, using clear and plain language and it should not contain unfair terms. For consent to be informed, the data subject should be aware at least of the identity of the controller and the purposes of the processing for which the personal data are intended. Consent should not be regarded as freely given if the data subject has no genuine or free choice or is unable to refuse or withdraw consent without detriment.

(43) In order to ensure that consent is freely given, consent should not provide a valid legal ground for the processing of personal data in a specific case where there is a clear imbalance between the data subject and the controller, in particular where the controller is a public authority and it is therefore unlikely that consent was freely given in all the circumstances of that specific situation. Consent is presumed not to be freely given if it does not allow separate consent to be given to different personal data processing operations despite it being appropriate in the individual case, or if the performance of a contract, including the provision of a service, is dependent on the consent despite such consent not being necessary for such performance.

(47) The legitimate interests of a controller, including those of a controller to which the personal data may be disclosed, or of a third party, may provide a legal basis for processing, provided that the interests or the fundamental rights and freedoms of the data subject are not overriding, taking into consideration the reasonable expectations of data subjects based on their relationship with the controller. Such legitimate interest could exist for example where there is a relevant and appropriate relationship between the data subject and the controller.
in situations such as where the data subject is a client or in the service of the controller. At any rate the existence of a legitimate interest would need careful assessment including whether a data subject can reasonably expect at the time and in the context of the collection of the personal data that processing for that purpose may take place. The interests and fundamental rights of the data subject could in particular override the interest of the data controller where personal data are processed in circumstances where data subjects do not reasonably expect further processing. Given that it is for the legislator to provide by law for the legal basis for public authorities to process personal data, that legal basis should not apply to the processing by public authorities in the performance of their tasks. The processing of personal data strictly necessary for the purposes of preventing fraud also constitutes a legitimate interest of the data controller concerned. The processing of personal data for direct marketing purposes may be regarded as carried out for a legitimate interest.

(58) The principle of transparency requires that any information addressed to the public or to the data subject be concise, easily accessible and easy to understand, and that clear and plain language and, additionally, where appropriate, visualisation be used. Such information could be provided in electronic form, for example, when addressed to the public, through a website. This is of particular relevance in situations where the proliferation of actors and the technological complexity of practice make it difficult for the data subject to know and understand whether, by whom and for what purpose personal data relating to him or her are being collected, such as in the case of online advertising. Given that children merit specific protection, any information and communication, where processing is addressed to a child, should be in such a clear and plain language that the child can easily understand.

(60) The principles of fair and transparent processing require that the data subject be informed of the existence of the processing operation and its purposes. The controller should provide the data subject with any further information necessary to ensure fair and transparent processing taking into account the specific circumstances and context in which the personal data are processed. Furthermore, the data subject should be informed of the existence of profiling and the consequences of such profiling. Where the personal data are collected from the data subject, the data subject should also be informed whether he or she is obliged to provide the personal data and of the consequences, where he or she does not provide such data. That information may be provided in combination with standardised icons in order to give in an easily visible, intelligible and clearly legible manner, a meaningful overview of the intended processing. Where the icons are presented electronically, they should be machine-readable.

(63) A data subject should have the right of access to personal data which have been collected concerning him or her, and to exercise that right easily and at reasonable intervals, in order to be aware of, and verify, the lawfulness of the processing. This includes the right for data subjects to have access to data concerning their health, for example the data in their medical records containing information such as diagnoses, examination results, assessments by treat-
ing physicians and any treatment or interventions provided. Every data subject should therefore have the right to know and obtain communication in particular with regard to the purposes for which the personal data are processed, where possible the period for which the personal data are processed, the recipients of the personal data, the logic involved in any automatic personal data processing and, at least when based on profiling, the consequences of such processing. Where possible, the controller should be able to provide remote access to a secure system which would provide the data subject with direct access to his or her personal data. That right should not adversely affect the rights or freedoms of others, including trade secrets or intellectual property and in particular the copyright protecting the software. However, the result of those considerations should not be a refusal to provide all information to the data subject. Where the controller processes a large quantity of information concerning the data subject, the controller should be able to request that, before the information is delivered, the data subject specify the information or processing activities to which the request relates.

(100) In order to enhance transparency and compliance with this Regulation, the establishment of certification mechanisms and data protection seals and marks should be encouraged, allowing data subjects to quickly assess the level of data protection of relevant products and services.

CHAPTER II. Principles

Article 7. Conditions for consent

2. If the data subject’s consent is given in the context of a written declaration which also concerns other matters, the request for consent shall be presented in a manner which is clearly distinguishable from the other matters, in an intelligible and easily accessible form, using clear and plain language. Any part of such a declaration which constitutes an infringement of this Regulation shall not be binding.

3. The data subject shall have the right to withdraw his or her consent at any time. The withdrawal of consent shall not affect the lawfulness of processing based on consent before its withdrawal. Prior to giving consent, the data subject shall be informed thereof. It shall be as easy to withdraw as to give consent.

CHAPTER III. Rights of the data subject

Section 1. Transparency and modalities

Article 12. Transparent information, communication and modalities for the exercise of the rights of the data subject

1. The controller shall take appropriate measures to provide any information referred to in Articles 13 and 14 and any communication under Articles 15 to 22 and 34 relating to processing to the data subject in a concise, transparent, intelligible and easily accessible form, using clear and plain language, in particular for any information addressed specifically to a child. The information shall be provided in writing, or by other means, including, where appropriate, by electronic means. When requested by the data subject, the information may be provided orally, provided that the identity of the data subject is proven by other means.
2. The controller shall facilitate the exercise of data subject rights under Articles 15 to 22. In the cases referred to in Article 11(2), the controller shall not refuse to act on the request of the data subject for exercising his or her rights under Articles 15 to 22, unless the controller demonstrates that it is not in a position to identify the data subject.

7. The information to be provided to data subjects pursuant to Articles 13 and 14 may be provided in combination with standardised icons in order to give in an easily visible, intelligible and clearly legible manner a meaningful overview of the intended processing. Where the icons are presented electronically they shall be machine-readable.

Section 2. Information and access to personal data

Article 15. Right of access by the data subject

1. The data subject shall have the right to obtain from the controller confirmation as to whether or not personal data concerning him or her are being processed, and, where that is the case, access to the personal data and the following information: (h) the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.

Section 3. Rectification and erasure

Article 17. Right to erasure ('right to be forgotten')

2. Where the controller has made the personal data public and is obliged pursuant to paragraph 1 to erase the personal data, the controller, taking account of available technology and the cost of implementation, shall take reasonable steps, including technical measures, to inform controllers which are processing the personal data that the data subject has requested the erasure by such controllers of any links to, or copy or replication of, those personal data.

Section 4. Right to object and automated individual decision-making

Article 21. Right to object

4. At the latest at the time of the first communication with the data subject, the right referred to in paragraphs 1 and 2 shall be explicitly brought to the attention of the data subject and shall be presented clearly and separately from any other information.

Article 22. Automated individual decision-making, including profiling

1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.
Expert opinions on making GDPR usable

Johanna Johansen

In preparation to be submitted.
A long version is published as Technical Report nr. 501 in the series with ISSN 0806-3036 managed by the University of Oslo Library.
Expert opinions on making GDPR usable

Johanna Johansen

September 2021

Abstract

We present the results of a study done in order to validate concepts and methods that have been introduced in [11]. We use as respondents in our interviews experts working across fields of relevance to these concepts, including law and data protection/privacy, certifications and standardization, and usability (as studied in the field of Human-Computer Interaction). We study the experts’ opinions about four new concepts, namely: (i) a definition of Usable Privacy, (ii) 30 Usable Privacy Goals identified as excerpts from the GDPR (European General Data Protection Regulation), (iii) a set of 25 corresponding Usable Privacy Criteria together with their multiple measurable sub-criteria, and (iv) the Usable Privacy Cube model, which puts all these together with the EuroPriSe certification criteria, with the purpose of making explicit several aspects of certification processes such as orderings of criteria, interactions between these, different stakeholder perspectives, and context of use/processing.

The expert opinions are varied, example-rich, and forward-looking, which gives an impressive list of open problems where the above four concepts can work as a foundation for further developments. We employed a critical qualitative research, using theory triangulation to analyze the data representing three groups of experts, categorized as ‘certifications’, ‘law’, and ‘usability’, coming both from industry and academia. The results of our analysis show agreement among the experts about the need for evaluations and measuring of usability of privacy in order to allow for exercising data subjects’ rights and to evaluate the degree to which data controllers comply with the data protection principles. However, the community still needs to find archetypal usability thresholds to guide both the businesses in what would be an acceptable level to attain and for the evaluators to assess the level of compliance against.

Regarding the above four concepts our results first show that, while agreeing with the Usable Privacy definition, the experts are more often interested in finding (and giving) instances of this definition, i.e., examples of what are the specific areas of privacy or data protection that usability is most relevant for, or what are the goals and characteristics of the users that have particular implications for privacy. Similarly, the experts are happy with the list of Usable Privacy Goals, seeing these as instances of the Usable Privacy definition. However, even if the Usable Privacy Criteria are seen as a good solution for starting to evaluate usability of privacy, these are not giving enough detail and method for a data controller to be able to understand how (what usability techniques and processes to use) to meet these criteria. Finally, the validation of the Usable Privacy Cube shows that this model captures at an abstract level many of the principles that are present in the existing processes of privacy evaluations. However, these
are present more in an implicit manner, and not as determining factors. Nevertheless, since these concepts are being familiar to the evaluators, and being already part of their practice, further work will be to make these weight more towards considering if an organization has achieved compliance or not. In addition, the movement towards achieving usability goals should also inspire a shift in perspectives from “what is enough to be compliant” towards “what would be the level of data protection that I would like to achieve in order to have a better competitive advantage”.

1 Introduction

Since privacy is a complex concept, with implications not only for individuals, but also for society at large, we have argued in [11, 10] that a multidisciplinary and pragmatic approach to usable privacy is needed. Particularly, [11] addresses the problem that existing data protection regulations, specifically the European General Data Protection Regulation (GDPR), are rather vague when describing the extent to which data protection principles and data subjects rights should be implemented as to be considered beneficial for the user. Therefore, we work towards a methodology that would produce measurable evaluations of the usability with which privacy goals of data protection are reached. Having a scale showing how well a product respects the privacy of its users, and how easy it is for the user to understand the level of privacy protection that a product offers, works towards fulfilling the goal expressed in the Recital (100) of the GDPR, i.e., that of “allowing data subjects to quickly assess the level of data protection of relevant products and services”.

Working towards the above expressed goals, we contribute in [11, 10] with defining, addressing and evaluating usability aspects of data protection.

First, [11] proposes a definition of Usable Privacy which extends and adapts the definition of usability from the ISO 9241-11:2018 [1] to privacy. The rest of that paper shows how to apply this definition to the GDPR.

Usable privacy refers to the extent to which a product or a service protects the privacy of the users in an efficient, effective and satisfactory way by taking into consideration the particular characteristics of the users, goals, tasks, resources, and the technical, physical, social, cultural, and organizational environments in which the product/service is used.

Second, a model called the Usable Privacy Cube (UP Cube) model is proposed in [11] to guide the process of evaluation of usable privacy in certifications. The UP Cube model has three axes of variability; the two at the base contain the existing EuroPriSe (European certification body) criteria, reorganized in two categories, i.e.:
(i) rights of the data subjects, and
(ii) data protection principles.

These also represent the two perspectives on privacy that one usually takes, i.e.:

(i) that of the users of which private information is being collected (and which the regulations usually seek to protect), and

(ii) that of the industry/controllers developing products or services that collect and process private information (and which must conform with regulations such as GDPR and show compliance by going through certifications such as the EuroPriSe).

The third vertical axis is composed of Usable Privacy Criteria intended for measuring the usability level of privacy in a specific context of use. The UP Cube model comes with other concepts that are beneficial for a certification process, such as the idea of ordering and prioritization of the criteria, as well as the possibility to identify intersections between the axes.

Third, [10] lists 30 Usable Privacy Goals (UP Goals)\(^3\) extracted from the GDPR text. One such goal is, e.g., found in the Article 12:

“...any information ...and communication ...relating to processing [to be provided] to the data subject in a concise, transparent, intelligible and easily accessible form, using clear and plain language, ...

How concise, transparent, or intelligible the form of presentation is, can be determined by measurements of efficiency, effectivity, and satisfaction, in a respective context of use. The emphasized words are those that can be interpreted differently based on the context they are used in, and can result in objective and perceived measurements when evaluated using usability methods.

Finally, based on UP Goals, [10] formulates a set of Usable Privacy Criteria (UP Criteria)\(^4\) meant to produce measurable evaluations of usability of privacy that can be translated into scales to be used in certifications. The goal from above, for example, is associated a criteria that contains several specific sub-criteria worded so to produce measurements, e.g., the one below would be requiring to measure efficiency:

How much time/effort/financial and material resources does the data subject need to invest in order to access the information related to the processing of his/her personal data?

1.1 Our Study

We would like through the present study to understand how experts in fields relevant for the above topics are currently dealing with usability aspects of the data protection. Moreover, we are particularly interested whether the solutions proposed in [11, 10] are aligned with their needs and practices. Summed up, the concepts that we aim to validate are:

\(^3\)The UP Goals are presented in short video that was used during the interviews in our current study, and can be found at the link: https://vimeo.com/569510999.

\(^4\)The UP Criteria are presented in short video that was used during the interviews in our current study, and can be found at the link: https://vimeo.com/556133682.
(i) the Usable Privacy Definition,

(ii) the Usable Privacy Criteria,

(iii) the Usable Privacy Goals, and

(iv) the Usable Privacy Cube model.

We have created three groups of participants, described in detail in Section 2.2. Each of these groups are formed around a specific expertise and are dedicated to validate one/two of the above four concepts as follows:

(A) the ‘usability group’ is used to study/validate the Usable Privacy Definition and the Usable Privacy Criteria,

(B) the ‘certifications group’ is used for the Usable Privacy Cube model, and

(C) the ‘law group’ for the Usable Privacy Goals.

We follow the classical way of structuring a qualitative research based study, by presenting how we collected and analysed the data, as well as information about our participants in a Method section (2). We continue with presenting our results in Section 3, where we conclude with Discussions. In the Results section we have a sub-section for each of our main key themes. We conclude in Section 4, where we also mention further work and limitations.

2 Method

To validate the four new concepts listed in Section 1.1 we employ a critical qualitative research [5], where we take an interrogative stance towards the meanings and experiences expressed in the data we collect through interviews. We are also interested in how the individual meanings reflect how usability is understood by the broader communities that the participant’s expertise are representing. To achieve this we involve three different theoretical perspectives in a “theory triangulation” manner [15]: certifications, law, and usability (Human-Computer Interaction/Interaction Design/UX experts). Special in our case is that the participants are not brought to discuss the data, but to discuss the theories and concepts introduced in [11, 10]. Their meanings then represent the data that we analyze in the rest of this paper.

With this approach we seek to validate our knowledge within the scientific and practice community [3], as represented by the experts we brought into the discussion. We aim at bringing forth a “disciplinary matrix” [14] of assumptions, theories, and practices shared on the topic of usability of data protection within the community of specialists on certifications, on data protection, and on usable privacy and security. To validate our knowledge claims, we created through interviews, an environment where a slice of the research and practice community could present their perspectives on the subject [13]. The perspectives are then analyzed in order to identify conflicting or agreeing interpretations, as well as possibilities for future development of the knowledge we claim in [11, 10]. Based on the feedback from the experts we seek to confirm weather our ideas are

Validation is defined, in the Oxford English Dictionary, as the process of confirmation.

4
valuable and useful enough to study, and whether the approach we adopted
to the research questions we address is the right one. We intend to find out
whether our research is both relevant and beneficial to those concerned, such as
the certifications experts. Intending to reach both an ethical and substantive
validation of our work, we have recruited people who have had experience with
the topic, establishing a cooperative approach between the researcher and the
researched in a social constructivism manner, which in our case is the community
working with certifications [3].

2.1 Collecting interview data

We chose to use the method of conducting interviews for collecting our data
because this method is best suited when needing, as in our case, to explore
understandings, perceptions, and constructions (which for us are of topics re-
lated to usability in data protection). We expect to generate rich and detailed
responses because the participants that we chose have a personal stake in these
topics, since they are working with privacy certifications and standards, and
they need to address in one way or another aspects that we talk about in the in-
terviews. Since [11] uses methods and terminology from the field of Ergonomics
of human-system interaction to evaluate usability of data protection, we also
invited experts from this field, especially those that have been working at the
interaction between usability and privacy, e.g., from the field known as usable
privacy and security.

We had meetings of 40 to 60 minutes duration with each of the interviewees.
The form of data collection employed in this study was approved by NSD -
Norwegian Centre for Research Data⁶ and each participant was asked to sign
a consent form prior to the interview. The interview was held online, using
Zoom, and the conversation was audio recorded. The data was transcribed and
anonymized and the recording deleted after transcription for maintaining the
anonymity of our participants.

We designed our interviews to be semi-structured, having a list of questions
to guide the conversation, while the participants were encouraged to talk freely
on the main topics of the interview. The topics and questions were adapted
to the different expertise the participants have, concerning different aspects of
our paper. All three interview types had two main parts. First we want to
learn about the participants’ current understanding of (and their relation with)
the usability aspects of the data protection legislation, without biasing them
with our views. Second, after we introduce our research through a short video
presentation, we than asked the participants to express their opinion directly
in relation with what we have presented. Before the interview we informed
the participants only about our general topic of research, as we did not want
to influence them with our opinions. We also wanted spontaneous, and not
preconceived responses, so to reflect ingrained knowledge of the respective fields
and areas of practice the participants represent.

We started the interview with a common topic for all three groups, where
the participants presented their understanding and experience with usability in
data protection (see Appendix B). This was done with the intention to reveal
the current understanding of usability in the respective domains, whether there

are differences or overlaps between them, and also to see how their perspectives are (or could be) related to our definition of ‘usable privacy’. Afterwords we had specific topics for each of the groups:

• With the ‘certifications group’ we discussed topics related to evaluating and measuring usability, as well as the Usable Privacy Cube model, because these participants have knowledge on which processes and methods are currently used in evaluation and certification processes. For the exact topics and questions see Appendix E.1 and E.2.

• The ‘usability group’ addressed topics related to defining usable privacy and Usable Privacy Criteria, because this group is acquainted with the ISO 9241-11:2018 standard on usability that was used as a basis for the definition of usable privacy in [11]. Moreover, this group knows well methods and processes of evaluating usability of digital products in general, as well as the process of formulating goals into criteria of evaluations. For the exact topics and questions see Appendix C.1 and C.2

• The participants in the ‘law group’, being well acquainted with the GDPR text, were asked to check if the list of the Usable Privacy Goals, appearing in [10], is complete, and whether the goals were correctly chosen to represent usability aspects. For the exact topics and questions see Appendix D.1.

2.2 The participants

The participants were sampled using convenience and snowball methods. The three groups, named ‘certifications group’, ‘law group’ and ‘usability group’, were establish based on the expertise. However, most of the participants have a composite background, a mixture of computer science, law, and human factors. Common for all is that they are working (or doing research) on aspects related to privacy and European data protection applied to IT services/products, thus all having knowledge of data and technology.

Another criterion in our selection process was diversity among the participants, aiming to bring expertise both from industry and academia, as well as age and gender diversity. Even though our initial recruiting of the participants tried to fulfill these criteria, in the interviews we also asked the participants to specify their primary and secondary area of expertise, as well as their work experience (type of position held, type of organization, and number of years) relevant for data protection/privacy (see Appendix A). One conclusion from the interviews, which strengthens one of our initial choices, is that all the participants in our study had some experience with usability related aspects from their work or at least a basic understanding of what usability is. This is apparent from the answers given to the first topic of our interview, which asked the participants to present their understanding and experience with usability. We attempted one interview with a participant that had expertise in data protection, but was not acquainted with the concept of usability at all, which in the end forced us to remove this participant because most of the questions could not be answered.

The ‘certifications group’ consists of six people working with standards, certifications, and data protection organizations. This is confirmed by their answers: 4 out of 6 have this as their main field of expertise whereas the remaining two
are working for DPAs. Moreover, all these participants have Law/data protection as part of their expertise (one as primary and 5 as secondary). The years of experience range from 6 to 32, and the gender is equally represented. The work experience ranges from leadership and research for DPAs, consulting, audit, or technical assessment for certification bodies and other governmental organizations, or board membership and other functions for standardization committees. We consider these backgrounds to represent well our target group.

One of the participants could have been placed, based on the main field of expertise, in the ‘usability group’, but a person with a Usability/HCI/IxD/UX background combined with certification is rare and we wanted to have this participant contribute to the topics especially chosen for the ‘certifications group’. As detailed later, the ‘usability group’ does not include experience with certifications.

The ‘usability group’ contains seven people working with Usability (sometimes also called HCI/IxD/UX), being confirmed also by their answers, i.e., 6 out of 7 have this as their main field of expertise. Their secondary expertise was somewhat more diverse, including law/data protection, privacy and security, cybersecurity, contract design, design thinking, and Information Systems Development from an organisational aspect. The remaining participant had Usability/HCI/IxD/UX as the second field of expertise, with computer security and privacy as main field. The years of experience range from 3 to 28, among 4 female and 3 male. Three of the participants have experience with work in industry as: freelance consulting on privacy as a competitive advantage, CEO & head designer for Legal design consultancy, and member of task group of usable security and privacy. All of the participants have academical positions ranging from PhD to Professor. Even though the academic roles are prevalent, we consider these backgrounds to represent well our target group.

The ‘law group’ consists of four people, three having Law/data protection as their main field of expertise. As the second field of expertise one chose again Law/data protection, another chose Certifications/ISO Standards/Regulations, and the other two chose Usability/HCI/IxD/UX. The fourth participant chose Usability/HCI/IxD/UX as primary field of expertise and Law/data protection as secondary expertise. The years of experience range from 5 to 14, with 3 female and one male. The balance here is skewed towards academic roles (three out of four) ranging from PhD to Professor, with one participant working for a privacy consultancy firm. For this group it was more difficult to find people that had also knowledge of usability, besides privacy and data protection.

2.3 Data analysis

We use thematic analysis (TA) for analysing the data, following [5]. We identify the themes in a “top-down” fashion, where we use data to explore the concepts of interest (i.e., those mentioned in Section 1.1). Since the analysis is guided by existing theoretical concepts, as well by our standpoints, disciplinary knowledge and epistemology, we adopt a theoretical variant of TA. However, we also employ experiential and constructionist variants of TA. For example, a critical and constructionist analysis is used to identify the concepts and ideas that underpin the assumptions and meanings in our data (e.g., we look at how the field of expertise of the participants influences the way they define and understand usability of privacy). We also use TA to develop a detailed descriptive account
of usable privacy and related concepts such as processes and criteria for evaluating usable privacy. At the same time, in an experiential TA fashion, we are interested in the participants’ standpoints towards evaluating and measuring usability, and how they experience and make sense of privacy/data protection aspects that we define as related to usability.

We adopted a researcher-derived approach while performing our coding. When analysing the data, we focused on identifying answers that can be understood as instances that fall (out)inside the concepts that were introduce in [11, 10], i.e.: (i) the Usable Privacy Definition, (ii) the Usable Privacy Cube model, (iii) the Usable Privacy Criteria, and (iv) the Usable Privacy Goals. Besides these four focus areas, we were also interested in validating our general research question, namely that of adding and integrating usability evaluations into the existing certification schemes. The themes have been created based on how meaningful the specific comments of the participants are for the elements we want to validate, how many of the participants have mentioned the specific aspect, but also on how strongly an opinion was articulated and argued for.

3 Results

The results are presented based on the concepts to be validate from the [11, 10]. We start by assessing the need of the certification and standardisation community for evaluating and measuring usability of privacy in Section 3.1. We continue with validating the UP Definition in Section 3.2 directly – by asking the participants in the ‘usability group’ their opinion about the definition we present –, but also indirectly – by asking all participants, irrespective of group, to present their understanding and experience with usability in privacy / data protection. In Section 3.3 we validate the list of our UP Goals with the ‘law group’ by asking them to choose the goals that have most relevance for usability, comment on their choices, as well as suggest other items to add to the list. The UP Criteria are validated in Section 3.4 with the ‘usability group’, where we give additional examples of criteria with their respective sub-criteria, asking the participants to assert weather they represent a good solution for measuring usability of privacy. Finally, as a way of putting all the above concepts together, we present the results from the validation of the UP Cube with the ‘certification group’ in Section 3.5. To conclude, at the end we provide a summary of the overall results and an overview of all themes and their interrelation.

3.1 The need to evaluate and measure usability of privacy

3.1.1 Evaluating usability of privacy

Our general research question – “evaluating and measuring on scales the usability of privacy” – was formulated in interview questions (Appendix E.1) that were addressed specifically to the certifications group, as they are best acquainted with the existing certifications, their needs, and practices. One of the interview questions aimed to elicit whether they find it important to evaluate usability aspects when certifying for compliance with data protection. Their answers are interpreted as we need evaluations of usability of privacy. A more detailed
look at the answers shows some variations. One of the participants did not use the word “important”, and sees the evaluation of usability as something that is “needed”:

“we need evaluations of usability”, “All the GDPR certification programs or schemas need to also look at usability” (CertP1).

Five of the respondents answered using the word “important” with variations such as:

“it’s very very important . . .” (CertP6),

“indeed is important . . .” (CertP3),

“I do think it’s important” (CertP2),

while in one of the answers the “important” label is not given as a general fact, but as connected to the data subjects right:

“it’s important also to focus on usability as one aspect related to data subjects rights” (CertP5).

Moreover, all participants identified several areas where the evaluation of usability is of special importance, or that evaluation should be done “at least” in these instances that they exemplified.

One outstanding example (i.e., mentioned by three out of four participants that specified cases where usability is important) is that “usability is important for exercising data subjects’ rights”. Usable transparency and usable intervisibility are presented by one of the participants as preconditions for the users to exercise their rights. Data Protection Authorities (represented by one of our participants) when evaluating the criteria of certifications schemes for accreditation would

“expect that something also tackling the usability aspects are in place, at least in the case of where transparency is necessary, like how easily understandable is something [pertaining to] Article 13 or 14 [i.e, Information to be provided . . .]. Is it really possible to make a total exercise of your data protection rights, or is it too complicated or not [possible at all]” and that “the end user has to see and understand what is happening” (CertP1).

The transparency principle is also mentioned in relation with giving consent:

“people need to understand what they’re signing or what they are committing to, . . .so that people understand the consent and that consent is not pushed onto them.” (CertP4)

The right to be informed is specifically mentioned:

“particularly when it comes to the part of the data protection law that talks about information to the user . . .and actually make them understandable and comprehensive enough for me to make a sort of informed decision” (CertP6).
Although not as strongly articulated as the above, data protection by default and by design are also mentioned by three respondents in relation with evaluating usability in EuroPriSe. At this point we can conclude that a sub-theme representative for the ‘certifications group’ is that

> evaluating usability is important for data subjects to exercise their rights and for data controllers to comply with the transparency principle.

### 3.1.2 Measuring usability of privacy

The other side of our general research question – “measuring on scales the usability of privacy” – is important for making evaluations of usability of privacy more objective and easy to follow by both the companies wanting to be certified and by the certification organizations. During the interview the respondents were asked whether they see as useful to concretely measure and evaluate how well the usability of privacy is dealt with by companies wishing to be GDPR compliant. We also explained to each participant that by measurements we meant some form of scale or score of the type used to indicate energy consumption for house appliances. The sub-theme that would be representative for the answers at this question is that Measuring is definitely useful but where do we start? That measurements are something desirable was very clearly and strongly stated by all participants. CertP1 even generalized the statement to the whole community:

> “Yeah, I think measurement is a good thing. It is something everybody or those who are in the community agree on.”

That the community is favorable towards scales based measurements, such as the use of traffic lights, is also exemplified through research work such as [4, 18] or by the work done on privacy icons [9, 6].

Even though the respondents were in favor of measuring privacy, they all brought up several challenges, and this without us encouraging them to do so. The concerns being raised are exemplary for indicating where the community is at the moment in terms of measuring (usability of) privacy and what are the challenges, and some possible solutions, that the community sees. Moreover, these answers offer confirmation for the choices taken in [11].

One of the more general challenges that one will be confronted with when measuring usability aspects of privacy comes from the fact that in privacy we do not deal with “stabilized knowledge” (according to one of the respondents). There are examples of actors such as the Stiftung Warentest\(^7\) who compares products/services based on aspects such as usefulness, functionality, or environmental impact, and that are using a scoring system based on percentages (example contributed by one of the respondents). However, usability of privacy is not as easy to measure as, for example, the “consistency for the shampoo” (CertP1). We have investigated this aspect in [10].

One conclusion from several of the participants is that we are still in a rather initial phase regarding measuring usability of privacy, where one still asks basic questions such as:

\(^7\)Stiftung Warentest is a German consumer organization and foundation involved in investigating and comparing goods and services. See [https://www.test.de](https://www.test.de)/ .
“how do you measure it and what do you measure” (CertP2)

or wheather

“is it now a score, is it value, is number ...or is it only one thing, we need that or not? ...is there something where we rather need, ‘This is a no go’?” (CertP1).

and we need to

“Come up with a set of objectively measurable aspects of what usability is and what we wish for”. (CertP6)

In our work [11, 10] we offer a definition of what usable privacy is and we suggest which aspects to evaluate and how to produce measurable results that can be translated later in a number or a score. We do not however answer what will be a “no go”, i.e., what could constitute a minimum level that needs to be reached in order to pass the evaluation. Moreover, we propose that usability evaluations should be coming on top of (just) GDPR compliance, with the goal of encouraging companies to improve the usability of privacy of their products (which is being supported also by one of the respondents),

“as some type of competition and advantage over others” (CertP5).

However, when promoting the ‘competitive advantage’ as a motivation, one has to think of aspects such as

“who’s going to pay for the process and it is actually very little happening because of people who can do it, in the sense of the consumer, of the citizen, who usually don’t have the resources.” (CertP4)

One continual source of discussion is deciding what can be considered as good enough usability in order to be compliant with data protection regulations, which in turn would define what it means to go beyond, and hence obtain a competitive advantage:

“at some point you would have to discuss if it’s still an obligation, a legal obligation, or if maybe some companies go beyond what is required by law” (CertP5).

In our work [10] we identify usable privacy goals that appear in the GDPR text, which means that we identify legal requirements for usability of privacy. CertP1 argues that

“as the data protection authority we also have to decide whether it’s sufficient, which is proportionate to whether it’s OK or not, and sometimes even with the bad usability, it’s still okay. It should be better. It could be better, but if we find a system and they only want to know whether they are legally compliant ...of course changes cost money ...is it okay or not because we have to decide whether we need to stop data processing because it’s unfair, whether we can sanction it or whether something has to be changed. ...in the end it is about being sufficiently compliant or not? This is where the court has to stop.”
At the same time CertP2 argues that usability is

“an element of success, and it’s also an element of failure, so it could lead to the failure of compliance. . . . If you make a system that is not usable by the user, then you cannot really have compliance. In the end, it’s going to be too difficult, too boring, too bureaucratic, too complicated to do something, and the system will fail.”

The conclusion that we can derive from these discussions is that one has to find what will constitute enough usability to be deemed compliant and what comes in addition as a competitive advantage.

Other aspects that were brought up by the participants and that resonate with our work [11, 10] are the context of use and the target group:

“what do you measure in what kind of context and who is the target group?”(CertP3);

“Who is the target group of this certification result? Is it the consumer themselves? Or is it maybe an intermediary professional like somebody from a consumer protection organization?”(CertP4);

“we need to know more about the target groups, who is, well, those who have to understand it all” (CertP1).

In our Usable Privacy Cube model [11] and the Usable Privacy Criteria [10] we account for the specific context of use, as well as the users with their goals and specific environments. The participants also give examples of where the specific context (including the users) is important for determining weather the usability is sufficiently addressed:

“And a very simple example coming to my mind is the number of clicks you need to do something or to get an information . . . because sometimes you can get to control or information in one click, but it doesn’t mean that the information is good . . . And sometimes if you have a proper layered approach, maybe you need three clicks, but then the user would be better informed or have better control” (CertP3)

Moreover, an important aspect raised is that different user groups would have different needs regarding the presentation of the evaluation results:

“you have certifications or certification results and they are to be looked at primarily by experts or semi-experts like the consultant or the adviser in a Consumer Protection Organisation. I mean they have some knowledge and they can be able to translate things to their respective target groups.” (CertP4),

which is a topic that we cover in depth in the paper [12]

3.2 Defining Usable Privacy

In [11, 10] we have suggested to adapt the definition of usability from the ISO standard 9241-11:2018 to privacy, which we validate here primarily with experts from HCI/IxD/UX community, as these are supposedly more acquainted
with this ISO standard. During the interview we presented the definition and explained how it is relevant for GDPR, after which the respondents from the usability group had to answer whether this definition captures their own (or their community’s) current understanding of usable privacy. The multiple-choice answers (i.e., ‘completely’, ‘partially’, ‘not at all’) were followed by an explanation of their choice (Appendix C.1). In addition to asking directly the usability experts to validate our definition, the participants in all three groups have been asked to explain their understanding of usability in the context of data protection, and to also anchor it in the reality of their practice (Appendix B). In order to gather unswayed perspectives, these questions were asked in the beginning of the interview, before presenting our definition; in the rest of the paper, we refer to these as ‘unswayed perspectives on usable privacy’.

From the answers related to our definition of usable privacy we can distill the following theme where all fit: We trust the usability definition from the ISO standard 9241-11:2018. The majority of the participants considered the definition as complete, and as an adaptation of the usability definition from the ISO standard, e.g.:

“...completely because it was almost the same definition of usability [as the ISO standard 9241-11:2018] but trying to have instead of general systems, systems that protect users privacy.” (UsabilityP1),

“I would say that it is a complete coverage of the different concepts that one could expect within the usable privacy domain because I think indeed there is quite a resemblance to the definition that comes from the ISO standard.” (UsabilityP2)

Moreover, besides agreeing with the definition itself, one of the participants also appreciated our exemplification of how the definition applies to GDPR.

“This was a more marvelous thing to see how well you related to the GDPR and to the ISO standard. I think that was wonderful.” (UsabilityP7)

We can conclude that all participants agreed that adapting the definition of the ISO 9241-11:2018 to privacy, as it was proposed by us, captures (the choice ‘completely’ being used by the majority, while the remaining chose the alternative ‘partially’) the current understanding of usable privacy in their field (i.e., in usable privacy and security, HCI, UX).

Further confirm of this conclusion (i.e., that it is good to base our definition of usable privacy on the ISO standard 9241-11:2018) is provided by two of the answers which are of the type ‘unswayed perspectives on usable privacy’, which mention the definition of usability from the ISO standard when they are asked to present their own understanding of usable privacy:

“I can refer to standard definition of usability in this context that they can take control over the data in an efficient way, while they are satisfied and also taking the effectiveness into consideration.” (UsabilityP1)

and
“So I define usability for me and for my students using this ISO definition of usability, where you have effectiveness, efficiency and satisfaction . . .” (UsabilityP4)

Moreover, these answers provide a good indication that our decision to test the usable privacy definition with the usability group was appropriate, as they are more acquainted with the ISO standard on usability.

For the respondents that checked the 'partially' choice, we can group their answers under the theme Instances of the usable privacy definition, as these are more specific cases or occurrences of the aspects that are represented at a higher level by the definition. For example, our definition mentions that the goals of the users should be considered, whereas UsabilityP4 says that there should also be a distinction between secondary and primary goals of the users:

“Partially because I didn’t see this distinguishing in primary and secondary goals. I think it is important. . . . one could also say that it’s some how included, but I think it’s better to mention it separately.”

When validating our definition with the usability experts we have found more such ‘instances’. One of these is privacy is a secondary goal to the user, which can be mapped to “taking into consideration the particular [goals] of the user” in our definition. This aspect was mentioned by the UsabilityP4 both when asked to evaluate our definition – as cited above – but also in the beginning of the interview, in the ‘unswayed perspectives on usable privacy’:

“I consider the three things: effectiveness, efficiency and satisfaction. They should be fulfilled for primary goals and for secondary goal which is in this case data protection. So for example, if some data protection mechanisms interferes with primary goals of the users,...too much is of course not good at all, not even if it’s interferes slightly, or more like middle-ish, I evaluate the things: OK, this can’t be usable.”

Privacy goals as interfering with the primary goals of the users was also mentioned by other participants:

“People would be able to do so [understand the privacy policies if they are written in a non legal way], but in practice they don’t [read] because it just doesn’t work with their lives and it doesn’t match the current goal of just signing up for the service and using it.”. (UsabilityP5)

“...there are conflicts between giving people enough time to think...[to make an informed decision], but they might reasonably have the opinion that they should be able to use something as quickly as possible.” (UsabilityP6)

“it’s a secondary task, and that means it should be as little. . . . the workload. . . . should minimise the workload as much as possible.” (UsabilityP7)
Since this topic was brought up by the participants in the usability group only, we mark it as a concern specific to this group. The significance of this topic is also acknowledged by the literature in the field of usable privacy and security, as for example in [19, 2].

In the comment of participant UsabilityP5 above we identify another ‘instance’ of our definition, which we code as *supporting the correct behaviour and characteristics of the users*, which can be mapped to the “Taking into consideration the particular characteristics of the users . . .” part of our definition. UsabilityP5 argues even more about this point:

“I think it is should not only be about the ability to access information because like you are able to access privacy policies, you’re totally able to do that, but nobody does...This is like the one of the most traditional things where HCI says yes, okay, formally you provided the information, but practically you didn’t reach the goal.”

Other two participants from the ‘usability group’ argument more lengthy on this topic:

“you need to understand the behavioural aspects of how people interact with information. Because on one hand, because of information overload, people are [?], so they’re using heuristics and proxies like . . . trust, or how famous an organisation is and they just hope that those organisations are doing something legal and reasonable. But people don’t really have neither the time nor the willingness to expend effort, many of them, to read what are their rights and what organisations are doing with their data?

To actually do disclosures that work for the users, I think that probably usability is also just wider than interacting with the with information, because there is still the problem. I believe that even if everyone, every organisation in the world would design these disclosures in a way that it’s understandable and user readable, still, we would need to read and interact with every single one of them, so people would be quite fatigued. . . . So are there standards or shorthand so we can guarantee data protection without people having to put any effort?” (UsabilityP3);

“And people don’t necessarily want to understand – some do – but a lot don’t really want to understand all the details. They want to know what they have to do and and what they shouldn’t do in order to do this correctly.

...what you see is really just responsibility being pushed down the chain, and basically exposing people to actually the legal text and various things . . . It’s really what I call responsibilization and not really actively enabling and supporting the correct behaviour. even if I do not agree, I still have to use it” (UsabilityP7)

While two of the ‘law group’ participants mention it as a fact:

“Nobody really wants or is able to give an informed consent for every single banner in every single website and application that he or she visits every day.” (LawP1)
“I know that people don’t read privacy policies . . .” (LawP3).

The literature on usable privacy and security covers well this topic, e.g., speaking of a privacy gap [8, 17] between what the user says that would do when asked or tested in the laboratory and what it actually does when in a concrete, real situation.

Another ‘instance’ that we have identified we call usability for transparency and data protection rights. This theme includes two elements that appear in the answers of the participant usually as ‘transparency’ and as different aspects of data protection rights such as ‘having control’, ‘self-determination’, ‘intervenability’, ‘having choices’, which are easily correlated. The certification group, in particular, is preoccupied with both aspects.

“end user has to see and understand what is happening and we distinguish also between having only the transparency and while having to agree, or also taking steps for your own control. The self determination, then the possibility to intervene.” (CertP1)

“how easily can they find information on the processing of their data, how easily can they control this and control what kind of data are processed for what kind of purpose, and also how easily they can exercise their rights?” (CertP2)

“I would say that the primary goal when it comes to data protection and usability is to create transparency, so that I as a user easily know what I’m doing and what I’m sending and what these things are used for. So that I actually have knowledge of the data I’m sharing with others and why.” (CertP6)

The transparency principle and data subjects rights, presented intertwined by the ‘certification group’ above, have been identified also in Section 3.1.1 with respect to evaluating usability of privacy. The ‘law group’ puts more emphasis on the transparency part, in the sense that they bring up challenges, but also solutions, related to how to translate legislative and even technical jargon so that they are more accessible (i.e., as in easy to find, quick to read, and easy to understand) to the regular users.

“And the difficulty is in really making the translation.” (LawP2)

“... almost everything that is in data protection. You know when you try to explain stuff there is a lot of technical jargon,” (LawP1)

In addition, the ‘law group’ is committed to suggesting or finding solutions, such as design patterns or privacy enhancing tools, to translate the transparency requirements into usable solutions for the user.

“Maybe the way to design concept experiences in which – [information about] how you protected data, how you process the data – can actually be designed in a way that makes sense for the user, and they can take more informed decisions, maybe through, for example design nudges, design friction.” (LawP1)
“I’m discussing ways that we can improve the way that policies are communicated. What I’m trying to do is to implement these policies through the design or through the interface of the website. So, for example, transparency enhancing mechanisms and or privacy enhancing mechanisms.” (LawP3)

Such solutions are being proposed also by participants from the ‘usability group’:

“we focus on the usable transparency in the consent forms and also the user satisfaction with the affirmative actions and the time that affirmative actions can take from users to handle content forms, and if these affirmative actions, for example can help users to pay more attention, to, for example, make it more accurate for them to remember to do exactly what they gave their consent afterwards.” (UsabilityP1)

“design patterns or some design frictions to help the users to actually get interested in reading” (UsabilityP3)

Another ‘instance’ that was mentioned by many participants, from different groups, is related to ‘dark patterns’ and practices where usability is used against the users. We have coded this instance as usability turned against users, which can be mapped to our whole definition, albeit negated, as follows: “Dark usable privacy refers to the extent to which a product or a service deceives the users to give up their privacy in an efficient, effective, and illusory/seemingly satisfactory way, by taking into consideration . . .”.

“There’s actually the opposite of usability or anyway usability being used for dark patterns. So design expertise being willfully used by big ones. So organisation that definitely know what good UX is like Facebook or Google that are thwarting the choices of users making it more difficult to change something, or to say no, you cannot track me and so on. So actually there is quite sophisticated work, but done against consumers.” (UsabilityP3)

The above ‘instances’ follow a pattern that we already have used when doing the work presented in [11, 10], i.e., that of finding out what are the usability aspects that need to be addressed and then find solutions for dealing with them. Our approach to identifying ‘instances’ of usable privacy was to go through the GDPR text and mark all the cases that could be considered vague and interpretable, and if evaluated in a usability tests could result in different levels of achievement of the usability goals: effectiveness, efficiency, and satisfaction. In the paper [11, 10] we identified most occurrences in the provisions concerning consent, then on the second and third place, provisions regarding the transparency principle, and respectively the rights of the data subjects. We can see here an overlap between the ‘instances’ we identified on the second and third polce in our paper and the ones that we surfaced in the analysis done in this section, called usability for transparency and data protection rights.

3.3 A comprehensive list of Usable Privacy Goals

In the previous section we identified aspects of privacy and data protection for which usability is considered relevant by our participants, with differences
between the three groups becoming apparent. What we called Usable Privacy (UP) Goals in [11, 10] can also be considered ‘instances’ of the usable privacy definition, in the same sense as above. Here we validate our UP goals with the ‘law group’, since this is well acquainted with the GDPR text, from which we have extracted our UP Goals.

The participants in the ‘law group’ were given a list with all our UP Goals (see Appendix D.1) and were asked to choose the ones that they thought relate to usability. We then discussed their choices and opinion about this list, whether they thought it was exhaustive, and whether they could provide additional goals.

Counting the numbers of goals from the list checked by the participants, the mean is 21.75 choices out of 28, i.e., a 77.67% coverage. Thus, the participants generally agree with our UP Goals, where particularly LawP1 checked all the goals, whereas LawP3 and LawP4 expressed directly their satisfaction with how well the list covers usability aspects:

“...your list was very complete. I cannot think of something that is not on this list. ...I think this list here is very broad and very comprehensive regarding usability. I cannot think of anything else.” (LawP3)

“I am happy with this list. I’m extremely happy with this list.” (LawP4)

However, the same two participants also said that if they would go thorough the GDPR text they will probably find more such goals, but they could not name any such goals in the interview.

“I think if I had more time to think and analyse and probably go back to GDPR I would find other usability goals.” (LawP3)

“I don’t think it’s exhaustive, but I would need to scroll all the GDPR.” (LawP4)

Therefore, from these comments and the large number of UP Goals selected (i.e., confirmed) we can derive the following theme: I am happy with the list of Usable Privacy Goals.

Analysing closer the not-checked UP Goals reveals uncertain UP goals since a direct relation to usability is not obvious. The UP Goal “Make the natural persons aware of risks, rules, safeguards and rights in relation to the processing of personal data. [Recital (39) of GDPR]” was not checked by LawP2 and LawP3. While having no explanation from LawP2, we can see that LawP3 associates this with other aspects than usability:

“it’s more related to language, and I tend to not associate so much the language with usability.”

Even LawP1 shows initial hesitation towards this goal, but agrees with it in the end:

“I’m not sure about how awareness goes hand in hand with usability. It’s an important ingredient, is an important aspect but I’m not sure whether one determines the other. ...But awareness of risks it’s different. I think it’s important. Awareness of the extent to which consent is given; I know it’s a usability issue.”
However, opposed to these three participants, LawP4 is strong on including this goal, together with “neutrality of privacy choices” and even arguments with literature references [7].

“I mentioned in the beginning like the user awareness of the risks, of the of the consequences, but I see here Recital (39) being quote, being here. So I’m very happy with this. … The neutrality aspect it’s not here. I think neutrality is very important. It’s also I think very related to usability. Actually, it’s one of the elements of the design of privacy choices, of the method that I mentioned from [Feng and Yao].”

Other goals that need further scrutiny, being left out by three of the participants are:

(UPG.11) “Personal data should be processed only if the purpose of the processing could not reasonably be fulfilled by other means. [Recital (39) of GDPR]” and

(UPG.30) “The data subject should have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her. [Article 22 (1) of GDPR]”

Despite not being given reasons for why the three participants did not choose the UPG.11, the UPG.30 is checked and argued for by the LawP1 with the following comment:

“And also the last one [number 28 in the list], I think there should be a very easy actionable choice to not be subject to decisions based on automated processing.”

LawP2 gives a more detailed account of the choices made in the list, where particularly interesting is mentioning ‘control’ as the more general area, while goals related to consent and data subject rights are seen as more specific. Since in [11, 10] we decide the level of achievement for a “generic” goals based on measurements coming from more specific criteria, it is also interesting to observe how the participant correlates consent and data subject rights to control:

“…more in general I have selected a number of sentences which relate to the control aspects . . . , but also the specific informed indication of your agreement. So about consent.”

and about how consent relates to data subject rights:

“having control of your personal data, … that’s really about being able to decide for yourself which data you’re sharing, having insights in these data. This also relates to the exercise of data subject rights, having access to your data – obtaining this access – in an easy way, being able to check whether things are correct indeed, and also having the option to object.”

Indicating such general usability related areas matches the approach we adopted in [11, 10] where the UP Goals and UP Criteria (see also Section 3.4)
are categorized based on their area of application from the GDPR text. The areas listed below are ranged based on the number of goals found in each category, the first item having the most occurrences.

1. Consent (lawful grounds for processing data principle),
2. Information and communication addressed to the public or to the data subject (transparency principle),
3. Rights of the data subjects (rights in general),
4. Purpose of processing, and
5. Legitimate interest of either the processor or the data subject (lawful grounds for processing data principle).

In addition to areas related to preexisting GDPR chapters, we also have a “generic” or higher level category, where we have placed UPG.2: “Natural persons should have control of their own personal data. [Recital (7) of GDPR]”.

“Having control” is the aspect that is valued highest by LawP2 in respect to usability. Although control is only mentioned in Recital (7) of GDPR, and mostly relates to consent

“It's difficult to get it from the GDPR, except for the Recital (7) I was mentioning. ... I think most control and most usability is related to situations where you are indeed basing the processing activities on informed consent. But then it should not only be about being informed, having a proper choice and really knowing what you're consenting to.” (LawP2)

LawP2 sees having control on which data one is sharing as the next level in usability of privacy:

“But the next step for me would really be to have more control and more fine grained control in which data then you are sharing, for which purposes and making more fine grained selections of how the data are used by a data controller. At least it would really be the next level on usability of privacy.”

### 3.4 Ways to meet the Usable Privacy Criteria

Having established the list of usability goals that GDPR stipulates, the practice in the Interaction Design field is to operationalize these by turning them into usability criteria formulated as questions [16]. Criteria can be seen as specific objectives to be reached by those that aim to reach the set of goals that the criteria relate to. In our case, the Usable Privacy Criteria enable one to assess the privacy related measures that a product or system provides in terms of how much these improve the control that the data subjects have over their data. Examples of commonly used usability criteria (i.e., not specific to only privacy) are:

(i) time used to complete a task (efficiency), such as reading a privacy statement, or
(ii) the number of errors made when carrying out a given task (effectiveness), such as choosing the desired privacy settings.

Usability criteria can provide quantitative indicators of the extent to which, for example, the data subjects understand the implications for her/his privacy from using a certain technology. In addition, given that certification bodies already use various forms of criteria in their evaluations, we too want to provide the Usable Privacy Criteria as a set of rules for assessing also usability, so to enable the certification bodies to integrate evaluations of usability within their existing certification schemes.

The UP Criteria are validated in this study with the ‘usability group’, as they are most acquainted with the process of formulating criteria to meet goals such as efficiency, effectiveness, and satisfaction.

The participants were given examples of the UP Criteria and were asked to comment on them (see Annex C.2). The UP Criteria have been assessed as good by most participants, using quick and simple statements, such as:

“I think they are a good solution for evaluating usability.” (UsabilityP1)
“I definitely see the reasoning behind it and it makes sense for me.” (UsabilityP2)
“The questions are good to me.” (UsabilityP3)
“I think it’s good that you have these measurement tools or dimensions so that you can start to negotiate how you should measure all these individual questions.” (UsabilityP5)
“Yeah, I think the criteria are good.” (UsabilityP6)

However, the participants were keen on the discussion to quickly turn towards another related topic that seems to be preoccupying the community at the moment, that of establishing standards, recommendations, and creating guidelines or design patterns, to help with meeting such criteria, e.g.:

“probably giving some sort of rubrics or somehow like figuring out what are the things that the organisation can possibly do, for example to give control to people of their personal data. You design a rubric with scenarios like is there a dashboard? Yes. Is the dashboard clear? Yes. Are all the buttons equal and there’s no dark patterns? And then you start scoring on a more grained way like that. Perhaps you can help also non-designers to evaluate these things or you can help the actors that would like to do it, but they don’t have a clue how to do it.” (UsabilityP3)

“To measure understanding of the consent text, you should do this. Whatever it is. For example, you should have to understand the questions about this, or you should have to make user interviews and see what users think when they read this text. Yeah, so something like concrete. What concretely to do in order to understand if the user understands?” (UsabilityP4)

“there is a Article 29 Working Party opinion on purpose limitation and implementation of the principle of purpose limitation and there they say a little bit more, or at least this could be a guideline for implementing information and unambiguous consent.” (UsabilityP5)
“your questions will really be questions to start with: How long, how much, et cetera? And then you start to see the patterns that are leading to good informed consent and then you say that, well, we take these as our standards now. . . .I would say that within a branch of industry or branch of services, you could say that after a while you will get established. You know this is sufficient for this kind of service. This kind of data request has become the standard. I think that the Data Protection Board or Consumer Agency are doing such evaluations.” (UsabilityP6)

The general tendency to transform the usability aspects into very concrete privacy practice is reflected also in the law domain, as directly expressed by one of our participants from the ‘law group’.

“I see a lot of development of usability aspects / usability concepts, how some principles of usability, if I can say so, are decomposing to very concrete privacy practises for data controllers and for users, so indeed there is a prolongation, there is some decomposition, some instantiation of it and not only at the hard law, but also on soft low level. . . .encompassing usability aspects mostly at the guideline level. So you will not see examples of icons in the GDPR, but you will see a lot in DPA guidelines. References to dashboards, or what does it mean intelligibility or clear and plain language you will see it in the guidelines, you will see it decomposed in case law. . . .there is some very important decision from that CNIL versus Google in 2019 that talks about ergonomics of information not only informed consent or informed choices, but how the information should be delivered.” (LawP4)

That such more concrete guidance is needed is confirmed by the participants in the ‘certifications group’ as well. Their assent is especially valuable as they are the ones that are actually performing the evaluation in practice.

“but it would need, for me, a very specific grading system that an auditor could check the effectiveness and the efficiency in the use of usability [against] because these terms are very experienced based, they’re very subjective, right? It’s not like if you’re asking them: Is the colour blue? You’re asking if it’s usable and they have to compare it to something and say it is usable. So unless you provide a very concrete scheme behind this, I don’t know how if it’s going to work.” (CertP2)

The focus of the participants was on the particularities of the evaluation, addressing questions such as who would perform the evaluation, what kind of expertise the evaluators would need to have, or which specific HCI methods should they use.

“What type of the methods should I use to evaluate this and then what type of users, the number of the users . . .” (UsabilityP1)

“So this kind of guidance that you have here would be very useful for somebody who is professional in both usability and privacy . . . I
could imagine that people who are experts in only one, or people who
are experts in law would need a lot of guidance . . .” (UsabilityP4)

Since the UP Criteria functioned more as a trigger for discussing other more
particular aspects of the evaluation process, a theme that would characterize
best the type of feedback we received from the participants is Ways to meet the
Usable Privacy Criteria.

One of the participants from the ‘certifications group’ indicates nicely where
we are at the moment with this endeavor.

“There’s some guidance right now, [and] I think it will be refined
more and more. Still it won’t be on the same level as: Your sen-
tences can only be half this length, or something which can also be
automated. On the other hand, I think we may have design pat-
terns which are okay, [as for example] some grouping like you use for
content management, and this [will be] approved by data protection
authorities or by court, and then perhaps 20 years later people think:
Let’s rephrase it a bit. Now it’s even better. . . . On the other hand,
we are not there yet. If you see the cookie banners there’s so many
different ways to give consent or [to present the information for the
data subjects] to understand what is essential.” (UsabilityP1)

3.5 Usable Privacy Model – an abstract representation
of known and implied principles of privacy / data
protection evaluations

Once having proposed a set of criteria, we are further interested in how these
would be integrated in the processes and with the criteria of evaluation of the
of EuroPriSe, and devises a higher level model, called the Usable Privacy Cube
(or UP Cube), that combines the evaluation criteria of EuroPriSe with the
Usable Privacy Criteria. Building the Usable Privacy Cube model on top of
the EuroPriSe criteria gives a good anchoring in this well established privacy
certification scheme, but even more, it provides a general reorganization of the
EuroPriSe criteria into rights and principles as found in GDPR, which is meant
to work as a guide for adding usability evaluations to other data protection
certifications that target GDPR.

In this study we validate whether the UP Cube model reflects the existing
privacy and data protection evaluation processes, and to what extend (i.e., to-
tally or partially). Specifically, we discuss with the participants the following
features of the UP Cube model:

(i) representing the perspectives of both data subjects and controllers/processors;
(ii) grouping, prioritization, and organization of the criteria;
(iii) interactions between the different criteria; and
(iv) context of use (or context of processing, as a term often used in GDPR).

To our question “Does the UP Cube model represent, at a high level, the
existing data protection and privacy evaluation processes?” two out of the five
participant chose “Completely”, while three chose “Partially”.

23
The answer of the CertP1 is exemplary:

“What I know best is EuroPriSe and the previous data protection seals from ULD [Landeszentrum für Datenschutz Schleswig-Holstein], so it’s quite very much related, but I think not completely. So I would say partially, although on the abstract level will be the same as the Standard Data Protection model that also uses the different axes for something like that. So the general principle I think is quite well known ...”

The Standard Data Protection model 9 has the notion of allocating the legal requirements of the German Federal Data Protection Act – BDGS (Data minimisation, Availability, Integrity, Confidentiality, Unlinkability, Transparency, Intervenability) to the protection goals, in a tabular manner. A cube (like our model) can be understood as a three-dimensional tabulation mechanism, i.e., represents three tables, each based on the combination of two of the axes. Therefore, the model mentioned by the respondent can replace the EuroPriSe in the base square of our UP Cube, and it is already fitting, to some extent, withing our two axes of organization in rights and principles. This example also confirms the versatility of the UP Cube model.

Moreover, CertP2 confirms the novelty and usefulness of the UP Cube model as it makes clear certification aspects that are not explicitly expressed anywhere, but more implied:

“I would say that it’s something that’s not actually ever written down, that’s more like implied. The auditors correlate these things when they’re doing the audit, but there’s nothing written down that says that they should do it like this. Yeah, so I would say partially, very abstract, actually.”

What is more interesting is that, even though CertP5 and CertP6 chose “Completely”, their answers are on the same line with the first two participants, in the sense that the ideas behind the UP Cube model are known, and that it represents these at an abstract level (as a model is meant to do). Therefore, the theme that we extract from these answers is that “UP Cube is an abstract representation of known, but implied or covert practices”.

“Yes, probably your model is representing everything at a more abstract level because it’s probably covering all four aspects. So I could say completely, but sometimes not very specific.” (CertP4)

As in the case of the Usable Privacy Definition presented in Section 3.2, we had questions preceding the presentation of our model (these are called the “unswayed perspectives”), asking if the participants know whether the certifications or standards that they are acquainted with have a high-level model to guide the process of evaluation. The conclusion from these answers is that it does not exists a published or well established model to guide the process of the evaluation, but there are some main guiding pillars. These are following the GDPR text, or in the case of the standards for evaluating management systems, the risk management or the Privacy Impact Analysis (PIA) is the focal point.

8https://www.datenschutzzentrum.de/guetesiegel/
“For the time being the guidance for certification against 27701\textsuperscript{10} is very generic. So you’re not going to find something that would be concrete for the evaluation process, so it’s just the they give you (?) with the requirements. They tell you that the organization should do these things and then it is asked to evaluate based on understanding and experience whether they achieve the goals of its control. They do not provide any more guidance than that. So they tell you, they [the organizations] should have a record of processing facilities, for example, this record could be in Excel, it could be a word file, it could be some paper that they write down. It could be anything as long as they have a record of processing facilities; from an audit point of view this is okay?” (CertP2)

“So what I know about the former and also the present certification schemes is that they are driven by the legal text and driven by additional criteria, perhaps from the Article 29 Working Group …” (CertP1)

“another tool that isn’t GDPR is the privacy impact assessment. This is a way to evaluate processing in regard of the risks that can happen or that are linked to it.” (CertP3)

“The ones I primarily know of are management system standards and … they are also risk based, which means that you choose measures based on your perceived risk.” (CertP6)

When discussing with the participants the individual components of our model, we obtain further confirmation of our previous observation that the elements of our model are present, to a certain degree, in the existing processes. Participants confirm that even elements that we deem indispensable for a usability evaluation, such as looking at the context of use or the different stakeholders, are also being considered to a certain extent in some of the certification schemes.

We next give an overview of the level of the consideration given at the moment to aspects related to usability evaluation in privacy certification. This overview can indicate the premises one would have to start from when considering evaluating to what extent products or services meet the usability goals of GDPR, identified by us in \cite{10}.

The idea that the UP Cube model should capture both the perspective of the data subjects and of the controllers/processor, resonates well with our participants.

“It is smart to have these different axes so that you can see at one glance how things are related to each other. We have that in our criteria catalogue. For example, we would have the legal obligations of, let’s say the controller regarding information duties and on the other hand we would also go to the other perspective of the data subject and we would have another requirement [related to the] right to be informed. So we have these connections also. But they are


25
We generally see in the discourse of our participants, irrespective of group – when discussing about usability in general or other topics not related to the UP Cube model – that they are preoccupied with both perspectives. In the case of the controllers, usability is considered as the ease with which they can implement data protection related requirements and become compliant.

So usability is always a battle in IT, we have to make sure, especially in IT security and of course in privacy, that whatever we create can be used by the people that are supposed to implement the specific activities without creating too much of a burden. Especially the early systems that we saw in compliance to either information security or data protection they used to be very paper based and they were very bureaucratic and they created a burden to anybody that had to perform these actions and especially the parts where they had to ask for specific permissions or consent, or they had to monitor the information that they were using through the record of processing activities. So for me, usability is to achieve your goals regarding information security and privacy, data privacy, while also allowing another person to perform their tasks seemingly unaffected by the fact that they are being compliant.” (UsabilityP5)

“I’m thinking about end users and how data privacy might be made more usable and a better user experience for them. But I’m also thinking about what practitioners and researchers have to do to deal with data protection issues, myself included, …we have been running a lot of projects internally in which we gather personal data, even sensitive data. So we have to navigate the complexities of data protection compliance for example, which is really unusable. It’s not a very good user experience for anybody as researcher, especially because you have a lot of obligations upon you, not only in better data protection but also in terms of ethical issues, in terms of how you manage data in general, not only personal data, but how do you protect and manage the data that you produce through the research. Of course is also a little bit broader, but then a core part of it is the personal data that you manage and it’s very very very complex, …for me it’s also research question: How can we make data protection compliance usable?” (UsabilityP2)

In practice, the rights and obligations of the parts involved need to be balanced, and the balance is achieved through a process of negotiation, where considering the specific context of data processing is seen as indispensable.

“what we also see is that very often there are some rights of the different parties, and …one is always already in a better power position. So usually the data controller. We would assume they are
From the examples above we see that it is necessary to consider the specificities of different “situations” being evaluated, in order to achieve the optimal power balance and protect the rights and interests of all parties involved. Similarly, we see that in the case of all existing evaluations the context of use/processing is considered in relation to, for example, the risk analysis or the type of data being processed, but this is not done in an explicit way as a usability evaluation would require.

“Not as such. I would say that’s usually implicitly added into the risk management model, as GDPR as well . . .” (CertP6)

“So we definitely considered it from the point of view of the data and the criticality of the data, but we go with the categories of data, we don’t go the with the the context. You look at what type of information is being processed and then you are judging whether it’s suitable for the use, whether the principle of data minimization for example and all the other principles are being imposed for that type of information.” (CertP2)

“In terms of talking about let’s say a product or also a processor service, as we call it, then it’s of course important who is addressed by that service. Is it, for example, exclusively offered to, let’s say lawyers or doctors, than you would of course have to consider: Since it’s a lawyer, I would mainly first of all consider legal specifics in a certain setting. But you could also of course go beyond that, you have to understand how the specific sector in which something is used, how this context works, because otherwise you you cannot really go for certification there. So you need at least a basic understanding of that. So these are definitely aspects that would matter here and in terms of the context of processing also and then it’s getting related somehow to the risk that is associated to the processing. This of course also important in terms of applying our criteria and
principles an it’s somehow similar probably, but also different if the data subjects would be the users; then of course it’s once again important to know, are they patients who want to make use of some, let’s say health app or some internet portal.” (CertP5).

3.6 Discussion

We present in Figure 1 an overview of the themes and their sub-themes that were identified in the analysis above. The themes are presented hierarchically and in relation with the concepts from [11] that are being evaluated. At the top of the hierarchy we have a theme that is closest related to our main research question. This is a confirmation of the importance of the topic we approach in our research, namely “We need to both evaluate and measure usability of privacy”.

Looking closer at both the “evaluation” and “measuring” parts we see that the participants are preoccupied with finding out how to go about measuring – “Measuring is definitely useful but where do we start?” – and also what are the privacy aspects that are relevant to usability and usability evaluations. “Transparency and data protection rights” seems to be the motif that appears mentioned in several context, for example as something important to measure, or
as an “Instance of the usable privacy definition”. The themes related to “Transparency and data protection rights” overlap with the categories that we form for our Usable Privacy Goals and Usable Privacy Criteria, both being evaluated favorably by our participants. We conclude that this is an important area to consider and prioritize when approaching these research questions.

We see one other overlapping area between our research in [11] and the answers of the participants, which is the tendency to identify ‘instances’ of usability relevant for privacy. The Usable Privacy Goals represent instances of usability aspects that we identified in the GDPR text, while the participants have been pointed to such instances both when asked to validate our Usable Privacy Definition, as well as when they presented their understanding of usable privacy/data protection.

The theme that we present at the top of the hierarchy “Measuring is definitely useful but where do we start?” resurfaces in relation to our UP Criteria, as the participants are preoccupied with finding and defining concrete ways of meeting the Usable Privacy Criteria, such as which exact HCI methods to use in the evaluation.

The Usable Privacy Cube is a model that brings together all our other concepts and shows how one could integrate evaluations of privacy in the existing certification schemes. The Cube is also our solution to how to approach the evaluation of privacy. Hence the grey box that encloses the UP Definition, Goals, and Criteria, and the arrow pointing from the “Measuring is definitely useful but where do we start?” to the UP Cube. We have validated our assumption that indeed the concepts in the model represent at a more abstract level elements that are already considered by the auditors in their evaluations. However, these are mostly “implied and covert principles”, and especially those that are important for usability have often little influence on the final appraisal.

## 4 Conclusion

In the present study we have validated concepts that were introduced in our paper [11], namely: (i) a Definition of Usable Privacy, (ii) a set of 30 Usable Privacy Goals (which can be seen as instances of the Usable Privacy definition), (iii) Usable Privacy Criteria, and (iv) the Usable Privacy Cube model. These concepts have been introduced as the first steps towards creating a methodology for evaluating, and measuring on scales, usability related aspects of the data protection law. We interviewed experts from three relevant fields of practice and research: data protection law, privacy/data protection certifications and standardization, and usability (spanning fields such as Human-Computer Interaction, Usable Privacy and Security, or User Experience). The experts were asked to express their knowledge, understanding, and/or opinions on the above four concepts/topics.

The study plan used one group of experts to address one specific topic; the expertise of the group was thus thought to match the topic. However, we did not hold interviews with focus groups, but with the individual participants. Therefore, the analysis of each of the four topics is done within the frame of one group. Nevertheless, the topic of usability in privacy being more general was addressed by all participants and was therefore analysed across the groups. Moreover, we sometimes found answers in one group to be relevant for another
topic than the one in focus. We thus often use such additional opinions to strengthen the findings within a group.

A second design aspect of the study was to have two main parts: one where the interviewees present their opinions without being influenced by our research, and another where we present to them our research after which we ask them to directly comment on what was presented. The answers from the first part were used often to corroborate the responses after the presentation, and we found that the participants were consequent with their opinions, the change being only in adapting their answers to what was relevant for our research topics.

The limitations of our study come from the fact that though we initially intended to have an equal number of participants for each group, it was a challenge to find people with a background in law that would also have a decent level of understanding of usability matters. One solution that we can see for future work is to organize workshops where we recruit participants with only knowledge of digital privacy and data protection for IT systems and use more time to introduce them to usability related theory. In addition, as we use an interpretative framework based on pragmatism, we wished for a better balance than we achieved between the industry and academic backgrounds in the case of the ‘law’ and ‘usability’ groups.

The opinions expressed in the interviews are substantially endorsing our research approach, both directly and indirectly. We give an example of an indirect confirmation: when the ‘usability group’ was asked to validate our definition of usable privacy, after agreeing in majority with our approach, they suggest more specific aspects to be taken in consideration. Looking for which are the specific areas of data protection that usability is relevant for, is what we do when compiling our list with Usable Privacy Goals. The analysis shows that there are overlaps between the UP goals and their groupings and the ‘instances’ proposed by the participants. One such prominent example being transparency and data protection rights.

A characteristic of all the participants was to look for more concrete, particular, and practical aspects to address in the future and to suggest possible solutions. For example, in conjunction with validating the UP Criteria, they were pointing out what is yet to be done to meet these criteria, and even proposing possible solutions, as for example: “...some sort of notes from reviewers or tips. Perhaps even – let’s say that this would be a formal evaluation – saying in advance what is it that you’re looking for, so people can start thinking: Okay, if these are the things that get scored, what are the practises that are considered good so that we can start doing them on our own?” (UsabilityP3). Some of these overlap with what we propose as further work in our paper [11], as giving examples of which HCI methods to use in the evaluation. Other important aspects mentioned as necessary in order to help the organizations meet these criteria was to give them some design patterns to follow, or example of best practices. These answers constitute a good list of open problems that the community can address.

References


A Demographic questions addressed to all participants, irrespective of group

1 Demographics - research or/and other work experience relevant for data protection

1.1 Research experience

* Do you have RESEARCH experience relevant for data protection?

NOTE: For other types of work experience, see next question.

☐ No
☐ Yes

* If you are NOW or have PREVIOUSLY been a researcher, please specify which role do/did you have.

☐ PhD
☐ PostDoc
☐ Associate Professor
☐ Professor
☐ Researcher in company/organization
☑ Other (please specify)

* Specify other research roles

* Number of research years (including PhD)

Only values of at most 60 are allowed

1.2 Work experience

Do you have WORK experience (other than research) relevant for data protection?

- No
- Yes

WORK EXPERIENCE (other than research) relevant for data protection.
- Any position (voluntary or paid) that includes some tasks relevant to data protection.
- TYPE of organization (e.g., Data Protection organization, ISO organization, certification organization, law firm, IT firm).

NOTE: Please do NOT mention the name of the organization.

<table>
<thead>
<tr>
<th></th>
<th>Position 1</th>
<th>Position 2</th>
<th>Position 3</th>
<th>Position 4</th>
<th>Position 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position title</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organization type</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of years you have had this position</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Please mention any other studies/membership/functions/practice (current or previous) that are relevant for data protection.

Formatting: position name / organization type / number of years
1.3 Field of expertise

* Your MAIN field of expertise/study is closest to
  - Law/data protection
  - Usability/HCI/fxD/UX
  - Certifications/ISO Standards/Regulations
  - Other

Specify other MAIN field of expertise

Your SECONDARY field of expertise/study is closest to
  - Law/data protection
  - Usability/HCI/fxD/UX
  - Certifications/ISO Standards/Regulations
  - Other

Specify other SECONDARY field of expertise

1.4 Gender

* Your gender
  - Female
  - Male
  - I prefer not to say
  - Other (please specify)
B  Topics discussed with the participants in all three groups

2 Your perspective/experience with usability in data protection

- Explain your understanding of what usability is in the context of data protection.
- What is your experience with usability for data protection?
- What do you associate the usability concept with, i.e., which situations?

C  Topics discussed with the ‘usability group’

C.1  Usable Privacy Definition

3.1 Our definition of Usable Privacy

The definition of Usable Privacy presented in the video captures the current understanding of usable privacy in your field:

<table>
<thead>
<tr>
<th></th>
<th>Completely</th>
<th>Partially</th>
<th>Not at all</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Check one of the following</td>
<td>☐</td>
<td>☺</td>
<td>☒</td>
</tr>
</tbody>
</table>

Please explain your choice.
C.2 Usable Privacy Criteria

6 Usable Privacy Criteria

Two main questions:

- Are our kind of criteria (see more examples below) and how they are formulated a good/relatively good/bad solution for evaluating usability of privacy in a way that can result in concrete measurements?
- Do you know/envision other solutions that could be used for this purpose, of concretely measuring usability of privacy?

More examples of Usable Privacy Criteria:

UPG.2 Natural persons should have control of their own personal data. [Recital (7) of GDPR]

UPC.1 What is the level of control the data subjects have over their data?
[Based on goal UPG.2][Type of criteria: generic]

UPC.1.1 How much is the user in control of the personal data? [Effectiveness][Measure:Objective]

UPC.1.2 What is the user’s perceived level of control? [Effectiveness][Measure: Perceived]

UPC.1.3 How much time, effort, financial and material resources does the data subject need to invest in order to have control over the processed data? [Efficiency][Measure:Objective]

UPC.1.4 How much time, effort, financial and material resources does the data subject perceives that s/he needs to invest in order to have control over the processed data? [Efficiency][Measure: Perceived]

UPC.1.5 What is the frequency the user makes use of the tools put to her/his disposition for controlling the data? [Satisfaction][Measure: Objective]

UPC.1.6 What is the level of satisfaction of the users with the achieved level of control? [Satisfaction][Measure: Perceived]

UPG.3 Consent should be given by a clear affirmative act establishing a freely given, specific, informed and unambiguous indication of the data subject’s agreement to the processing of personal data relating to him or her. [Recital (32) of GDPR]

UPC.8 Is consent given by a clear affirmative act establishing a freely given, specific, informed and unambiguous indication of the data subject’s agreement to the processing of personal data relating to him or her? [UPG.3][Type: Consent]

UPC.8.1 How much of the consent text does the data subject understand? [Effectiveness: Completeness]

UPC.8.2 How much of the implications of consenting does the data subject understand? [Effectiveness: Completeness]

To what degree the data subjects perceive the agreement to be [UPC.8.3 freely given, UPC.8.4 informed, and UPC.8.5 unambiguous]? [Satisfaction: Cognitive responses]

UPG.27 Facilitate the exercise of the data subjects rights under Articles 15 to 22 – right of access, right to rectification, right to erasure, right to restriction of processing, right to data portability, right to object and automated individual decision-making. [Article 12 (2) of GDPR]
Are the rights under Articles 15 to 22 of the data subjects – right of access, right to rectification, right to erasure, right to restriction of processing, right to data portability, right to object and automated individual decision-making – facilitated? [UPG.27]

[Type: Rights]

UPC.16.1 How much [Time / Effort / Financial / Material resources] does the data subject spend in order to exercise her/his rights? [Efficiency: Time used, Human effort, financial and material resources expanded]

UPC.16.2 How many of the rights under Articles 15 to 22 are facilitated, and to what degree? [Effectiveness: Completeness]

UPC.16.3 To what degree the data subjects perceive that their rights are facilitated? [Perceived][Satisfaction: Cognitive responses]

UPC.16.4 What is the percentage of data subjects able to exercise their rights with ease, when probed? [Effectiveness: Accuracy]

6.1 The process of evaluating usability of privacy

- How would you, if at all, integrate these criteria in a process of evaluation?
- What would be the challenges/advantages with doing that?
D Topics discussed with the ‘law group’

D.1 Usable Privacy Goals

6 Usable Privacy Goals

Choose the goals that you think are related to usability (can be multiple or all of them).

☐ Ensuring a high level of protection of personal data. [Recital (6) of GDPR]
☐ Natural persons should have control of their own personal data. [Recital (7) of GDPR]
☐ Consent should be given by a clear affirmative act establishing a freely given, specific, informed and unambiguous indication of the data subject’s agreement to the processing of personal data relating to him or her. [Recital (32) of GDPR]
☐ If the data subject’s consent is to be given following a request by electronic means, the request must be clear, concise and not unnecessarily disruptive to the use of the service for which it is provided. [Recital (32) of GDPR]
☐ Any information and communication related to the processing of personal data should be easily accessible and easy to understand. [Recital (39) of GDPR]
☐ Any information and communication related to the processing of personal data should use clear and plain language. [Recital (39) of GDPR]
☐ Make the natural persons aware of risks, rules, safeguards and rights in relation to the processing of personal data. [Recital (39) of GDPR]
☐ The specific purposes for which personal data are processed should be explicit. [Recital (39) of GDPR]
☐ The personal data should be adequate, relevant and limited to what is necessary for the purposes for which they are processed. [Recital (39) of GDPR]
☐ Personal data should be processed only if the purpose of the processing could not reasonably be fulfilled by other means. [Recital (39) of GDPR]
☐ In the context of a written declaration on another matter, safeguards should ensure that the data subject is aware of the fact that and the extent to which consent is given. [Recital (42) of GDPR]
☐ A declaration of consent pre-formulated by the controller should be provided in an intelligible and easily accessible form, using clear and plain language and it should not contain unfair terms. [Recital (42) of GDPR]
☐ The data subject should have genuine and free choice in giving the consent. [Recital (42) of GDPR]
☐ The data subject should be able to refuse or withdraw consent without detriment. [Recital (42) of GDPR]
☐ Carefully assess the existence of a legitimate interest of a controller taking into consideration the reasonable expectations of data subjects based on their relationship with the controller. [Recital (47) of GDPR]
☐ Any information addressed to the public or to the data subject should be concise, easily accessible and easy to understand. [Article 12 (1) and Recital (58) of GDPR]
☐ Any information addressed to the public or to the data subject should use clear and plain language. [Article 12 (1) and Recital (58) of GDPR]

☐ Any information addressed to the public or to the data subject should use, when appropriate, visualization. [Recital (58) of GDPR]

☐ Provide information of the intended processing in an easily visible, intelligible and clearly legible manner. [Article 12 (7) and Recital (60) of GDPR]

☐ Provide a meaningful overview of the intended processing. [Article 12 (7) and Recital (60) of GDPR]

☐ A data subject should have the right of access to personal data which have been collected concerning him or her, and should exercise that right easily and at reasonable intervals, in order to be aware of, and verify, the lawfulness of the processing. [Recital (63) of GDPR]

☐ Allow the data subjects to quickly assess the level of data protection of relevant products and services. [Recital (100) linking to Article 42 of GDPR]

☐ The request for consent should be presented in a manner which is clearly distinguishable from the other matters, in an intelligible and easily accessible form, using clear and plain language. [Article 7 (2) of GDPR]

☐ It should be as easy to withdraw as to give consent. [Article 7 (3) of GDPR]

☐ Facilitate the exercise of the data subjects rights under Articles 15 to 22 – right of access, right to rectification, right to erasure, right to restriction of processing, right to data portability, right to object and automated individual decision-making. [Article 12 (2) of GDPR]

☐ The data subject should obtain from the controller meaningful information about the logic involved, as well as the significance and the envisaged consequences of automated decision-making, including profiling to which s/he is object to. [Article 15 (1) (h) of GDPR]

☐ The right to object should be explicitly brought to the attention of the data subject and should be presented clearly and separately from any other information, at the latest at the time of the first communication with the data subject. [Article 21 (4) of GDPR]

☐ The data subject should have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her. [Article 22 (1) of GDPR]

• Do you have any comments regarding the given list, or the choices you have made?
• Can you give other examples of usable privacy goals that you know of from the GDPR text or point to areas in GDPR where such goals could be found?
• Do you think that the examples you gave are exhaustive, covering all GDPR text?
• Do you think that there might be other similar goals that you do not remember now exactly?
E  Topics discussed with the ‘certifications group’

E.1 Evaluating and measuring usability of privacy

2.1 Evaluating usability of data protection

- Do you think it is important/not important to evaluate usability aspects and why?
- How much weight should be put on usability evaluations, compared with other aspects of the data protection law?

- Do you know of certifications that include usability aspects in their processes of evaluation of compliance with data protection?
  
  ☐ Yes
  ☐ No

- Which usability aspects do they consider?
- How much weight do they put on such evaluations?
- Are you satisfied with the way usability aspects are currently integrated in these processes?

2.1.1 Measurements

- Do you see as useful to concretely measure and evaluate how well the usability of privacy is dealt with by companies wishing to be GDPR compliant?
E.2 Usable Privacy Cube

4 Our model

Video with the presentation of the UP Cube model: https://vimeo.com/571358474

4.1 Your opinion on the UP Cube model -- A general view.

Does the UP Cube model reflect the processes that you know of, at a more abstract level?

<table>
<thead>
<tr>
<th>Completely</th>
<th>Partially</th>
<th>Not at all</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Check one of the following" /></td>
<td><img src="image" alt="Check one of the following" /></td>
<td><img src="image" alt="Check one of the following" /></td>
</tr>
</tbody>
</table>

• What are the differences and/or the similarities between the UP Cube model and certification models you know of?
• Which aspects of the UP Cube model are considered in certification processes you know of?
• What do you think might work well/not that well with the UP Cube model?
• Are there any aspects of the UP Cube model that you think might facilitate usability evaluations?

4.2 You opinion on the UP Cube model -- A detailed view.

• Are data subjects’ rights considered in the certification processes you know of? How and where in these processes are they included?
• Are data protection principles considered in the certification processes you know of? How and where in these processes are they included?
• How are the certification processes you know of dealing with areas of interactions between the data protection principles and rights?
• How and where in their processes, the certification schemes that you know of, consider the context of processing? What is the logic behind these considerations? From which practical needs do they arise?
Paper IV

Revealing Human-to-Computer Bias Transference

Johanna Johansen, Tore Pedersen, Christian Johansen

To appear in Springer's journal *AI & Society: Knowledge, Culture and Communication.*

Revealing Human-to-Computer Bias Transference

Johanna Johansen, Tore Pedersen, Christian Johansen

1Department of Informatics, University of Oslo
2Department of Psychology, Bjørknes University College
3Norwegian University of Science of Technology

Abstract

It is generally agreed that one origin of machine bias is resulting from characteristics within the dataset on which the algorithms are trained, i.e., the data does not warrant a generalized inference. We, however, hypothesize that a different ‘mechanism’, hitherto not articulated in the literature, may also be responsible for machine bias, namely that biases may originate from (i) the programmers’ cultural background, including education or line of work, or (ii) the contextual programming environment, including software requirements or developer tools. Combining an experimental and comparative design, we study the effects of cultural and contextual metaphors, and test whether each of these are ‘transferred’ from the programmer to the program, thus constituting a machine bias. Our results show that (i) cultural metaphors influence the programmer’s choices and (ii) contextual metaphors induced through priming can be used to moderate or exacerbate the effects of the cultural metaphors. Our studies are purposely performed with users of varying educational backgrounds and programming skills stretching from novice to proficient.

1. Introduction

A cognitive bias is a systematic tendency to make an incorrect judgment. This is particularly prominent in situations characterized by uncertainty, e.g., when processing information that are too voluminous or too complex for the human brain to handle, or when forced to make a rapid judgment in a time-frame that is too short to review the information at hand, or when there is insufficient information for making the decision, like in underspecified software requirements in programming. This is because the brain’s preferred cognitive mode is the automatic and non-conscious System 1, also termed Intuitive thinking. In situations characterized by certainty, this usually works well because we are on a “familiar terrain” where the useful mental shortcuts employed by System 1 are adaptive and functional. The problem is that our brain employs System 1 also in situations characterized by uncertainty, when it should in fact activate System 2, often termed Analytic thinking, used for controlled and conscious cognitive processing (Kahneman, 2011).

Two key concepts in mental shortcuts, also termed heuristic processing, are mental accessibility (resulting from the availability heuristic) and mental representativeness (Tversky and Kahneman, 1974; Gilovich et al., 2002; Thaler and Sunstein, 2009). When something is easily retrievable from memory, we have a tendency to wrongfully regard it as something that is also occurring frequently, even if it is not. We may also make an incorrect judgment about an unfamiliar phenomenon by identifying superficial resemblances to a familiar phenomenon. Because it is cognitively effortful to identify substantial similarities between two phenomena, particularly in situations characterized by incomplete information and uncertainty, superficial similarities are more easily identified. In many instances this results in an incorrect judgment.

1 Corresponding author

Email addresses: johanna@johansenresearch.info (Johanna Johansen), tore.pedersen@bhioslo.no (Tore Pedersen), christian.johansen@ntnu.no (Christian Johansen)
1.1. Algorithmic Bias: Data or the Programmer

Media and the general public seem to assume that machines and algorithms are neutral and objective. However, it has been known for quite some time that complex algorithms, such as those from artificial intelligence, may exhibit biases such as: racial bias (Schlesinger et al., 2018), gender discrimination (Zou and Schiebinger, 2018) and other socially relevant types of biases (Friedman and Nissenbaum, 1996; Boyd and Crawford, 2012; Jobin et al., 2019), when processing information in the support of decision making (Corbett-Davies et al., 2017; Dressel and Farid, 2018; Grgić-Hlača et al., 2019; Vaccaro and Waldo, 2019). This phenomenon is commonly labeled machine/algorithmic bias (Chouldechova and Roth, 2020), and has been confirmed in different areas, e.g., in big data (Hajian et al., 2016), web (Baeza-Yates, 2016, 2018), autonomous systems (Danks and London, 2017). Among institutions that have raised concerns about the existence of “biased algorithms” are: the ACM US Public Policy Council; the EU Parliament; the New York City Council bill on “Accountability and transparency in algorithms for public agency support”; ERCIM (Rauber et al., 2019); World Wide Web Foundation and many more (Cath et al., 2018), joined by major publication venues such as Science and Nature (Obermeyer et al., 2019; Zou and Schiebinger, 2018; Gianfrancesco et al., 2018) and by scholarly books (Boden, 2008; O’Neil, 2016).

All the works above focus, however, only on the data that AI algorithms train on, and show how the data contains biases. We are not aware of works that study empirically the transfer of biases from the human programmer to the algorithm, although we have found related ideas mentioned in two recent articles (Baeza-Yates, 2018; Silva and Kenney, 2019). The present work provides insights into why the bias transfer phenomenon may occur. We operate within the same paradigm and with a similar agenda as those who study human behavior in multidisciplinary research themes such as Behavioral Economics (Tversky and Kahneman, 1974; Kahneman et al., 1991), Behavioral Transportation Research (Pedersen et al., 2011; Gärling et al., 2014) or Behavioral Artificial Intelligence (Pedersen and Johansen, 2019).

1.2. Two Main Questions

The present study investigates methodically, experimentally, and empirically the hypothesis of bias transfer in programming, providing a first convincing argument for inspiring more empirical studies to be taken in the same direction. As such, in this study our main focus is to find support for (or against) the hypothesis that people may unknowingly and inadvertently transfer biases to computer programs that they build. However, we do not study specific biases (such as gender or racial), nor do we test or suggest specific programming methods and tools that could counteract bias transfers, as this would be a task for future research. Some may argue that robust quality assurance procedures eliminate any instances of biases in algorithms, at least in professional programming environments. We leave out for now testing whether the quality assurance procedures themselves have inherent biases or miss some forms of biases, and focus only on proving that programmers may be a source of biases, and not only the data given to the program. We set to investigate two main hypotheses:

I Are biases being transferred from the human programmer to the program artefact?

This is studied in a basic form with a bias revealing test that we detail in Section 2, which we impose on the subjects of our study, as described in Section 4. The biases that we study are of both cultural and contextual nature.

II Can programmers be manipulated, i.e., primed by inducing a new bias, and is this new bias then transferred to the program?

In Section 3 we describe the method that we use to prime our study subjects towards the same biases studied for the first question. Subsequent sections then describe how we used the priming in our studies and their outcomes.

---


1.3. Contributions

The contributions of this paper are presented schematically in Figure 1. The vertical axis of the diagram represents types of biases (or degrees of influence) with the arrow indicating a direction from more general, long-term, and strong biases to more contextual, short-lived, and weak forms of biases. We study two extremes, namely biases originating from the background of a person (including socio-cultural, educational, and professional influences), and biases resulting from manipulations such as priming. However, in between, one may study other forms of biases originating, e.g., from working cultures (Blackwell et al., 2019) or from media and propaganda.

Our results will be presented in Section 7, showing how these two types of biases are being transferred from humans to programs, providing evidence that transfer of both cultural and contextual biases exists. In order to support such a new claim, the rest of the paper is devoted to the development of our studies and analysis of the data.

We study users with different programming skill levels, i.e., from professional to amateur (Markopoulos et al., 2017a; Paternó and Santoro, 2019). This is motivated by the observation that increasingly more lay people (wrt. programming) are interacting and designing rather complex systems (Manca et al., 2019). Nowadays it is not only expert developers that program, but people with all levels of expertise carry out various programming-like tasks, from simple configurations of IoT systems in their smart home (Ur et al., 2016; Markopoulos et al., 2017b; Brich et al., 2017), to more complex installation and management of technology systems in their work, to more unconventional forms of programming using visual languages (Erwig et al., 2017; Akiki et al., 2017) (such as Fraunhofer’s IoT Programming Language NEPO6 or Google’s Blockly) or domain specific languages, and even assembling ready-programmed components into a final software system as done, e.g., in the IBM’s IoT development environment. This is because of the proliferation of simple (abstract, graphical, etc.) programming languages and interfaces aimed at non-programmers to design domain specific information systems, e.g.: a biologist programming a DNA search or an oil-engineer programming a complex database search. Therefore, in our study we use a simple programming task presented as fictitious, i.e., a proxy task, where the participants are imagining that they are programming.

Thus, the first important element of our study is presented in Section 2 where we develop a bias revealing cognitive task, which can be used for both types of biases. This cognitive task allows respondents to answer only with one of the three rationales that are listed on the middle line of the diagram. In Section 6.2.1 we analyze how well our test worked, using one of our control questions.

Section 3 details the manipulation method that we used, involving priming participants with metaphors hidden in a fictitious ‘Philosopher’ story. How well these manipulations worked is studied again with a control question involving listing of ‘similar words’ in Section 6.2.2. We created three metaphors to match the three rationales, which in turn match with three kinds of educational lines (or views on life). This correspondence is reflected in the vertical alignment of these elements in the diagram. With this we study the influence of contextual metaphors, in addition to the cultural bias.

Thus, since we aim to investigate whether inducing a bias is effective it is important, in order to avoid any intrinsic de-biasing, to “hide” from the subjects the real goal of the study behind a seemingly unrelated goal; in our case we used the title “Study of natural language in programming”. This is a standard study approach in research on biases because many types of biases can also be experimentally induced using priming (Tulving and Schacter, 1990; Yonelinas, 2002). Once we have established in this paper whether or not priming also works in the setting of programming biases, other works can carry out more detailed studies about whether such priming already exists “out there” (intentionally or not) and what types of priming would work and to what degree.

We have thus chosen our participants to represent these three different backgrounds, as detailed in Section 5. To test the assumptions about our participants’ backgrounds we used one control question (analyzed in Section 6.2.3) asking them to rank the three ‘life-aspects’ listed on the bottom line of the diagram.

We thus also investigate whether people educated in programming exhibit less biases and are less prone to manipulation. This is meant to investigate a considerable opposition that our idea has encountered, i.e., that programmers cannot be biased when writing code. Moreover, we also aim to study whether it is possible to

---


experimentally induce a bias on this category of users, or if this particular category is more resistant to priming and bias-transfer to programs.

The rest of the paper is devoted to presenting in Section 4 the major phases of designing our surveys and our studies using usability testing, and analyzing the data and demographics, in Section 6.

Our present work is motivated by the need to prove or disprove the idea that human biases could be transferable to the programming artifacts. However, which types of biases and how ‘dangerous’ these might be are not the subject of this study. Other specific studies would have to be devised, maybe similar to the research on human biases developed in the psychology field.

2. A Bias Revealing Cognitive Task

Over the years, we have used a simple cognitive task (Townsend, 2003) (originally called Alice’s Alphabet Puzzle) in lectures on judgment and decision making. In Townsend’s book, this particular puzzle lists the letters on a horizontal line, where straight lined letters are placed above and curved lined letters are placed below the line. We changed the puzzle in a vertical position to trigger more the infinity of the line, and the balance of the sides. In the task, the audience is first shown, as in Figure 2, a sequence of the letters divided by a vertical line and then asked to decide on which side of the vertical line the next letter H should be placed and why it should be there.

When asked why, the respondents provide rationales that seem to fall into three categories, which we categorized as: (I) ‘balance’, (II) ‘shapes’, or (III) ‘algorithm’. The quintessence of their arguments are as follows:

I  Some argue that there should be an equal number of letters on each side of the line: since there are already four letters on the right side and only three on the left side, the next letter, H, should go on the left side, thus indicating a sense of ‘balance’.

II  Others argue that the straight-lined letters A, E, and F are on the left side of the line, whereas the curved-shaped letters B, C, D, and G are on the right side, something which makes it perfectly reasonable that H should be together with its “kin” on the left side, given the different ‘shapes’ of the letters.

III  Yet others argue that there is an inherent order (or pattern) in the sequence: e.g., some indicate the sequence Left-1, Right-3, Left-2, Right-4, thus suggest placing H on the right side due to perceiving the image/puzzle as having the characteristics of an ‘algorithm/pattern’.

This exercise is simple enough to reveal cognitive tendencies of System 1, instead of consciously engaging the System 2 analytical thinking, as usually done by more complicated tasks. More importantly, the bias cannot be avoided because there is nothing else in the picture to help the person when making the decision, and any placement is correct; therefore only something from either (i) the background of the subject, alternatively, (ii) an experimentally induced prime, could help with making the decision – or random choice.
3. Metaphors as Priming Method

We use the above task to ‘forcefully’ reveal a bias, albeit an innocent one (compared to racial or gender), which would have its origin in the cultural background of the person (e.g., education, line of work, hobbies). This will be used to test our first main hypothesis that we explained in the introduction. Besides that, in order to test our second hypothesis, we want to prime the subjects to non-consciously make a decision in one specific “direction” – being one of the three rationales that we identified in Section 2.

Our motivation for hypothesizing that programmers would non-consciously be affected by the prime, comes from the well-known effect of cognitive heuristics (Gilovich et al., 2002). Under conditions of uncertainty, where one does not know, but nevertheless has to make a judgment or a choice, one will nonconsciously base one’s judgments either on instances that spring easily to mind (i.e., the cultural background or the contextual prime triggers the availability heuristic), or on instances that resemble the current problem (triggering the representativeness heuristic). The judgment can also be made as an approximation to the most recent, the most related or the most relevant information (anchoring heuristic). In terms of such associative reasoning, the metaphors and metaphorical thinking are a strong source of influence on how we as humans view the world.

The essence of a metaphor is, according to (Lakoff and Johnson, 2008), simply that “we understand and experience one kind of thing in terms of another”. For example, an argument may be understood and experienced in terms of the metaphor war, where we may “attack weak points in others’ arguments”, we may “shoot down” others’ arguments, and we may “win or lose” arguments. In fact, metaphors are so pervasive and ubiquitous in our lives that we simply cannot do without them.

Metaphorical thinking is something that can even be manipulated, e.g., (Thibodeau and Boroditsky, 2011) study of the effect of metaphors on the general public’s preferences for crime-prevention measures. When reporting crime-rates in a fictitious city, crime was either described in terms of “a beast” or in terms of “a virus”. When exposed to the metaphor crime is a beast, the general public argued for harsher and more severe crime-preventing measures than what was the case when they were exposed to the metaphor crime is a virus. This can make us alter our view of the world, and most of the time we are not aware, neither of the fact that we think metaphorically, nor that our metaphorical thinking can be manipulated by governments, media, our employers, or others, either for commercial or political purposes.

3.1. Experimental Manipulation Using Metaphors

Our experimental manipulation is in the form of ‘a story about a philosopher who invented a puzzle’, in which we vary the embedment of a different ‘life-aspect’, i.e., forming three different versions of the story. We also had one control condition, i.e., the story without any life-aspects (no metaphor), intended for a comparison to the experimental groups. The three different life-aspects are:

A harmony and equality
B aesthetics and arts
C order and continuity

Figure 2: The puzzle game.
The metaphors include four words, placed in two groups, two words in the beginning of the story, and the other in the end, following indications from relevant literature (Lakoff and Johnson, 2008; Thibodeau and Boroditsky, 2011). The words are:

A. harmony and balance; then equality and fairness

B. aesthetics and beauty; then forms of arts

C. order and structure; then linearity and continuity

We hypothesized that each of the above life-aspects would metaphorically influence the participants in the respective group A/B/C to provide an explanation that could be interpreted as one of the rationales from Section 2, respectively rationale I/II/III (i.e., ‘balance’, ‘shapes’, ‘algorithm’).

The metaphorical primes were embedded in the following fictional brief story about the philosopher who was presented as the one who originally created the puzzle from Section 2. Each subject will read a story that differs only in the words shown inside square brackets below.

“A philosopher who lived a life filled with [harmony and balance | aesthetics and beauty | order and continuity] created the riddle used in the game that we ask you to imagine that you program on the next page. Although the philosopher is nearly forgotten today, we know that the philosopher influenced many contemporary philosophers’ view of the world. The most prominent influence seems to have been the importance of maintaining [equality and fairness | forms of arts | linearity and continuity] in life and in society.”

4. Designing the Studies

Since the hypothesis that we want to test appears to be hard to accept (at least by programmers), we need to build a case backed by strong, data-driven evidence; therefore the complexity of our studies.

First, we incorporate the two instruments described in the previous sections into a programming task. We use a “paper-task” (described in Section 4.1) where the subject imagines to be programming, so that we can easily involve non-programmers, since part of our hypothesis is that people with various backgrounds (outside computer science) are involved in various “types of programming”.

We then incorporate the programming task into the survey described in Section 4.2. This contains additional questions to collect information for different purposes, e.g.: identifying ‘unserious subjects’, i.e., subjects that did not pay sufficient attention to the task, but instead responded randomly; or for helping with the interpretation of the subjects’ explanations of their rationales and their background.

We carried out our work in two stages. First we performed pilot studies, which we used to make improvements to the design. In particular, we first carried out specific usability testing in one pilot survey (described in Section 4.3). Then we improved the survey by using eye tracking technology to make sure that the priming is being read and to see more of how the subjects would interact with the survey (see Section 4.4).

4.1. The Programming Task

We designed a fictitious programming task in which actual programming was not undertaken during the session, but where the focus was on the subject’s reasoning about the programming task. Thus, we informed the subject that she should imagine herself in the role of a programmer.

The task was to ‘program a game for children’ where the image from Figure 2 would be the game board. The game would consist of the player (which would be different from the subject/programmer) having to place the next letter H into one of the two designated empty boxes. Upon correct placement, the game (i.e., the programmer) would reward the player. The design of the boxes was purposely made in order for the game to be perceived as continuing downwards. This was done to reduce the risk of being confounded by unintended biases (i.e., to avoid the subject perceiving the game board as finite, with letter H being the last one).
The task description text can be seen as the “requirements” that programmers receive from their clients (or elicited during a requirements engineering process); sometimes these include, so called, “user stories”, which are realistic descriptions of the functionality of the software in terms of how a user (in our case, a player) would interact/work (in our case, play) with the software (in our case, the game). Our requirements contain one major intended omission (i.e., it is incomplete) in that it does not say what would constitute a “correct” placement of the letter H. In consequence, the subjects need to decide for themselves to which side of the line they should give the reward. We hypothesized that the uncertainty inherent in the task would elicit heuristic thinking prompted by either cultural or contextual metaphors.

In order to introduce the priming metaphor, the game board image was linked to the story of the philosopher by saying that this “puzzle” was created by the philosopher. This link was made after the pilot testing (see details in the respective subsection below). We hypothesized that, if the participants were offered a simple explanation of the origins of the puzzle, then the philosopher story, containing the primes, would prompt the subject to non-consciously choose an explanation similar to the inherent rationale in the respective prime.

The task description that we used is the following (see (Johansen et al., 2020, Appendix A) for exact layout).

“First, spend one minute imagining how you would be programming the simple task below. Then proceed to answer the following questions.

Imagine that you are a non-expert programmer who is developing a simple puzzle game. The game is based on a riddle made by the philosopher that you read about previously. Imagine that you have already drawn the game board that you can see below:

[The image from Figure 2]

Now you are going to program the player’s interaction with the game.

The player (not you, you are the programmer of the game) has to solve the puzzle by drag-and-dropping the letter H onto one of the two dotted boxes. The player is rewarded if the program accepts the placement of the letter H as the correct placement.”

4.2. The Survey

The survey is created in SurveyMonkey®, bilingual, the Norwegian respondents having the possibility to choose between English and Norwegian. Screenshots of all the pages of the survey are given in (Johansen et al., 2020, Appendix A).

‘Page 1: Introductory text’: presents the goal of the survey and how the data is going to be dealt with. The goal of the experiment is only partially disclosed, and the true hypothesis remains completely undisclosed. Since the respondents have various backgrounds, other than computer science, it was also important to mention that no prior knowledge of computer programming is required for taking part in the survey/task/exercise.

‘Page 2: Instructions’: contains information that we consider important for the respondents to know before starting the survey:

“The back button is disabled. You will not be able to go back to a previous question, so we ask you to read each question carefully, because some depend on the previous ones.

Please put effort into reading carefully everything on each page.”

Note that some text is being emphasized, in the case of skim-reading. We need the participants to actually read the texts in the survey for the primes to work and for understanding the requirements in the programming task. For the mTurk and SurveyMonkey respondents, who were paid, we also added information about required minimum time for completion (average completion time was 6 minutes).

‘Page 3: Philosopher story’: contains our story intended for priming, which we have detailed in Section 3. We experienced during the pilot tests that the participants might not read a text if the information there cannot be used for answering questions in the survey. Therefore, we added one question meant as extra motivation (see Fig. 3).

8 Platform for creating online surveys: https://www.surveymonkey.com
‘Page 4: Programming task’: contains the text from the previous Section 4.1. We had three questions on this page, only one of these being important for the study, i.e., it asks about the placement of the letter H. In order to conceal the importance of this question we added two more questions completely irrelevant for our experiment. However, all three questions are made to look like questions that concern the programming task, i.e., it makes the task more realistic. If we would have left only the question about the choice of placement then the subject could have observed the missing information in the requirements that we gave and thus perceived the task as less realistic.

‘Page 5: Self explanation of choice’ and ‘Page 6: Alternative explanations’: where the respondents give, respectively, choose, an explanation for the choices they made in the programming task. We detail these two pages in the next subsections.

The rest of the questions on the following pages are meant to gather more information that could influence the results of the experiment, i.e., one’s view on life, hobbies, educational background, and demographics (age, gender).

‘Pages 7: Ranking life-aspects’: where the three alternatives from Section 3 could be ranked.

"Please rank the following three pairs of life-aspects in the way that best reflects how you view life yourself (where 1 is the highest while 3 is the lowest).
[ Options: harmony and equality | aesthetics and arts | order and continuity. ]"

This is a form of self-evaluation, where the subjects express directly their order of preference for the three instances of priming metaphors (this is done after they have completed the main task, and they are not aware that they were themselves randomly exposed to one of the metaphors). If they rank the prime that they were exposed to highest, this might indicate that the prime has had an influence. The UI for ranking questions is made well by SurveyMonkey so that when the question is required, then the subject must indeed provide a ranking, and not just leave the default.

‘Pages 8: Words suggestions’: where the subjects could suggest one to three words characterizing each of the three life-aspects, from the “Ranking life-aspects” question. The open-ended format chosen for this page has several reasons.

• We wanted to have another way to check the metaphorical priming effect by looking whether, and how often, our priming words appear among the answers (see details in Section 6.2.2).

• We also wanted to gather more data for future studies; i.e., others could use some of these words in future metaphor studies.

‘Pages 9: Demographics’: where the subjects had 5-7 questions about age, gender, years of education, field of study, and leisure activities.

In the following section we explain the reasoning behind the way the questions are composed, as a result of the discoveries we have made during the pilot testing.

4.3. Pilot Testing for Usability

To improve the usability of the survey we performed several pilot tests. The first pilot test used the method of usability testing (Dumas and Redish, 1999), with our survey being the product under test. One goal with testing the survey for its usability is to see whether the explanatory texts, requirements and questions are written in a clear and easy-to-understand language. Moreover, since we intended to prime the subjects, we needed to make certain that the story of the philosopher was read carefully and not just skimmmed through.

The usability study (see more details in (Johansen et al., 2020)) involved five participants. Four of these participants have a background in computer science and one in arts & design. Subjects with these two types of backgrounds were going to be used in our full-scale experiments as well.

The participants were asked to take the survey while being observed by us, sitting next to them (one of us took the role of a moderator, while another researcher was only an observer). The test was run with one participant at the
time. Before taking the survey, the subjects were explained verbally the purpose with the test session, which was to help us improve the face-value quality of the survey, although the hypothesis was not revealed. They were also presented the order of the tasks: first they were to take the survey, without any interruption, and then were supposed to answer questions meant to elicit suggestions for improvements of the survey.

The test also helped with validating our initial decision of disabling the back button in the survey. For the priming to work, the participants should not realize the connection between their choice in the programming task and the ‘Philosopher story’. If the participants would understand at a later stage in the survey that such a connection existed, they should not be allowed to navigate back and read the ‘Philosopher story’ again. In this pilot study one of the participants had the back button purposely left enabled. This participant did just what we expected, s/he navigated back to the ‘Philosopher story’, read it again, and adjusted her/his answers to reflect the view of the philosopher and not her/his own as the question required.

4.4. Eye Tracking for Better Insights

A more exact way to reveal the behavior of the participants when reading the information, and which flow they follow, is by using eye tracker technology (Bojko, 2013). We created two versions of the ‘Philosopher story’ – a ‘Short story’ and a ‘Long story’. We employed summative research\(^9\) in combination with eye tracking methods for comparing and deciding which version was more effective.

The test was done in a usability laboratory set up with eye tracking equipment. We used a combination of single-subject and between-subject design, where each participant (ten in total) was exposed to only one of the test stimuli, so to avoid any carryover effects between the stories. Both stories contained the same priming words. The ‘Long story’ was created with the purpose of helping the reader to immerse in the story – by giving more background information on the philosopher – and preparing the participant for the ‘Programming task’. Since we intended to prime the subject, we needed to hide the priming words well in the story, so that unintended debiasing (e.g., reactance) would not occur. At the same time, a too long story could make the subject not read the whole text and thus possibly skip the priming words. A shorter version of the story would also reduce the cognitive burden on

\(^9\) Summative research implies comparing an interface or product to its other versions, competitors, or benchmarks (Bojko, 2013).
The eye tracking testing was thus meant to help us identify whether the subjects skip our priming words, and also how much cognitive effort (i.e., how much time) they put into reading the stories.

The heatmaps and gaze plots visualizations provided both spatial and temporal insight into how the participants interacted with the text on each page of our survey. We obtained information about which areas of the text were fixated and for how long, the number of fixations and the order in which the fixations occurred.

Interpreting this data we concluded that there was no noticeable difference in how the text, and especially the priming words, were read between the long and short version of the story. For both cases, the participants read the text thoroughly, line by line (Figure 3). This shows that the instruction on the ‘Philosopher story’ page about reading the story “carefully” had the wished effect. The difference in reading the long story in 1:10 minutes compared to 40 seconds for the short one, meant a reduction of ca. 50 percent in cognitive load and time, and thus we decided to use the ‘Short story’ in our full-scale studies.

Another aspect that we analyzed with the help of eye tracking was whether the question about the philosopher being a man or a woman works as extra motivation for the participants to read the story. We found out that in order to answer this question, the participants returned to reading the story several times. In addition to the motivational aspect, questions such as this one help in drawing potential attention away from our true hypotheses.

More aspects that we investigated with eye tracking are detailed in (Johansen et al., 2020, Sec.5.4).

5. Methods

5.1. The Participants

The participants were chosen based on their educational or occupational background, to span three main domains. This is meant to cover well different computer programming skill levels as well as socio-cultural influences, properties and preferences. We reason that, when enrolled in a certain university study line or field of work, people have already developed predominant skills and characteristics needed for the specific education or occupation.

We had three main cohorts of respondents, totaling ca. 300 respondents:

A ‘Social sciences’ cohort – composed of students studying psychology;
B ‘Natural sciences’ cohort – composed of students studying computer science; and
C ‘Arts and Culture’ cohort – composed of a group of participants working in the field of arts and design, a group of students studying theatre, and another group studying music.

This categorization based on the educational and professional background is confirmed by the analysis of the data obtained from the control question on the ‘Demographics’ page, specifically about which field of study or line of work the respondents affirm their background to be mainly consistent with. (See (Johansen et al., 2020, Sec.7.2) for details.)

Based on the conditions to which the respondents were exposed, we also categorize the three cohorts into:

I ‘helped’ and ‘confined’,

II ‘helped’ and ‘not confined’,

III ‘not helped’ and ‘not confined’.

The ‘not confined’ respondents took the survey in the environment of their choice, which was unknown to us, whereas ‘confined’ means taking the survey in a more controlled environment (i.e., the university auditorium). The ‘(not) helped’ classification refers to whether the respondents were (not) given alternative explanations to pick from, regarding their choice for the placement of the letter ‘H’. The ‘Social sciences’ cohort belongs to the category

---

10 The gaze was offset vertically by approximately one line. This was due to the mismatch between the Operating System version and the version of Eye tracking software at the time of testing. The offset has been consistent across the participants and did not affect our interpretations.
(III), as the survey they were given did not contain the ‘Alternative explanations’ page and they could take the survey at the time and place of their choice. The mTurk and SurveyMonkey respondents from the ‘Arts and Culture’ cohort were helped with ‘Alternative explanations’ and were free to choose the environment where to take the survey. The ‘Cultural studies’ students were also ‘helped’ but confined to a classroom, where the course leader and one of the authors were also present. The ‘Natural sciences’ cohort was both ‘helped’ and ‘confined’ as the survey was taken as part of their regular course-work.

The environment of the participants and the support they received is related to three main types of environment where (future) programming activities can take place in:

A **Typical professional programming environment**, where the programmer is ‘confined’ to an office space and has to her disposal all the professional resources necessary to fulfill her tasks. In our case, for the programming task and the required explanations, we tried to reproduce this type of environment for the group of computer science students, by both confining them to the classroom and course hours, and offering them helping answers.

B **Semi-professional environment**, where an expert in some technical field (other than programming, e.g., railway engineering) has professional tool support for simple programming/configuration, e.g., by using a GUI based programming tool or a graphical programming language. However, programming is not their main task or responsibility and thus are not supposed to put too much effort into it, which we consider as ‘not confined’. The mTurk and SurveyMonkey respondents were thus ‘not confined’ but ‘helped’.

C **Non-professional environment**, where people, e.g., in their homes, are configuring an IoT system without any professional support nor prior knowledge. The ‘Social sciences’ respondents were neither ‘helped’ nor ‘confined’, and can thus be seen to some extent as fitting this profile.

For the purpose of studying the influence of the priming, we further group the respondents from each cohort by the metaphor they have been exposed to (or not), according to Section 3.1:

A a control group which is not primed in any way,
B a group primed as in Section 3.1.A (which we call, ‘primed with harmony and equality’),
C a group primed as in Section 3.1.B (i.e., with ‘aesthetics and arts’), and
D a group primed as in Section 3.1.C (i.e., with ‘order and continuity’).

The control group is meant to serve as a baseline to observe what the programmers’ preferences for task-solutions are in the absence of primes. This is relevant for our first main question.

The three primed groups are meant to help us test whether the bias can be induced upon the programmer, and subsequently transferred from the programmer to the algorithms.

The cohort with students from the computer science study line is also meant to help us test whether programmers shut away the other two biases, except the pattern/infinite way of thinking, which is sometimes assumed that the programmers do.

5.2. Methods Employed

The studies employ a combination of experimental design and comparative design. In the analyses of both (i) the comparative aspect, i.e., differences between the three cohorts, and (ii) the experimental aspect, i.e., differences within each cohort, resulting from the experimental manipulation, we employed both (a) inferential statistics, more specifically chi-square analyses of categorical data, as well as (b) descriptive statistics to report frequencies and percentages. We performed an experiment on each cohort, as well as compared the three cohorts to each other, regardless of the experimental manipulation. Since the three cohorts were different in terms of cultural and educational background, we were able to study the unique effect of background per se.

Conforming to the true experimental design method (Lazar et al., 2017; Cook et al., 2002), we first assigned the participants of each cohort randomly to one of three experimental conditions where we induced one specific type of contextual metaphorical thinking in each, or to a control condition containing neither of the three primes. The
control condition contained the neutral non-prime story and was meant to serve as a “baseline” to establish whether the participants, without being primed, were inclined to favor one of the three “rationales” over the other.

The subjects are given the programming exercise described in Section 4.1. The programming task, the educational/professional background of the subjects, and the story containing the primes, are the independent variables in our experiment. The choice of what will be the right solution for the puzzle is the dependent variable. We are interested in finding out if the primes and the background of the participants (the independent variables) influence how the puzzle is programmed (the dependent variable), following the rationale that it is the programmer who decides to give the player a prize based on what the programmer thinks qualifies as the right answer.

The conditions (or treatments) that we intend to compare are reflected in the explanations that the subjects provide, being under the influence of three contextual metaphor primes and three types of cultural background.

The experimental conditions are controlled and kept constant to the extent that we recorded the time spent on the tasks and thus ensured that the tasks were completed within a reasonable time-frame. Thus, we excluded the effect of any seriously potentially confounding variables, such as diffusion of experimental manipulations (i.e., we reduced the possibility of participants sharing the contents of the tasks with other participants). Participants completed the task individually and received identical instructions, and the hypotheses were not revealed to the participants. Such non-disclosure of hypotheses is the most robust experimental procedure, and it is employed in around 87 percent of all experimental-psychology research (Hertwig and Ortmann, 2008) because it allows for the elicitation of valid measures of behavior instead of relying on less valid measures by means of other methods, such as self-reports (Bröder, 1998; Christensen, 1988; Kimmel, 1998; Trice, 1986; Weiss, 2001).

For analyzing the second main hypothesis that we proposed in the Introduction, pertaining to the potential influence of the context, the research hypothesis is that the manipulation (“prime”) will increase the number of the corresponding explanations the participants give. The participants’ explanations for their respective choices were qualitatively coded according to the three predefined categories. Explanations conforming to one of the three predefined categories were categorized both according to their discrete category (i.e., ‘balance’, ‘shapes’ or ‘algorithm’) as well as whether they were ‘sensical’ (i.e., eligible for inclusion in the predefined categories) or ‘nonsensical’. Non-interpretable explanations were thus labeled ‘nonsensical’ and discarded (see Section 6.1 for a thorough analysis of this). If the rationale of primemanipulation in the respective condition is chosen significantly more than the other rationales, this would imply that the participants were influenced by external features that are not relevant to the programming task itself.

We implemented one additional variable to control for the bias, resulting from an observation in our practical use of the cognitive task from Section 2, that the choice of placing the letter H is also an indication of the rationale. Particularly, participants choosing Left would be those using the rationales I and II from Section 2, whereas participants choosing Right would be those using the rationale III for ‘algorithm’. This is analyzed in Subsection 6.2.1.

Even though we chose the subjects based on their educational and professional background, we also asked them to provide information about their educational background themselves, as well as information about their preferred free-time activities. This is done in order to disclose a possible relation between this particular aspect of the background of the participants and their choices in the programming task. Moreover, this information from free-text questions can also help detect respondents that did not relate seriously to the task, as well as to control our qualitative coding of their explanations and background.

Alternative Explanatory Variables

The age and gender of the participants are analyzed as alternative explanatory variables. Other alternative explanatory variables that might occur could result from the subjects not understanding the task well, the task being too difficult, or the prime not being strong enough as a result of superficial reading. However, these factors were something that we detected and removed through our pilot tests.

By implementing the questionnaire questions related to individual preferences and extracurricular activities (‘Ranking life-aspects’ and ‘Words suggestions’ pages, and the question about hobbies on the ‘Demographics’ page), we expect to be able to clearly identify if the choice was dictated by the bias. Moreover, the programming task is mean to be very simple, thus requiring very little cognitive effort. For such cases, it is empirically proven that the individual differences have a small impact (Lazar et al., 2017).
5.3. Learning from the First Full-scale Study

The first full-scale study, also referred to as the ‘Social sciences’ cohort, consists of undergraduate students enrolled in a psychology study program. The link to the survey on SurveyMonkey was sent through email by the study program administrators and resulted in 77 responses. Observations made after the first study helped with improving the following studies. An analysis of the incomplete responses (31 out of 77) from the first study is shown in Figure 4.

The high number of dropouts on the ‘Programming task’ page could be explained by the fact that these psychology students may have deemed this task as not relevant, not interesting, or maybe too difficult. Based on this reasoning, in the subsequent studies we introduced on the first page the mention “It is not required to have any prior knowledge of computer programming.”, and on the ‘Programming task’ page we wrote that the puzzle is simple.

Another solution for further motivating the respondents to finish the survey was to add a progress bar indicating how much of the survey was left until completion. For the last three pages we also added a page-footer informing, consecutively, that ‘there are three, two, and one pages left’, i.e., aiming to reduce the two latter types of dropouts.

Recall that the first study was conducted with volunteer social science students that were neither paid nor participating during their normal class hours. In contrast, the second study was run in a lecture hall, before the break, as part of a first year computer science course. In the case of mTurk and SurveyMonkey respondents from the third study, we consider the payment as an important motivating factor (see Section 5.4).

To reduce the cognitive effort required (and maybe the dropout rate) on the ‘Self explanation of choice’ page we added, immediately after this page, the page called ‘Alternative explanations’. Moreover, adding these alternatives in the second study reduced the number of uninterpretable answers significantly (see Section 6).

For the second study, the total number of responses received was 53. Of this total number, one respondent dropped out on the ‘Programming task’, one on the ‘Self explanation of choice’, one on the ‘Ranking lifeaspects’, and two on the ‘Words suggestions’ page. The small number of dropouts in this second study indicates that the adjustments made after the first study were successful.

5.4. Transitioning from Volunteering Students to Professional Respondents

For the third cohort, we recruited people with a background in arts and culture in general. We started the third study with two groups of students, studying music and theater. Though we had no dropouts from these groups, the numbers of students in the classes were small (which is specific to these kinds of studies), i.e., 10 respondents from music and 10 from theater. To increase the number of responses we also recruited respondents through the specialized platforms Amazon Mechanical Turk and SurveyMonkey. These would no longer be volunteers but professional respondents who are paid for their participation and do such tasks often.

From the total of 128 responses we removed 13 respondents that spent less than four minutes on completing the survey (average response time from the previous studies was eight minutes). Additionally, seven more respondents were rejected as we deemed them unserious (e.g., computer generated answers). Out of the remaining 108 responses, one participant dropped out on the ‘Self explanation of choice’ page. Moreover, six respondents that spent more than four minutes were still deemed unreliable and thus removed from the analysis. This was decided based on the quality of the responses given in the open-ended questions. (See more details in (Johansen et al., 2020, Sec.6.4).)
6. Analyzing the Data

6.1. Sensical vs. Nonsensical Answers

The participants’ explanations were analyzed qualitatively and coded into one of the three rationales from Section 2. There were still many answers that could not be categorized, either because they did not make much sense, or the reason given was no reason at all. However, many of these answers were recurrent, transcending even the language differences, and this allowed us to group them in categories. (See (Johansen et al., 2020, Sec.7.1) for details.) To reduce the number of uninterpretable answers, starting with the second full-scale survey, we introduced the alternative answers which were formulated based on the wordings that we encountered among the responses from the first study. Thus, the first study provided a type of ‘saturation’ of alternatives.

We thus define as Sensical those answers that were interpretable and allowed for category inclusion in one of the three rationales from Section 2, and we define as Nonsensical the remainder of the answers.

We observe in Figure 5 a substantial increase of answers that allowed for interpretation when the respondents were offered the alternative explanation choices.

One can observe that as soon as the participants were confined their explanations became even more sensical. Here we look at the two cohorts to whom alternative explanations were offered, and we notice the increase from 70.50% to 84.31% in the case of the respondents from the ‘Natural sciences’ cohort who were confined to the classroom and course working-hours. This bears evidence that the transition from a non-professional towards a more professional programming environment would trigger the programmers to be more careful about their choices. Another aspect of the confinement is that it triggers the System 2 thinking, which is known to result in a reduction of human biases. We have also observed instances of System 1 vs. System 2 thinking, i.e., ‘starting’ as a System 1 response, but then ‘self-apprehended’ and activated a System 2 response; e.g.: a participant wrote in the ‘Self explanation of choice’ page “I choose right previously but actually left makes more sense. Balancing the sides; 4 letters on the left, 4 letters on the right.” (pID M:11272410463, third cohort).

6.2. Control Questions

In the study we included several additional questions with the purpose to control for various aspects. Three questions were of particular importance, as they were meant to control, or to reinforce, three important assumptions that we have. Essentially, Section 6.2.1 reinforces our bias revealing test from Section 2 as a good instrument; Section 6.2.2 tests how well our priming metaphors from Section 3 worked, since such story-based metaphors may be revealed within listings of words/synonyms; whereas Section 6.2.3 reinforces our beliefs and categorization of the backgrounds of the three cohorts that we study, thus confirming that the categories/labels we provided in Section 5.1 are appropriate, and the bias transfer results that we report in Section 7.1 are well informed.

6.2.1. Left/Right Placement

On the ‘Programming task’ page of the survey, the respondents are asked to decide whether to reward the player for the placement of the letter ‘H’ on the left or right side of the vertical line on the game board. This is one of the three questions on this page, intended as a control question for the hypothesis that we made in Section 2, i.e., that
choosing to place the letter to the ‘right’ should indicate a preference for the ‘algorithm’ rationale, while choosing ‘left’ a preference for the ‘balance’ or ‘shapes’ rationales.

An analysis of the ‘left/right’ placement wrt. each of the three rationales confirms this initial assumption, see Figure 6b for numbers. In particular, observe that in the case of the ‘algorithm’ rationale the choice of placement to the ‘right’ is overwhelming for each cohort; and similarly, ‘left’ is the preferred choice when answering with the ‘balance’ or ‘shapes’ rationales in all cohorts.

Moreover, the analysis of the ‘left/right’ placement overall inside each cohort, which we summarize in Figure 6a, confirms our earlier observation that the background of the participants is reflected in their preference for one choice of placement (and thus for one type of rationale).

6.2.2. Words Suggestions vs. Priming Metaphor

In (Johansen et al., 2020, Sec.8.2) one can find a thorough analysis of the words given by the respondents to the ‘Words suggestions’ question of the survey, which was meant to reveal which of the primes worked and how well.

The participants were given three pairs of words to suggest synonyms for, each containing two of the four priming words used in the ‘Philosopher story’. In analyzing the responses for the ‘Words suggestions’ question, we looked for the occurrence of the other two words that were used in the ‘Philosopher story’ as primes (cf. Section 3 also).

The numbers from our analysis shows that the priming metaphors for ‘balance’ and for ‘shapes’ were chosen well, whereas the words for the ‘algorithm’ metaphor were too difficult, which diminishes the strength of the priming.

This analysis also indicates two other factors that might have had influence on the priming effect. One is how familiar the respondents are with the priming words. If the words are very little known or not understood, people will not be primed by them, as it is the case of the ‘algorithm’ words. For the second factor, if the respondents have a large vocabulary at their disposal, the System 1 will be less inclined to use the priming words in this synonyms question. Such observations can be made in the case of the ‘Social sciences’ students in comparison with the ‘Natural sciences’ students: 147 unique words compared to 95. We see how the priming was stronger in the latter cohort compared to the former (they strive to find similar words, and the availability heuristic retrieves the primes from the short term memory).

The words the participants choose the most can also be affected by other immediate contextual elements. In the case of the ‘Natural sciences’ cohort the survey was taken by the students as part of a course on logic. This made the word ‘logic’ occur the most for ‘order and continuity’.

Figure 6: Overview of the choice of placement of the letter ‘H’ to the Left or Light.
6.2.3. Life-aspects Ranking

This was meant as a control question for the way we identify the background in our cohorts. Moreover, we want to also look at the coded answers from the ‘Self explanation of choice’ compared to the ‘Ranking life-aspects’ because if the correlation is similar to the one we have observed previously from the background, then this would reinforce our perception of background.

One’s view on life is, among others, highly influenced by society and culture (Cialdini and Goldstein, 2004; Schultz et al., 2007; Cialdini, 2009). For children this may be the main influence (e.g., through their parents), whereas for young adults (like many of our respondents who are young students) other factors of their own life-experience start to influence their views, including their education when they are studying and their professional environment when they start working. We summarize the three types of influences in Figure 7, organized as a pyramid to suggest the strength and time of the influence.

From the data analyzed in detail in (Johansen et al., 2020, Sec.8.3) we can conclude that the control question about ‘Ranking life-aspects’ confirms our assumptions about the backgrounds for our three cohorts and the fact that we have associated each of these cohorts with the life-aspect that is most predominant for those respondents. Therefore, we consider adequate the claims that we make throughout the paper where we correlate the background of a cohort with one specific life-aspect, and thus with one specific corresponding bias/rationale.

7. Results

The data is analyzed both quantitatively and qualitatively. The qualitative analysis is done to detail the quantitative data, by analyzing the participants’ responses to the open-ended questions. Since our study is exploratory, we employ a combination of statistical and descriptive analysis. Statistical analyses were not possible in all situations because of the small number of respondents in those categories.

7.1. Influences from the Cultural Background

Students with a cultural background from social sciences differed significantly from students with a cultural background from computer science. Social sciences students were significantly more prone than computer science students to describe their choices matching the rationale ‘balance’, whereas computer science students were significantly more prone to describe their choices matching the rationale ‘algorithm’: $X^2(1, N = 71) = 8.1686, p < .05$ (with calculated p-value of .004262). These results support the hypothesis that the cultural background influences people when they carry out programming tasks under conditions of uncertainty.

The statistical significance test, as well as the graph in Figure 8a consider the total number of responses, from all four treatments. The same observations about the cultural background influence are confirmed also when looking only at the control group (see the graph in Figure 8b), though a statistical test is not relevant in this case, given the small number of responses. For both graphs the percentages are calculated from the ‘sensical’ answers only.
When analyzing the results from the ‘Arts and Culture’ cohort in comparison with the other two cohorts (see Figure 9), we see that the influence of their artistic background makes them choose much more the ‘shapes’ rationale.

7.2. Influences from the Priming Metaphors

We observed influences of our experimental manipulations, albeit not reaching statistical significance. Thus, since we can neither rule out a Type I error nor a Type II error, in the rest of this section we present results from quantitative analyses of the priming effect and whether or not this transferred to the programs.

The graph in Figure 10 shows the influence of the three groups of priming metaphors when the responses from all the cohorts are put together. This shows the priming effect irrespective of the participants’ background. We compare each group with the control group.

First, we observe that the ‘algorithm’ group gives answers that cannot be readily seen as being influenced by priming. The same inconclusive observation is found also when looking inside each cohort, comparing the ‘algorithm’ group there with the respective control group. This conforms with the observations made in Section 6.2.2, where the words used for this priming are little known or maybe not understood by the participants, and thus cannot have an impact on their choice. Therefore, we focus in the rest of the section on the other two groups of priming.

Secondly, when we analyze the other two groups we clearly observe priming influences, albeit of different kinds as explained further. For the ‘balance’ group we see that the ‘balance’ rationale increases from 28.26% (for the

---

\[\text{Figure 8: Comparison of background influences (cultural biases transferred).}\]

\[\text{Figure 9: Comparison of number of answers categorized in the ‘shapes’ rationale.}\]
control group) to 43.18% (for the ‘balance’ group), whereas in the case of the ‘shapes’ group the ‘shapes’ rationale increase from 8.70% to 15.70%; irrespective of the background.

Besides these observations about the general strength of the priming metaphors, we are to a greater extent interested in their interactions with the educational/professional background of the participants, as discussed in the previous subsection. In the chart from Figure 11a, related to the ‘Social sciences’ cohort, we observe that the background of the respondents is further strengthened by the priming metaphors, with an increase from 58.33% to 100%. Quite the opposite, in the case of the ‘Natural sciences’ cohort in Figure 11b, we see a weakening effect of their background since the ‘algorithms’ rationale decreases from 71.43% to 58.85% in the group that was influenced with the ‘balance’ metaphor, in favor of the ‘balance’ rationale. For the ‘Arts and Culture’ cohort we again see in Figure 11c that the ‘shapes’ metaphor strengthens their background since the ‘shapes’ rationale increases from 15% to 22.22% inside the group that was primed with the ‘shapes’ metaphor. For this cohort also the ‘balance’ metaphor has an influence (from 15% to 34.62%), due to the fact that this metaphor’s words were well chosen, as we have observed in Section 6.2.2.

We can conclude that the contextual metaphors that have been deemed as strong enough in the control question ‘Words suggestions’ are also found to have an effect in strengthening or weakening the influence from metaphors in the cultural background of the respondents. Contextual metaphors have a strengthening effect when the words are representative of the respective cultural background. At the same time, well chosen contextual metaphors can
weaken the effect from the cultural background when they go against it, e.g., ‘balance’ metaphor applied in the ‘Natural sciences’ cohort.

8. Conclusions and Discussions

The aim of this study, as well as its implications, are manifold. The study can be categorized as both (i) a comparative/experimental study of how biases from cultural and contextual metaphors can be transferred from programmers to programs, and (ii) an exploratory study on how to develop ergonomically valid and reliable instruments, procedures and testing conditions to empirically study such biases transfer. This paper was intended as a foundation for future research endeavors to improve and diversify these instruments, procedures and testing conditions.

The strengths of this work reside in its exploratory nature in studying a hitherto not researched phenomenon, namely the transfer of human biases from the (not necessarily expert) programmer to the artifact that is developed (or configured). Concretely, we have exposed (in Section 7) interesting aspects of our main hypothesis, namely that machine bias may originate also from the programmer’s biases in terms of influences from the cultural background as well as contextual influences from the programming environment. Interestingly, attempts to moderate the strong influence from the cultural metaphors by means of experimentally introducing ‘hidden’ contextual metaphors, were only successful to a certain extent. When the priming metaphor was chosen well we observed influences in both directions of strengthening the cultural background as well as moderating it, each time tipping the balance of answers in the direction of the metaphor.

8.1. Possible Future Research Directions

One possibility is to perform similar studies focused on specific categories of subjects that can be seen as programmers, e.g., one playful possibility could be to study children as programmers – programming languages/environments specific for children abound, such as Google’s Blockly (Trower and Gray, 2015; Weintrop and Wilensky, 2017) or MIT’s Scratch (Resnick et al., 2009; Maloney et al., 2010; Armoni et al., 2015).

One useful refinement of our work could be to study professional programmers in a professional environment, both (i) classical programming environments guided by software development life cycle methods and tools, (ii) non-expert programming environments, such as complex configurations, DSLs, graphical programming, or curating of big data. One question can be: What are the avenues that bring biases into the programming environment? We have assumed that biases are a result of underspecified requirements. This is a common form of uncertainty in programming; but there are others as well. It is thus important to know which of these give way to biases, so that one can build debiasing techniques (Jolls and Sunstein, 2006; Blackwell et al., 2009; Cheng and Wu, 2010), maybe incorporated in the IDEs of the programmers (similarly to how others have developed culturally adaptive user interfaces (Reinecke and Bernstein, 2011)).

We have studied two sources of biases, namely cultural metaphors and priming, that we consider situated at the two extremes on the vertical axis from Figure 1, which indicates the strength of the bias, and also a temporal aspect regarding the persistence of these biases (e.g., priming may not be as strong as the culture, and acts on a short time scale, usually minutes after the priming is applied). One could study other sources of influence that would lie in between on our vertical axis, e.g.: propaganda (i.e., misinformation (Mintz et al., 2012; Kumar and Geethakumari, 2014) and disinformation (Graham and Metaxas, 2003)), which may be done on limited but considerable stretches of time; or working cultures which can influence a programmer in different ways when changing jobs.

References


Studying the Transfer of Biases from Programmers to Programs

Johanna Johansen\textsuperscript{a,∗}, Tore Pedersen\textsuperscript{b,c}, Christian Johansen\textsuperscript{d}

\textsuperscript{a}Department of Informatics, University of Oslo
\textsuperscript{b}Center for Intelligence Studies, Norwegian Defence Intelligence School
\textsuperscript{c}Department of Psychology, Bjørknes University College
\textsuperscript{d}Department of Technology Systems, University of Oslo

Abstract

It is generally agreed that one origin of machine bias is resulting from characteristics within the dataset on which the algorithms are trained, i.e., the data does not warrant a generalized inference. We, however, hypothesize that a different ‘mechanism’, hitherto not articulated in the literature, may also be responsible for machine’s bias, namely that biases may originate from (i) the programmers’ cultural background, such as education or line of work, or (ii) the contextual programming environment, such as software requirements or developer tools. Combining an experimental and comparative design, we studied the effects of cultural metaphors and contextual metaphors, and tested whether each of these would ‘transfer’ from the programmer to program, thus constituting a machine bias. The results show (i) that cultural metaphors influence the programmer’s choices and (ii) that ‘induced’ contextual metaphors can be used to moderate or exacerbate the effects of the cultural metaphors. This supports our hypothesis that biases in automated systems do not always originate from within the machine’s training data. Instead, machines may also ‘replicate’ and ‘reproduce’ biases from the programmers’ cultural background by the transfer of cultural metaphors into the programming process. Implications for academia and professional practice range from the micro programming-level to the macro national-regulations or educational level, and span across all societal domains where software-based systems are operating such as the popular AI-based automated decision support systems.

Keywords: Biases, programmers, AI, cultural background, metaphors, priming, randomized controlled trial.

1. Introduction

Biases are manifestations of incorrect judgments resulting from cognitive tendencies that humans exhibit in situations of uncertainty or when there is insufficient information but a judgment must be made nevertheless. Biases are often difficult to study because of the complex thinking ‘machinery’ that makes up the human brain and because of the human’s interaction with its complex social environment. Moreover, people are usually unaware of their own biases and they may even be prone to rationalize their own biased tendencies. Nevertheless, by employing carefully designed experiments, both the specific psychological mechanisms that lead to biases and the occurrences of specific human biases have successfully been identified and well established (e.g. Gilovich et al. (2002); Tversky & Kahneman (1974); Oliver (2014); Wilson & Gilbert (2003)).

Due to the increase of power and societal penetration of artificial intelligence (AI) algorithms, the occurrence of artificial biases have been observed and described, something which is rather counter to our intuition that machines are unbiased and objective. Biases in AI have a strong negative impact in society, prompting organizations like ACM and the European Parliament to issue strong statements ACM Policy Council (2017); STOA (2019) and prominent researchers to publish lengthy reports Brundage et al. (2018) warning against biased AI.

The occurrence of biases in AI have been observed to appear due to biased training data on which these algorithms are built Feldman et al. (2015); Mittelstadt et al. (2016); Caliskan et al. (2017); Silva & Kenney (2019). For example, if the limited sample-data that the algorithms are trained on are not sufficiently representative of the larger population-data that the algorithms are subsequently unleashed upon, then the algorithmic judgments would not be valid but instead biased. Or, for example, if the training data shows a strong relationship between two variables, say, ethnicity and crime Zou & Schiebinger (2018); Dresel & Farid (2018), a bias may occur because it is assumed that the relationship is causal when the relationship between the two variables may in fact be non-causal and instead be caused by a third variable, say, poverty.

However, we have reasons to believe that there is a second source of biases in algorithms/programs, which has been little investigated Kirkpatrick (2016). We believe...
(and argue in this paper) that biases may also be transferred from the human programmer into the final artifact, i.e., the program. Transfer (or contagion) of biases between humans is well-known, such as, for example, the conformity bias (e.g., Moscovici & Fauchex (1972)). Additionally, a transfer of biases due to influence on humans from social and cultural institutions like media or education is equally known (e.g., Bourdieu & Passeron (1977); Lakoff & Johnson (2008)). Given the human cognitive tendencies (explained in Section 2) to employ inappropriate mental judgment-modes in situations that are “uncertain”, combined with influences from institutional agendas, human biases are ubiquitous. There is, however, yet no evidence to support the related possibility of programmer biases being encoded, in some way, in their programming artefacts. In fact, some argue against this view by claiming that programmers are “immune” to biases due to their technical training and tools/methods used.

**Target Audience.** The work reported in this paper is relevant for researchers from several fields. First of all, people working in AI and machine learning can be interested in our proposal that biases in machine learning can come not only from the data but also from the people programming the algorithms. We study this to some considerable detail, as we explain in the rest of this introduction. Second, people working in psychology and cognitive sciences can be interested in this new application that we propose, where they can apply their skills and methods to study this new form of human bias and its transfer to machines. Third, practitioners working with software engineering or managing software development teams can be interested in studying more various programming environments and tools to see how much human bias is transferred to the programs in each situation. Finally, at a macro level, both governments, private business enterprises, and NGOs would become aware of machine bias originating from human programmers who unknowingly transfer the influences from their own cultural backgrounds to the machine programs. Thus, the target audiences are diverse and would benefit both on a micro level, e.g., in research and development, and in (computer science) education, as well as on a macro level, e.g., in issuing improved knowledge-informed national regulations on the domains where automated decision-support systems operate.

### 1.1. Two Main Questions

The present study investigates methodically, experimentally, and empirically the hypothesis of bias transfer in programming, providing a first convincing argument for inspiring more empirical studies to be taken in the same direction. As such, in this study our main focus is to find support for (or against) the hypothesis that people may unknowingly and inadvertently transfer biases to computer programs that they build. In particular, this work supports the hypothesized existence of an alternative mechanism than the one that is already known, that may render AI biased, i.e., in addition to the well-studied cases of biases originating from data. However, we do not study specific biases (s.a., gender or racial), nor do we test or suggest specific programming methods and tools that could counteract bias transfers, as this would be a task for future research.

The two main hypotheses that we set to investigate are:

I. **Are biases being transferred from the human programmer to the program artefact?**

This is studied in a basic form with a bias revealing test that we detail in Section 3, which we impose on the subjects of our study, as described in Section 5. The biases that we study are of both cultural and contextual nature.

II. **Can programmers be manipulated, i.e., primed by inducing a new bias, and is this new bias then transferred to the program?**

In Section 4 we describe the method that we use to prime our study subjects towards the same biases studied for the first question. Subsequent sections then describe how we used the priming in our studies and their outcomes.

One can see several immediate benefits of the present study alone. For example, in education one could measure how well programming courses train the students, by measuring the bias transfer-rate at the start and end of the courses. Another example is to measure how effective some technology quality assurance method is at removing or identifying programmer’s biases, like testing frameworks, peer programming, abstract/detailed specifications, code generators, etc. Moreover, regarding the growing population of ‘lay’ programmers in the smart-living and IoT-ubiquitous programming environments of today (i.e., almost everyone in technologically ‘modern’ societies) both business companies and consumers would benefit from more insight into the non-conscious influence of culture and context on the programming choices that are made by the ‘novice’ programmer that has no formal training. In terms of education and learning, we argue that this insight could be used to help consumers become more aware of the cultural and contextual influences that shape their cognitive tendencies when they are programming.

### 1.2. Key Methodological Aspects

We employ several different methods to help us divide and detail these two general hypotheses. In short, here are the main aspects that are specific to our studies; details about our methods can be found throughout the paper and especially in Section 6.

A. We develop a cognitive task intended to reveal biases originating from the cultural background of a programmer, such as education, line of work, and free-time activities.
B. We investigate users with different programming skill levels, i.e., from professional to amateur.

C. We investigate a very general form of bias, but well hidden inside a programming task. This is because the programming task needs to be appropriate also for people with little or no programming skills.

D. We investigate whether inducing a bias is effective.

E. We investigate whether people educated in programming exhibit less biases and are less prone to manipulation.

Aspect A. To reveal the existence or nonexistence of cultural biases, we decided that the respondents in our studies should have different educational/professional backgrounds, i.e., from social and natural sciences, as well as from arts and cultural studies; see Section 6.1.

Aspect B. is motivated by the observation that increasingly more lay people (wrt. programming) are interacting and designing rather complex systems. Nowadays it is not only expert developers that program, but people with all levels of expertise carry out various programming-like tasks, from simple configurations of IoT systems in their smart home, to more complex configurations of technology systems in their work, to more unconventional forms of programming using visual languages and/or domain specific languages, and even assembling ready-programmed components into a final software system as done, e.g., in the IBM's IoT development environment. This is happening mainly because of the proliferation of simple (abstract, graphical, etc.) programming languages and interfaces aimed at non-programmers to design domain specific information systems, e.g.: a biologist programming a DNA search or an oil-engineer programming a complex database search. Therefore, in our study we use a simple programming task presented as fictitious, i.e., imagining to be programming. This allows us to perform our study both on programmers and non-programmers.

Aspect C. is important in order to avoid any intrinsic debiasing, and thus we must “hide” from the subjects the real goal of the study behind a seemingly unrelated goal, i.e., the “study of natural language in programming”. We must avoid giving any indications or hints to the study subjects about our intention to study their cognitive tendencies during the programming task.

Aspect D. is a standard study approach in research on biases because many types of biases can also be experimentally induced, sometimes termed priming Tulving & Schacter (1990); Yonelinas (2002). We thus also study the influence of contextual metaphors, in addition to the cultural bias, from A. Once we have established in this paper whether or not priming also works in the setting of programming biases, other works can carry out more detailed studies about whether such priming already exists “out there” (intentionally or not) and what types of priming would work and to what degree.

Aspect E. is meant to investigate a considerable opposition that our idea has encountered, i.e., that programmers cannot be biased when writing code. Moreover, we also aim to study whether it is possible to experimentally induce a bias on this category of users, or if this particular category is more resistant to priming and bias-transfer to programs other than user categories.

Our present work is motivated by the need to prove or disprove the idea that human biases could be transferable to the programming artifacts. However, which types of biases and how dangerous these might be are not the subject of this study. Other specific studies would have to be devised, maybe similar to the research on human biases developed in the psychology field.

1.3. Overview of the Paper

The contributions of this paper are presented schematically in the diagram from Figure 1. The vertical axis of the diagram represents types of biases (or degrees of influence) with the arrow indicating a direction from more general, long-term, and strong biases to more contextual, short-lived, and weak forms of biases. In this paper we study two extremes, namely biases coming from the background of a person (including socio-cultural, educational, and professional influences), and biases coming from manipulations such as priming. However, in between these two extremes one may study other forms of biases coming, e.g., from working cultures (think of programmers working for Google compared to a startup), or from media or propaganda. These biases form over a considerable period, e.g., several years, which is shorter than a life-time or childhood period, but longer than minutes as it is the case with priming.

Our results will be presented in the end of the paper, in Section 9, showing how these two types of biases are being transferred from humans to programs, providing evidence that transfer exists in both cases. In order to support such a new claim, the rest of the paper is devoted to the development of our studies and analysis of the data. Importantly, Section 8 contains the analysis from our control questions, which were meant to help us understand better whether the main elements (and assumptions) of our study were correct. The first important element of our study is presented in Section 3 where we develop a bias revealing test, which is used for both types of biases. This cognitive task allows respondents to answer only with one of the three rationales that are listed on the middle line of the diagram. In Section 8.1 we analyse how well our test worked, using one of our control questions. Section 4 details the manipulation method that we used, involving

---

priming participants with metaphors hidden in a fictitious ‘Philosopher’ story. How well these manipulations worked is studied again with a control question involving listing of ‘similar words’ in Section 8.2. We created three metaphors to match the three rationales, which in turn match with three kinds of educational lines (or views on life). This correspondence is reflected in the vertical alignment of these elements in the diagram. We have thus chosen our participants to represent these three different backgrounds, as detailed in Section 6. To test the assumptions about our participants’ backgrounds we used one control question (analysed in Section 8.3) asking them to rank the three ‘life-aspects’ listed on the bottom line of the diagram. The rest of the paper is devoted to presenting in Section 5 the major phases of (analyzing the data and demographics, in Section 7).

2. Cognitive Biases in Human Judgment

Numerous robust experiments in psychological science studying human judgment and decision making over nearly five decades have unanimously arrived at the now well-established fact that humans have a tendency to exhibit cognitive biases. Contrary to making an error, which represents a single incident in which one makes an incorrect judgment, a bias arises from a systematic tendency to commit the same type of error over time or in different situations. Thus, a cognitive bias can be understood as a sustained tendency to make an incorrect judgment. The tendency to exhibit a cognitive bias is particularly prominent in situations or contexts that are characterized by uncertainty, e.g., when processing information that are either too voluminous or too complex for the human brain to handle, or when forced to make a rapid judgment in a time-frame that is too short to review the information at hand, or when there is insufficient information for making the required decision, like in underspecified software requirements in programming. This last aspect, i.e., underspecification, is what we study empirically in this paper. Uncertain situations have a tendency to prompt the use of automatic and non-conscious cognitive processing. Whereas controlled and conscious cognitive processing, often termed Analytic thinking, System 2-thinking, or “slow thinking”, is the preferred cognitive mode for arriving at a correct judgment when the situation is in fact uncertain, this mode is not always easily attained. This is because the brain’s preferred cognitive mode is the automatic and non-conscious System 1, also termed Intuitive thinking or “fast thinking”. In situations characterized by certainty, “fast thinking” usually works well because we are on a “familiar terrain”. In uncertain situations, however, System 1 has a tendency to kick in and spark our automatic thinking mode even if it should not. But how does System 1 actually arrive at incorrect judgments? The reason is simply because our brains employ mental shortcuts, which are very useful in situations characterized by certainty. The problem is that our brain employs System 1 also in situations that are characterized by uncertainty, when it should in fact consciously activate System 2 Kahneman (2011).

Two key concepts in mental shortcuts, also termed heuristic processing, are mental accessibility (closely related to the availability heuristic) and mental representativeness Tversky & Kahneman (1974); Gilovich et al. (2002); Thaler & Sunstein (2009). When something is easily retrievable from short-term or long-term memory, we have a tendency to wrongfully regard it as something that is also occurring frequently, although it may not be occurring frequently at all. We may also make an incorrect judgment about an unfamiliar phenomenon by identifying superficial resemblances to a familiar phenomenon. Because it is cognitively effortful to identify substantial similarities between two phenomena, particularly in situations characterized by incomplete information and uncertainty, superficial similarities are more easily identified. In many instances this results in an incorrect judgment. To summarize, situations or contexts characterized by uncertainty in one way or another, prompts an inappropriate cognitive processing in System 1-mode, where the employment of mental shortcuts leads us to arrive at an incorrect judg-
ment, thus exhibiting a cognitive bias.

2.1. Algorithmic Bias: the Data or the Programmer

Media and the general public seem to have the assumption that machines and algorithms are in themselves neutral and objective. However, it has been known for quite some time (and recently also came to the attention of several public actors) that complex algorithms, such as those from artificial intelligence among others, may exhibit biases s.a.: racial bias Schlesinger et al. (2018), gender discrimination Zou & Schiebinger (2018) and other socially relevant types of biases Friedman & Nissenbaum (1996); Boyd & Crawford (2012); Jobin et al. (2019), when processing information in the support of human and institutional decision making Corbett-Davies et al. (2017); Dresdel & Farid (2018). This phenomenon is commonly labeled as machine bias or algorithmic bias, and has been confirmed in different areas (published in respective top venues), e.g., in big data Hajian et al. (2016), web Baeza-Yates (2016, 2018), autonomous systems Danks & London (2017). Among institutions that have raised concerns about the existence of “biased algorithms” are: the ACM US Public Policy Council²; the EU Parliament³; the New York City Council which passed a bill on “Accountability and transparency in algorithms for public agency support”⁴; ERCIM Rauber et al. (2019); World Wide Web Foundation⁵ and many more Cath et al. (2018). An influential research report Brundage et al. (2018) has raised even more concerns about harmful algorithms, and has recently been joined by articles in major publication venues such as Science and Nature Obermeyer et al. (2019); Zou & Schiebinger (2018); Gianfrancesco et al. (2018) and by scholarly books Boden (2008); O’Neil (2016).

All the works above focus on the data that AI algorithms train on, and show how the data contains biases. We are not aware of works that study empirically the transfer of biases from the human programmer to the algorithm, although we have found related ideas mentioned in recent articles: Silva & Kenney (2019) describe nine types of biases (present at five different algorithmic stages: input, algorithmic operations, output, users, and feedback), some of which can be studied in conjunction with the general bias transfer that we demonstrate in this paper (we consider it useful to demonstrate a type of bias through empirical studies as we carry out here). Baeza-Yates (2018) brings up the users and producers of the web content as sources of bias related to the data, but also points out different forms of bias coming from the user interface made by interaction designers whom could be regarded as ‘programmers’. We consider it particularly useful to study empirically the different forms of biases described by Baeza-Yates (2018) in the light of our hypothesis of ‘bias transfer’ and using methods similar to what we present in this paper, especially so since the author recognizes in the conclusion the same general sources of biases as we study here, i.e.: “each program probably encodes the cultural and cognitive biases of their creators”, and points in the introduction “measuring bias” as a major challenge, which is what we do here.

To state that algorithms are biased, or to assert that algorithms systematically produce an output that is biased, must as a consequence lead scholars to pose the question of whether a biased output could be proven and whether potential causal mechanisms leading to the bias could also be tested and studied.

Although an awareness of biases in algorithms has arisen, including the awareness of biases originating from data, no research programs seem to have undertaken the aim to study empirically the (cognitive) mechanisms that may lead to biases in algorithms, i.e., biases that do not originate from data itself, but from the programmers’ cultural backgrounds or from contextual influences in the immediate programming environment.

Thus, rather than pointing once more to the problem itself, the present study will instead provide explanations and insights, derived from our scientific study containing empirical evidence of actual human programming behaviour, into why this phenomenon may occur. We operate within the same paradigm and with the same agenda as those who study human behaviour, that is, we follow in the path of other multidisciplinary research themes such as Behavioral Economics Tversky & Kahneman (1974); Kahneman et al. (1991), Behavioral Transportation Research Pedersen et al. (2011); Gärting et al. (2014) and our own recent contributions termed Behavioral Artificial Intelligence Pedersen & Johansen (2019) and Behavioural Computer Science Pedersen et al. (2018).

As mentioned in the introduction, some may argue that robust quality assurance procedures eliminate any instances of biases in algorithms, at least in professional programming environments. We leave out for now testing whether the quality assurance procedures themselves have inherent biases or miss some forms of biases, and focus only on proving that programmers may be a source of biases, and not only the data given to the program.

Because heuristic thinking is seen as the main psychological “engine” for generating cognitive biases, our experiments will also employ a heuristic approach, that is, relying on mental shortcuts such as “accessibility/availability” when inducing a bias on the participants in the study.
3. A Bias Revealing Cognitive Task

Over the years, the first author has used a simple cognitive task Townsend (2003) (originally called Alice’s Alphabet Puzzle) in lectures on judgment and decision making. In Townsend’s book, this particular puzzle lists all the letters in the English alphabet (i.e., latin, roman) on a horizontal line, where straight lined letters are placed above and curved lined letters are placed below the line. We changed the puzzle in a vertical position to trigger more the infinity of the line, and the balance of the sides. The idea is to make students and professionals in various disciplines reflect on the possibility that facts that we “see” in the real world are generally a result of a “theory” of how we view the world. In the task, the audience is first shown, as in Figure 2, a correct sequence of the letters A to G – divided by a vertical line where A, E, and F is on the left side, and B, C, D, and G is on the right side – and then asked to decide on which side of the vertical line (left or right) the next letter H should be placed and why it should be there.

![Diagram of Alphabet Puzzle](diagram.png)

**Figure 2: The puzzle game.**

Although some want to place H on the left side of the line, whereas others want to place it on the right side, when asked why, they provide distinctly different rationales for their decision. Their rationales always seem to fall into three categories, which we categorized as: (I) ‘balance’, (II) ‘shapes’, or (III) ‘algorithm’. The quintessence of their arguments are as follows:

I. Some argue that there should be an equal number of letters on each side of the line: since there are already four letters on the right side and only three on the left side, the next letter, H, should go on the left side, thus indicating a sense of ‘balance’.

II. Others argue that the straight-lined letters A, E, and F are on the left side of the line, whereas the curved-shaped letters B, C, D, and G are on the right side, something which makes it perfectly reasonable that H should be together with its “kin” on the left side, given the different ‘shapes’ of the letters.

III. Yet others argue that there is an inherent order (or pattern) in the sequence: e.g., some indicate the sequence Left-1, Right-3, Left-2, Right-4, thus suggest placing H on the right side due to perceiving the image/puzzle as having the characteristics of an ‘algorithm/pattern’.

This exercise is simple enough to reveal cognitive tendencies of System 1, instead of triggering the System 2 analytical thinking, as usually done by more complicated tasks. More importantly, the bias cannot be avoided because there is nothing else in the picture to help them when making the decision, and any placement is correct; therefore only something from the background of the subject, alternatively, an experimentally induced prime, could help with making the decision. Moreover, since the subjects are asked “Why?”, they most often find an explanation in the memory related to the decision, or make up an explanation after the fact; though sometimes they answer that they choose at random.

4. Metaphors as Priming Method

We use the above task to force revealing a bias, albeit an innocent one (compared to racial or gender), which would have its origin in the cultural background of the person (e.g., education, line of work, hobbies). This will be used to test our first main hypothesis that we explained in the introduction, namely that cultural metaphors would influence the programmers’ choices. Besides that, in order to test our second hypothesis, we want to test whether we can prime the subjects to non-consciously make a decision in one specific “direction” – being one of the three rationales that we identified in Section 3.

Our motivation for hypothesizing that programmers would non-consciously be affected by the prime, comes from the well-known effect of the availability heuristic and the representativeness heuristic, together with the anchoring heuristic Gilovich et al. (2002). Under conditions of uncertainty, where one does not know, but nevertheless has to make a judgment or a choice, one will non-consciously base one’s judgments either on instances that spring easily to mind (i.e., the cultural background or the contextual prime triggers the availability heuristic), or on instances that resemble the current problem (the cultural background or the contextual prime triggers the representativeness heuristic). The judgment can also be anchored as an approximation to the most recent, the most related or the most relevant information; the context of the prime triggers the use of the anchoring heuristic. Thus, as regards cognitive processing, heuristic thinking in System 1 mode is very much an associative reasoning mode influenced by cognitive availability and perceived representativeness. However, one needs to also consider the content of the heuristic processing mode, e.g., what is actually easily accessible in memory. In terms of associative reason-
ing, the metaphors and metaphorical thinking are a strong source of influence on how we as humans view the world. The essence of a metaphor is, according to Lakoff & Johnson (2008), simply that “we understand and experience one kind of thing in terms of another”. For example, an argument may be understood and experienced in terms of the metaphor WAR, where we may “attack weak points in others’ arguments”, we may “shoot down” others’ arguments, and we may “win or lose” arguments. Similarly, we may perceive the concept of time in terms of the metaphor MONEY, where we may “waste each other’s time”, or “save time”, or even “budget our time”, and “borrow” time. In fact, metaphors are so pervasive and ubiquitous in our lives that we simply cannot do without them. Think for example of how the term “happy” is understood, experienced, and communicated in terms of UP, whereas “sad” is understood in terms of DOWN, showing that affective states, and even human health, consciousness, or control, are understood and experienced in terms of directionality. The good things are UP and the bad things are DOWN.

The concept of metaphor, “understanding one kind of thing in terms of another” as Lakoff and Johnson put it, particularly the way we perceive the world in terms of metaphors, is something that can even be manipulated. Consider for example Thibodeau & Boroditsky (2011) study of the effect of metaphor on the general public’s preferences for crime-prevention measures. When reporting crime-rates in a fictitious city, crime was either described in terms of “a beast” or in terms of “a virus”. When exposed to the metaphor crime is a beast, the general public argued for harsher and more severe crime-preventing measures than what was the case when they were exposed to the metaphor crime is a virus. Thus, not only are these “metaphors that we think with” Lakoff & Johnson (2008) something that we employ constantly in order to make sense of the world, but the way we use metaphors to experience the world is also something that can be manipulated. This can make us alter our view of the world, and most of the time we are not aware, neither of the fact that we think metaphorically, nor that our metaphorical thinking can be manipulated by governments, media, our employers, or others, either for commercial or political purposes.

The employment of metaphors, as well as the manipulation of our employment of metaphors, are processes that mostly go on outside our conscious awareness. Thus, since metaphorical thinking is so pervasively and ubiquitously present in our understanding, our experiencing, and our sense-making of the world, metaphors are exceptionally well suited for studying how the ultimate purpose of a computer program may be affected both by the programmer’s cultural metaphors, as well as whether one can influence the programmer’s initial understanding of the program’s purpose by a manipulation, i.e., inducing new metaphors that would alter the initial metaphors that the programmer employs.

In the case that the programming specifications are too sparse to completely fulfill the formal purpose of the computer program, how does the programmer ‘fill in the blanks’? We suggest that this is done by employing metaphorical thinking, either in terms of ingrained metaphors, i.e., cultural background, or in terms of cues in the proximate contextual environment that triggers metaphorical thinking, i.e., our manipulation of contextual metaphors.

4.1. Experimental Manipulation Using Metaphors

The experimental manipulation that we use consists of three experimental conditions in the form of ‘a story about a philosopher who invented a puzzle’ which we manipulated by varying the embedding of a different ‘life-aspect’, i.e., one specific metaphor, forming three different versions of the story. We also had one control condition, i.e., a story that did not contain any life-aspects/metaphor, intended for a comparison to the experimental groups. The three different life-aspects were:

A. HARMONY AND EQUALITY
B. AESTHETICS AND ARTS
C. ORDER AND CONTINUITY

The metaphor, which is given in detail below, included four words, placed in two groups of two words, one in the beginning of the story, and the other in the end, following indications from relevant literature Lakoff & Johnson (2008); Thibodeau & Boroditsky (2011). The words were:

A. harmony and balance; then equality and fairness
B. aesthetics and beauty; then forms of arts
C. order and structure; then linearity and continuity

We hypothesized that each of the above life-aspects would metaphorically influence the participants in the respective group A/B/C to provide an explanation that could be interpreted as one of the rationales from Section 3, respectively rationale I/II/III (i.e., ‘balance’, ‘shapes’, ‘algorithm’).

The metaphorical primes were embedded in the following fictional brief story about the philosopher who was presented as the one who originally created the puzzle from Section 3. Each subject will read a story that differs only in the words shown inside square brackets below, i.e., the first word pair alternative is provided to Group 1, the second pair to Group 2, and the third pair to Group 3.

“A philosopher who lived a life filled with [harmony and balance | aesthetics and beauty | order and continuity] created the riddle used in the game that we ask you to imagine that you program on the next page. Although the philosopher is nearly forgotten today, we know that the philosopher influenced many contemporary philosophers’ view of the world. The most prominent influence seems to have been the importance of maintaining [equality and fairness | forms of arts | linearity and continuity] in life and in society.”
We also had a control group in order to compare the effect of the different primes in each of the three experimental groups to a neutral condition without any metaphor-priming. In this control group, ‘ethics’ was embedded in the story, as this concept is unrelated to the three metaphors. The following story was given to the control group.

“A philosopher created the riddle used in the game that we ask you to imagine that you program on the next page. Although the philosopher is nearly forgotten today, we know that the philosopher influenced many contemporary philosophers’ view of the world. The most prominent influences seem to have been the importance of ethics in life and society.”

5. Designing the Study

We have spent considerable effort on designing our studies. Since the hypothesis that we want to test (i.e., that programmers can transfer their biases to the programming artefacts) seems to be quite hard to accept (at least we have seen this as the common reaction), we needed to build a case backed by strong, data-driven evidence.

First, we incorporate the two instruments that we described before, i.e., the bias revealing cognitive task and the priming metaphor, into a programming task. For now, we focus on a “paper-task”, where the subject imagines to be programming, so that we can easily involve non-programmers, since part of our hypothesis is that in the near future (if not already so) people with various backgrounds (outside computer science) would be involved in various “types of programming”. Examples that we are aware of include, e.g., “configuration” tasks as when managing a cloud environment with AWS or when using the IBM’s Watson IoT Platform for building IoT systems from software components plugged together on a graphical interface; or DSLs (domain specific languages) used in many disciplines that interact with software; or visual programming languages such as Fraunhofer’s IoT Programming Language NEPO or Google’s Blockly.

We then continue to describe in Section 5.2 the survey in which we have incorporated the programming task described in Section 5.1. We explain each of the survey’s questions and their purposes, as well as our rationale for “hiding well” the priming metaphor that is hypothesised to result in a bias. Besides the programming task, the survey contains additional questions meant to collect information intended for different purposes, from identifying ‘unserious subjects’, i.e., subjects that did not pay sufficient attention to the task, but instead responded more or less randomly, to information intended to help with the interpretation of the subjects’ explanations of their rationales or their background (see also Section 7).

We carried out our work in two stages. First we performed pilot studies, which we used to learn of flaws and to make improvements to the design. In this section we describe the steps and techniques that we used to arrive at our final studies. In particular, we first carried out specific usability testing in one pilot survey (described in Section 5.3). Then we improved the survey by using eye tracking technology to make sure that the priming is being read and to see more of how the users/subjects would interact with the survey (see Section 5.4).

In the second stage, three studies were done in full-scale with all the planned groups/cohorts from different cultural backgrounds: people with background in – or/and working in a field related to – social sciences (psychology), natural sciences (informatics), and visual arts & design, and performing arts (theatre and music). The results that we report in Section 6 are from these three full-scale studies. We adapted the questionnaire to the respective target group of subjects; particularly, after the first full-scale study, we made a few improvements that helped reduce the number of uninterpretable and nonsensical answers that needed to be discarded.

5.1. The Programming Task

We designed a fictitious programming task in which actual programming was not undertaken during the session, but where the focus was on the subject’s reasoning about the programming task. Thus, we informed the subject that she should place herself in the role of a programmer, i.e., imagine being the programmer to whom this task is given to.

The task was to program a game for children where the image from Figure 2 would be the game board. The game would consist of the player (which would be different from the subject/programmer) having to place the next letter H into one of the two designated empty boxes, i.e., on either left/right side of the vertical line. Upon correct placement, the game (i.e., the programmer) would reward the player. The boxes designated for placing the letter H are drawn with a dotted line with a large space between the dots. This design choice was made in order for the game to be perceived as continuing downwards and thus not blocked by a solid and continuous line/box. This was done to reduce the risk of being confounded by unintended biases (i.e., in this case, to avoid the subject to perceive the game board as being finite).

What we gave to the subjects as task description can be seen as the “requirements” that programmers receive from their clients (or elicited during a requirements engineering process); sometimes these include, so called, “user stories”, which are realistic descriptions of the functionality of the software in terms of how a user (in our case, a
player) would interact/work (in our case, play) with the software (in our case, the game). Our requirement contains one major intended omission (i.e., it is incomplete) in that it does not say what would constitute a “correct” placement of the letter H. In consequence, the subjects (as the programmer) need to decide for themselves to which side of the line they should give the reward (i.e., in other words, we wanted to elicit on which side they would prefer to place the letter H, themselves). As one can read further below, the subjects would be asked to explain their choice in a subsequent question of the survey.

We hypothesized that the uncertainty inherent in the task would elicit heuristic thinking prompted by either (i) cultural or (ii) contextual metaphors, in the same way people are influenced by metaphors in real-life situations. Specifically, the subjects’ cultural metaphors (from their cultural background) would influence the choices they would make in the programming task, and also that contextual metaphors (coming from our experimental manipulations described in Section 4) would have the potential to moderate their choices. We expect the choices to be one of the three different rationales from Section 3, i.e., the ‘balance’, ‘shapes’, or ‘algorithm’ rationale.

In order to introduce the priming metaphor, the game board image was linked to the story of the philosopher by saying that this “puzzle” was created by the philosopher. This link was made after the pilot testing (see details in the respective subsection below). We hypothesized that, if the participants were offered a simple explanation of the origins of the puzzle, then the philosopher story, containing one of the primes, would prompt the subject to non-consciously choose an explanation similar to the inherent rationale in the respective prime.

The task description that we used is the following.

“First, spend one minute imagining how you would be programming the simple task below. Then proceed to answer the following questions.

Imagine that you are a non-expert programmer who is developing a simple puzzle game. The game is based on a riddle made by the philosopher that you read about previously. Imagine that you have already drawn the game board that you can see below:

[The image in Figure 2]

Now you are going to program the player’s interaction with the game.

The player (NOT YOU, you are the programmer of the game) has to solve the puzzle by dragging-dropping the letter H on to one of the two dotted boxes. The player is rewarded if the program accepts the placement of the letter H as the correct placement.”

We want to conceal the bias revealing task behind the programming task in order to avoid debiasing. We want that the subjects focus on the programming aspect, and do not realize that we ask them to make a choice that has no wrong answer. This is relevant for what we discussed in 1.2.C.

5.2. The Survey

The survey is created in SurveyMonkey\textsuperscript{7}. The respondents are directed by a link to this online platform to complete the questionnaire. SurveyMonkey was also used for the respondents from mTurk. The survey was created bilingual, in both English and Norwegian; the Norwegian respondents had the possibility to choose between English and Norwegian (we considered the possibility of international students being enrolled in the studies). Screenshots of all the pages of the survey are given in Appendix A.

The survey starts with ‘Page 1: Introductory text’, presenting the goal of the survey and how the data is going to be dealt with. The goal of the experiment is only partially disclosed, and the true hypothesis remains completely undisclosed:

“There is now a growing number of programmers using programming languages that do not require programming skills or education. We are therefore conducting research regarding programming done by non-experts and we invite you to participate in a brief exercise about the use of natural language to explain the function of computer programs.”

We did not disclose the fact that we are studying biases. This nondisclosure was done in order to not influence the effect of priming and in order not to make subjects aware that we are interested in the potential relation between their cultural background and the choices they make in the programming task. We also inform the subjects about how long, on average, the completion of the survey should take and that the study is completely anonymous. Since the respondents have various backgrounds, from other disciplines than computer science, it was also important to mention that no prior knowledge of computer programming is required for taking part in the survey/task/exercise.

‘Page 2: Instructions’, contains information that we consider important for the respondents to know before starting the survey. There is also a reason for choosing to have such guidelines on a separate page, which we will expand on later in the Section 5.3. The respondents are informed that:

“The back button is disabled. You will not be able to go back to a previous question, so

\textsuperscript{7}Platform for creating online surveys: https://www.surveymonkey.com
we ask you to read each question carefully, because some depend on the previous ones.

Please put effort into reading carefully everything on each page."

Note that some text is being emphasized. In the case of skim-reading the minimum information the respondent perceives is: “back button disabled”, “read carefully”, and “put effort”.

We need the participants to actually read the texts in the survey for the primes to work and for understanding the requirements in the programming task. For the mTurk and SurveyMonkey respondents, who were paid, we also added information about required minimum time for completion.

‘Page 3: Philosopher story’. contains our story intended for priming, which we have detailed in Section 4. We experienced during the pilot test that the participants might not read a text if the information there cannot be used for answering questions in the survey. Therefore, we added one question meant as extra motivation for the respondents to read the text that contained the primes that were hypothesized to lead to a biased programming choice. (See Appendix A.)

‘Page 4: Programming task’. contains the text from the previous Section 5.1 (see also Appendix A). We also added three questions to this page, partly meant as extra motivation for the respondents to read the text with the programming task. One of these questions is important because it is directly related to the bias, i.e., it asks about the placement of the letter H. In order to conceal the importance of this question we added two more questions completely irrelevant for our experiment. However, all three questions are made to look like questions that concern the programming task, i.e., it makes the task realistic. If we would have left only the question about the choice of placement then the subject could have observed the missing information in the requirements that we gave and thus perceived the task as less realistic.

‘Page 5: Self explanation of choice’ and ‘Page 6: Alternative explanations’. is where the respondents give (or choose) an explanation for the choices they made in the programming task. We detail these two pages in the next subsections.

The rest of the questions on the following pages are meant to gather more information that could influence the results of the experiment, i.e., one’s view on life, hobbies, educational background, and demographics (age, gender).

‘Pages 7: Ranking life-aspects’. where the three alternatives from Section 4 could be ranked. We ask the subjects to

"Please rank the following three pairs of life-aspects in the way that best reflects how you view life yourself (where 1 is the highest while 3 is the lowest).
[ Options: harmony and equality | aesthetics and arts | order and continuity. ]"

This is a form of self-evaluation, where the subjects have the chance to express directly their order of preference for the three instances of priming metaphors (this is done after they have completed the main task, and they are not aware that they were themselves randomly exposed to one of the metaphors). If they rank the prime that they were exposed to highest, this might indicate that the prime has had an influence. If so, this would strengthen the results (i.e., the prime has influenced both programming and subsequent statements). If they do not rank “their own” prime highest, this could mean that the prime had an effect on the programming task, even if it did not have an effect on the participant’s statements.

The UI for ranking questions is made well by SurveyMonkey so that when the question is required, then the subject must indeed provide a ranking, and not just leave the default, since some action is required before the next button is enabled (e.g., either provide a ranking number or move the choices with the mouse).

‘Pages 8: Words suggestions’. where the subjects could suggest one to three words characterizing each of the three life-aspects, from the “Ranking life-aspects” question. The open-ended format chosen for this page has several reasons.

- We wanted to have a way to identify unserious subjects or robot-generated answers (as detailed later in Section 6.4).
- We also wanted to have another way to check the metaphorical priming effect by looking whether, and how often, our priming words appear among the answers.
- We also wanted to gather more data for future studies; i.e., others could use some of these words in future metaphor studies.

‘Pages 9: Demographics’. where the subjects had 5-7 questions about age, gender, years of education, field of study, and leisure activities (i.e., hobbies).

In the following section we explain the reasoning behind the way the questions are composed, as a result of the discoveries we have made during the pilot testing.

5.3. Pilot Testing of Usability

To test the usability of the survey we performed several pilot tests.

The first pilot test used the method of usability testing Dumas & Redish (1999), with our survey being the product under test. One goal with testing the survey for
The usability study involved five participants. Four of these participants have a background in computer science and one in arts & design. Subjects with these two types of backgrounds were going to be used in our full-scale experiments as well.

The participants were asked to take the survey while being observed by us, sitting next to them (one of us took the role of a moderator, while another researcher was only an observer). The test was run with one participant at the time. Before taking the survey, the subjects were explained verbally the purpose with the test session, which was to help us improve the face-value quality of the survey, although the hypothesis was not revealed. They were also presented the order of the tasks: first they were to take the survey, without any interruption, and then were supposed to answer questions meant to elicit suggestions for improvements of the survey.

After a participant completed the survey, s/he was asked to verbally answer a set of open-ended questions (post-test questionnaire/interview). The questions listed below represented the starting point for a discussion where the participants were free to give comments on the usability problems s/he encountered, while the moderator was recording their comments and asking supporting and clarifying questions.

- Were the questions/tasks easy or difficult to understand? Which of these?
- Did you find the instructions in the programming task clear or confusing?
- Was the text of the questions or explanatory text too long or too short? Which of these?
- Would you have needed/wanted to go back and read the previous question/texts? In which case? (we had the back button disabled so that they could not navigate back to the previous questions).
- How much time did you need to spend on reading the task about programming the puzzle? (they were asked to spend 30 seconds before answering the questions related to the programming task).
- Did you spend more time on certain questions than on others? Which of these? Why?

The participants were asked to recall what was experienced as difficult or unclear by retrospectively going through each page in the survey and reading again the content of the page.

The findings were marked with a severity level of high or medium. An example of a “high severity” is related to the ‘Ranking life-aspects’ question where the participants were asked to rank the three pairs of life-aspects. We wanted the participants to do the rating so that it reflects their own way of viewing life, but in the first pilot we did not emphasize this. By not specifying this, the subjects’ answers could potentially reflect their perceived view of others. In the case of one participant, s/he chose the answer that reflected the view of the philosopher presented in the story on the second page, as s/he surmised that this is what we might have wanted.

In general the findings from this first pilot test helped with shortening and simplifying the text and making the requirements/questions more clear.

The test also helped with validating our initial decision of disabling the back button in the survey. For the priming to work, the participants should not realize the connection between their choice in the programming task and the ‘Philosopher story’. If the participants would understand at a later stage in the survey that such a connection existed, they should not be allowed to navigate back and read the ‘Philosopher story’ again. In this pilot study one of the participants had the back button purposely left enabled. This participant did just what we expected, s/he navigated back to the ‘Philosopher story’, read it again, and adjusted her/his answers to reflect the view of the philosopher and not her/his own as the question required.

Another observation from this first pilot test was that on the ‘Philosopher story’ page the participants scrolled down quickly to the questions and then read only the part of the text that helped them to answer those questions. This was however difficult to establish with accuracy only through observation and the participants might not want to recognize that they did not put enough effort into reading the whole text. Only one of them acknowledged that. We concluded that the best way to reveal the exact behaviour that the participants have in reading the information, and also which flow they follow, was by using eye tracker technology Bojko (2013).

### 5.4. Eye Tracking for Better Insights

We created two versions of the ‘Philosopher story’ – a ‘Short story’ (Figure 3a) and a ‘Long story’ (Figure 3b). We employed summative research\(^8\) in combination with eye tracking methods for comparing the two versions and deciding which was more effective.

The test was done in a usability laboratory set up with eye tracking equipment. The survey was displayed to the participants on a desktop computer to which a remote eye tracker was attached. Using a remote type of eye-tracker,
required to calibrate the device before starting the testing with each of the participants, as well as instructing the participants to find a comfortable and stable position which to maintain during the whole testing, so to avoid head and especially body movement. To test the effectiveness of the two versions, a combination of single-subject and between-subject design was used, where each participant (ten in total) was exposed to only one of the test stimuli – five to the ’Short story’ and five to the ’Long story’. This was to avoid any carryover effects between the stories. For their help, the participants (students from Karlstad University) received a small reward in terms of a coupon usable in the student cafeteria.

Both stories contained the same priming words. The ’Long story’ was created with the purpose of helping the reader to immerse in the story – by giving more background information on the philosopher – and prepared the participant for the task on the next page of the survey, the ’Programming task’. Since we intended to prime the subject, we needed to hide the priming words well in the story, so that unintended debiasing (e.g., reactance) would not occur. At the same time, a too long story could make the subject not read the whole text and thus possibly skip the priming words. A shorter version of the story would also reduce the cognitive burden on the subject. The eye tracking testing was thus meant to help us identify whether the subjects skip our priming words, and also how much cognitive effort (i.e., how much time) they put into reading the stories.

The heatmaps and gaze plots visualizations\(^9\) provided both spatial and temporal insight into how the participants interacted with the text on each page of our survey. We obtained information about which areas of the text were fixated and for how long, the number of fixations and the order in which the fixations occurred. More specifically, we obtained insight into which lines and words were read, whether some of these were reread, and also how many times.

Interpreting this data we concluded that there was no noticeable difference in how the text, and especially the priming words, were read between the long and short version of the story. For both cases, the participants read the text thoroughly, line by line (Figure 4).

This shows that the instruction on the ’Philosopher story’ page about reading the story “CAREFULLY” had the wished effect. Thus, this requirement was carried over to the full-scale experiments as well. The difference between the reading of the two stories was that the long one required more time and effort, 1:10 minutes compared to 40

\(^9\)The gaze was offset vertically by approximately one line. This was due to the mismatch between the Operating System version and the version of Eye tracking software at the time of testing. The offset has been consistent across the participants and did not affect our interpretations.
seconds for the short one. Reducing cognitive load and time by almost 50 percent would be of help to the subjects, and thus we decided to use the ‘Short story’ in our full-scale studies.

Another aspect that we analysed with the help of eye tracking was whether the question about the philosopher being a man or a woman works as extra motivation for the participants to read the story. We found out that in order to answer this question, the participants returned to reading the story several times. In addition to the motivational aspect, questions such as this one will help in drawing potential attention away from the true hypotheses.

We also tracked the reading of the ‘Instructions’ text, as this is where we ask the participants to “put effort into reading” the texts in the survey. During the eye tracking phase, the instructions were part of the ‘Introductory text’ page, as a separate paragraph at the end of the page, separated by the rest of the text by a capitalized and emphasized header “INSTRUCTIONS” (Figure 5). In addition, we also included a sentence at the start of the ‘Philosopher story’ asking the respondents: “Please read CAREFULLY the story below”, again with capitalized and emphasized letters (Figure 3).

The findings from the recordings revealed that one participant did not read the instructions text, but only looked at the words “instructions” and “put effort”. These words were standing out through their typography design. This participant was found to not put effort in several other places. Three of the participants were only reading parts of the instructions, while the remaining six were reading carefully, line by line. In a laboratory context the participants are known to be putting in more effort than they might do on their own in other more natural contexts.

By analysing the behaviour exhibited by the participants in reading the instructions on both pages, we concluded that we should create a dedicated page for the instructions, just before the ‘Philosopher story’ where we should include the “put effort into reading” text. Thus, we drew attention to the importance of the instructions by creating a separate page for them. Here we also included the “… put effort into reading carefully …” text, in order to avoid repeating it on the ‘Philosopher story’ page. This was done with the intention of preventing the possibility of the participants guessing our priming intention by marking the story as something to pay special attention to. We also decided to typographically emphasize the important words or sentences used for instructions throughout the whole survey, as we have seen that they are always read by the participants.

6. Performing the Study

6.1. The Participants

Here, we describe the reasons behind the way we chose and grouped our subjects, and how the different cohorts of participants are meant to help study the research questions that we elaborated on in the Introduction.
The participants were chosen based on their educational or occupational background, to span three main educational and professional domains, which is relevant for the 1.2.A aspect presented in the Introduction. This is also meant to cover well different computer programming skill levels as well as socio-cultural influences, properties and preferences. We reason that, when enrolled in a certain study line or field of work, people have already developed predominant skills and characteristics needed for the specific education or occupation.

We will thus talk about three main cohorts of respondents:

A. ‘Social sciences’ cohort – composed of students studying psychology;
B. ‘Natural sciences’ cohort – composed of students studying computer science; and
C. ‘Arts and Culture’ cohort – composed of a group of participants working in the field of arts and design, a group of students studying theatre, and another group studying music. The last two subgroups we term the ‘Cultural studies’ participants.

This categorization based on the educational and professional background of the respondents is confirmed by the analysis of the data obtained from the control question on the ‘Demographics’ page, i.e., specifically about which field of study or/and line of work the respondents affirm their background to be mainly consistent with.

Based on the conditions to which the respondents were exposed, we can also categorise the three cohorts into:

I. ‘not helped’ and ‘not confined’,
II. ‘helped’ and ‘not confined’,
III. ‘helped’ and ‘confined’.

The ‘confined’ or ‘not confined’ categorization describes the environment where the respondents were at the time of taking the survey. The ‘not confined’ respondents took the survey in the environment of their choice, which was unknown to us, whereas ‘confined’ means taking the survey in a more controlled environment (i.e., the university auditorium). The ‘helped’ or ‘not helped’ classification refers to whether the respondents were supported or not by being given alternative explanations to pick from, in order to explain their choice for the placement of the letter ‘H’. Starting with the second full-scale study we introduced an extra page in the survey, offering such alternative explanations with possible answers to choose from, meant to reduce the number of uninterpretable answers. The respondents that were given the ‘Alternative explanations’ to help them with explaining their choice are termed as ‘helped’. The ‘Social sciences’ cohort belongs to the ‘non helped’ and ‘not confined’ category (I), as the survey they were given did not contain the ‘Alternative explanations’ page and they could take the survey at the time and place of their choice. In the case of the ‘Arts and Culture’ cohort we had a combination of ‘helped/confined’ and ‘helped/not

Figure 5: Image of one of the recordings, watched in half of the real speed, showing that the instructions were carefully read, with an even distribution of fixations.
confined’ (II, III) with a predominance of the latter. The mTurk and SurveyMonkey respondents were helped with ‘Alternative explanations’ and were free to choose the environment where to take the survey. The ‘Cultural studies’ students were also ‘helped’ but confined to a classroom, where the course leader and one of the authors were also present. Similarly, the ‘Natural sciences’ cohort was both ‘helped’ and ‘confined’ (III) as the survey was taken as part of their regular course-work. The ‘confined’ / ‘not confined’ and ‘helped’ / ‘not helped’ are categories that are going to be used for analysing the sensical vs. nonsensical data in Section 7.1.

The environment of the participants and the support they received is related to three main types of environment where (future) programming activities can take place in:

A. Typical professional programming environment, where the programmer is ‘confined’ to an office space and has to her disposal all the professional resources necessary to fulfill her tasks. In our case, for the programming task and the required explanations, we tried to reproduce this type of environment for the group of computer science students, by both confining them to the classroom and course hours and offering them helping answers.

B. Semi-professional environment, where an expert in some technical field (other than programming, e.g., railway engineering) has professional tool support for simple programming/configuration, e.g., by using a GUI based programming tool or a graphical programming language s.a. Blockly. However, programming is not their main task or responsibility and thus are not supposed to put too much effort into it, which we consider as ‘not confined’. The mTurk and SurveyMonkey respondents were thus ‘not confined’ but ‘helped’.

C. Non-professional environment, where people, e.g. in their homes, are configuring an IoT system without any professional support nor prior knowledge. The ‘Social sciences’ respondents were neither ‘helped’ nor ‘confined’, and can thus be seen to some extent as fitting this profile.

For the purpose of studying the influence of the contextual-metaphor priming, we further group the respondents from each cohort by the priming they have been exposed to (or not), according to Section 4.1:

A. a control group which is not primed in any way,
B. a group primed as in Section 4.1.A (which we call, ‘primed with harmony and equality’),
C. a group primed as in Section 4.1.B (i.e., with ‘aesthetics and arts’), and
D. a group primed as in Section 4.1.C (i.e., with ‘order and continuity’).

The control group is meant to serve as a baseline to observe what the programmers’ preferences for task-solutions are in the absence of primes, i.e., when presented with a description that is neutral with regard to the task at hand and the task’s inherent possible solutions. This is relevant for our first main question.

The three primed groups are meant to help us test whether the bias can be induced upon the programmer, and subsequently transferred from the programmer to the algorithms. This is relevant for our second main research question presented in the Introduction and the aspects discussed in 1.2.D.

The cohort with students from the computer science study line is meant to help us test whether programmers shut away the other two biases, i.e., resulting from the cultural metaphor or the induced contextual metaphor, except the pattern/infinite way of thinking, which is sometimes assumed that the programmers do. This is relevant for what we discussed in 1.2.B and 1.2.E.

We also collect information about the hobbies, years of higher-education, and gender of the participants. The hobbies and years of higher-education are meant to help with fine-graining the analysis of the educational and occupational background of the participants. The age data is presented in the Section 7, with the purpose of giving an overview of the population distribution. We also look for indications of whether the bias is more pronounced in different age categories, and if this is reflected in how much it is transferred into the program.

6.2. Methods Employed

The main research hypothesis to be tested is whether biases can be transferred from the programmer to the program. This hypothesis is tested throughout all our three experimental cohort-studies, and it contributes to answering our first main question from the Introduction.

The studies employ a combination of experimental design and comparative design. In the analyses of both (i) the comparative aspect (i.e., differences between the three cohorts) and (ii) the experimental aspect (i.e., differences within each cohort, resulting from the experimental manipulation), we employed both (a) inferential statistics, more specifically chi-square analyses of categorical data, as well as (b) descriptive statistics to report frequencies and percentages. We performed an experiment on each cohort, as well as compared the three cohorts to each other, regardless of the experimental manipulation. Since the three cohorts were different in terms of cultural and educational background, we were able to study the unique effect of background per se.

Conforming to the true experimental design method Lazar et al. (2017); Cook et al. (2002), we first assigned the participants of each cohort randomly to one of three experimental conditions where we induced one specific type of contextual metaphorical thinking in each, or to a control condition containing neither of the three primes. The control condition contained the neutral non-prime story from the end of Section 4.1 and was meant to serve as a “baseline” to establish whether the participants, without being
primed, were inclined to favor one of the three “rationales” over the other.

The subjects are given the programming exercise described in Section 5.1. The programming task, the educational/professional background of the subjects, and the story containing the primes, are the independent variables in our experiment. The choice of what will be the right solution for the puzzle is the dependent variable. We are interested in finding out if the primes and the background of the participants (the independent variables) induce changes in how the puzzle is programmed (the dependent variable), following the rationale that it is the programmer who decides to give the player a prize based on what she (i.e., the programmer) thinks qualifies as the right answer.

The conditions (also known as treatments) that we intend to compare are reflected in the explanations that the subjects provide, being under the influence of three contextual metaphor primes and three types of cultural background.

A true experiment requires randomization and controlled trials (Randomized Controlled Trial – RCT), as well as one or more distinct experimental manipulations. First, our study conforms to these requirements due to random assignment of each participant to the four conditions (i.e., three experimental conditions and one control condition). Second, the experimental conditions are controlled and kept constant to the extent that we recorded the time spent on the tasks and thus ensured that the tasks were completed within a reasonable time-frame. Thus, we excluded the effect of any seriously potentially confounding variables, such as diffusion of experimental manipulations (i.e., we reduced the possibility of participants sharing the contents of the tasks with other participants). Participants completed the task individually and received identical instructions. In addition, the hypotheses were not revealed to the participants. Such non-disclosure of hypotheses is the most robust experimental procedure, and it is employed in around 87 percent of all experimental-psychology research Hertwig & Ortmann (2008) because it allows for the elicitation of valid measures of behaviour instead of relying on less valid measures by means of other methods, s.a. self-reports Bröder (1998); Christensen (1988); Kimmel (1998); Kron (1998); Trice (1986); Weiss (2001).

For analysing the second main hypothesis that we proposed in the Introduction, pertaining to the potential influence of the context, i.e., metaphors in the programming environment, the research hypothesis is that the manipulation (“prime”) will increase the number of the corresponding explanations the participants give. In the event that the number of explanations does not increase as a result of the prime, we discard the research hypothesis and retain the null hypothesis.

Supporting evidence of whether the prime has induced a specific metaphorical thinking and thus has produced a biased judgment in the prime’s direction should be shown in the participants’ explanations, given after they have finished the programming task. The participants’ explanations for their respective choices were qualitatively coded according to the three predefined categories. Explanations conforming to one of the three predefined categories were categorized both according to their discrete category (i.e., ‘balance’, ‘shapes’ or ‘algorithm’) as well as whether they were ‘sensical’ (i.e., eligible for inclusion in the predefined categories) or ‘nonsensical’. Non-interpretable explanations were thus labeled ‘nonsensical’ and discarded (see Section 7.1 for a thorough analysis of this). We coded the explanations qualitatively and categorized them into one of the three possible rationales. If the rationale of prime-manipulation in the respective condition is chosen significantly more than the other rationales, this would imply that the participants were influenced by external features that are not relevant to the programming task itself.

We implemented one additional variable to control for the bias, and this stemmed from our assumption, resulting from an observation in our practical use of the cognitive task from Section 3, that the choice of placing the letter H is also an indication of the rational. Particularly, participants choosing Left would be those using the rationales I, II from Section 3 for ‘balance’ or ‘shapes’ (similarly those primed with the life-aspects A, B from Section 4); whereas participants choosing Right would be those using the rationale III for ‘algorithm’ (similarly those primed with C).

This is analyzed in Subsection 8.1.

We choose the subjects based on their educational and professional background. However, in the questionnaire we ask the participants to provide information about their educational background themselves (because some have multiple) as well as information about their preferred free-time activities. This is done in order to disclose a possible relation between this particular aspect of the background of the participants and their choices in the programming task. Moreover, this information from free-text questions can also help detect respondents that did not relate seriously to the task, as well as to control our qualitative coding of their explanations and background.

We hypothesize that the results from the analyses will show a statistically significant relation between (i) cultural metaphors (i.e., the subjects’ cultural background) or (ii) contextual metaphors (i.e., the experimental prime) and the choices made (and the explanations provided) regarding the programming task. The majority of the answers are expected to fall in one of these three categories: biased by the prime only, biased by the background only, or biased by both. If neither of this is the case, our hypotheses are discarded, and we conclude that biases are not being transferred from the programmer to the program.

6.2.1. Alternative Explanatory Variables

The age and gender of the participants are analysed as alternative explanatory variables.

Other alternative explanatory variables that might occur could result from the subjects not understanding the task well, the task being too difficult, or the prime not being strong enough as a result of superficial reading. How-
ever, these factors were something that we detected and removed through our pilot tests.

The nature of our study requires an (i) experimental and (ii) comparative between-group design, where each participant in the experimental part is only subjected to one experimental condition. In the comparative part of the design, the three cohorts represent three distinctly different cultures. Thus we were able to study the potential effect of both the cultural metaphors and the contextual metaphors on programming choices as two different sources of bias.

In order to effectively exclude the potentially confusing noise caused by the individual differences that may occur in small samples, as well as potentially nonsensical explanations that would need to be discarded, we decided to use a sufficient number of participants in each condition. Thus, we assigned at least ten participants from each cohort to each condition in order to arrive at least at five sensical output-explanations and thus conform to the conventional requirements of chi-square comparisons, i.e., including at least five participants in each group.

We believe, however, that the individual differences will not affect the study results to a large extent. By implementing the questionnaire questions related to individual preferences and extracurricular activities (‘Ranking life-aspects’ and ‘Words suggestions’ pages, and the question about hobbies on the ‘Demographics’ page), we expect to be able to clearly identify if the choice was dictated by the bias. Moreover, the programming task is thought to be very simple, thus requiring very little cognitive effort. For such cases, it is empirically proven that the individual differences have a smaller impact Lazar et al. (2017). For the same reason, i.e., reducing the factors influencing the main conditions to be compared in the experiment, the groups of respondents were composed of respondents studying/working in the same field, something which reduces potentially confounding individual differences at least to some extent.

6.2.2. Ethical Aspects

All three full-scale studies were done with the responses being anonymized. All questions are registered on the random ID that the system generates. It is thus not possible to identify any of the respondents.

6.3. Learning from the First Full-scale Study

The first full-scale study, also referred to as the ‘Social sciences’ cohort, consists of undergraduate students enrolled in a psychology study program. The link to the survey on SurveyMonkey was sent through email by the study program administrators and resulted in 77 responses. Observations made after the first study helped with improving the following studies. An analysis of the incomplete responses (31 out of 77) from the first study shows that: 6 dropped out right away; 12 dropped out on the ‘Programming task’ page; 6 on the ‘Self explanation of choice’ page; 5 on the page with ‘Words suggestions’ page; and 2 on the ‘Demographics’ page. The dropout rates are respectively 19.4%, 38.7%, 19.4%, 16.1%, and 6.5% (i.e., Figure 6).

![Figure 6: Overview of the dropout number – frequencies and percentages – of respondents from the ‘Social sciences’ cohort; including the names of the pages where the dropout happened.]

The high number of dropouts on the ‘Programming task’ page could be explained by the fact that the students in this cohort are studying psychology and they may have never been exposed to computer programming, thus they may have deemed this task as not relevant, not interesting, or maybe too difficult. Based on this reasoning, in the second study, also referred to as the ‘Natural sciences’ cohort, we introduced on the first page more text where we explain that “It is not required to have any prior knowledge of computer programming.”; moreover, on the ‘Programming task’ page we wrote that the puzzle is simple, i.e., ”... who is developing a simple puzzle game”.

Another solution for further motivating the respondents to finish the survey was to add a progress bar indicating how much of the survey was left until completion. For the last three pages we also added a page-footer informing, consecutively, that ‘there are three, two, and one pages left’, i.e., aiming to reduce the two latter types of dropouts.

Recall that the first study was conducted with volunteer social science students that were neither paid nor participating during their normal class hours. In contrast, the second study was run in a lecture hall, before the break, as part of a first year computer science course. In the case of mTurk and SurveyMonkey respondents from the third study, also referred to as the ‘Arts and Culture’ cohort (see Section 6.4), we consider the payment as an important motivating factor.

To reduce the cognitive effort required for explaining the choice for the ‘H’ letter placement on the ‘Self explanation of choice’ page we added, immediately after this page, an additional page called ‘Alternative explanations’, containing a list of predefined explanations to choose from. This was meant to reduce the high dropout rate that we saw on the ‘Self explanation of choice’ page. Moreover, adding these alternatives in the second study reduced the number of uninterpretable answers significantly. A detailed analysis of the uninterpretable answers (i.e., explanations) is done in the Section 7.

For the second study, the total number of responses received was 53. Of this total number, one respondent dropped out on the ‘Programming task’, one on the ‘Self
explanation of choice’, one on the ‘Ranking life-aspects’, and two on the ‘Words suggestions’ page. In total there were five dropouts. The small number of dropouts in this second study indicates that the adjustments made after the first study were successful.

Regarding the language in the first study, the students could choose between English and Norwegian. The majority of participants in the first cohort were Norwegian speakers. Nevertheless, some of the students could be, for example, exchange students and thus more comfortable with English than Norwegian, therefore an English version was made available as well. Though we did not have any students choosing the English version in this case, we encountered three such cases in the second study, testifying to the relevance of providing both a Norwegian and an English version. Moreover, the English version of the survey was necessary for the mTurk and SurveyMonkey respondents in the third study.

6.4. Transitioning from Volunteering Students to Professional Respondents

For the third cohort, we planned to recruit people with a background in arts and culture in general. We intended to continue with the same style of set-up as in the second study, where the respondents take the survey during their course, confined to a classroom. It proved difficult to find a large number of respondents to comply with these requirements. We started the third study with two groups of students, studying music and theatre. Though we had no dropouts from these groups, the numbers of students in the classes were too small (which is specific to these kinds of studies), i.e., 10 respondents from music and 10 from theatre. To increase the number of responses we decided to recruit respondents through specialized platforms, specifically through the Amazon Mechanical Turk and SurveyMonkey. These would no longer be volunteers but professional respondents, i.e., who are paid for their participation and do such tasks often.

The data from Amazon Mechanical Turk (mTurk) comes from three batches. The first batch of respondents were chosen based on the following qualification requirements: working in the Arts & Design field and having a Masters qualification. The respondents in this first batch were rewarded $2 plus fees $1.35 (Mechanical Turk Fee: $0.80, Masters Fee: $0.10, US High School Graduate Fee: $0.05, Job Function – Arts & Design: $0.40). The reward sum was decided based on what other requesters on mTurk were paying, which was generally under $1. We received only one response. Therefore, we increased the reward to $3 (plus appropriate fees) and reopened the mTurk HIT. However, this second batch brought only three additional responses. The number of respondents with the wished qualifications proved to be small, irrespective of the remuneration. In a third batch we kept only the job function – Arts & Design – requirement, and the same payment of $3. This generated 115 responses, which was close to the number of respondents we were aiming for (i.e., 100 respondents).

The respondents in this batch come from the mTurk. However, given that the questionnaire was already made in SurveyMonkey, we directed the mTurk respondents to SurveyMonkey through a link. There is a mismatch between the number of respondents registered in mTurk (115 plus four from the first two batches) and SurveyMonkey (128). This is due to the fact that we rejected (and not paid) some of the respondents, and thus their places were reopened in mTurk for new respondents, but SurveyMonkey kept generating new IDs, recording also those that we have rejected in mTurk.

We asked the respondents (i.e., in the “Instructions” section of the survey) to spend at least eight minutes on their response and read all text with attention. However, quick readers could have managed to complete the survey in no less than four minutes. The respondents that spent less than four minutes could not have spent the necessary time on reading the texts in the survey and their answers cannot be considered reliable. This was established also by checking the answers in the open fields, where we could see how much effort the respondent put into writing an explanation on the ‘Self explanation of choice’ page and giving examples of words on the ‘Words suggestions’ page.

Thus, from the total of 128 responses registered in SurveyMonkey, collected from all three batches, we removed 13 respondents that spent less than four minutes on completing the survey. Additionally, seven more respondents were rejected as we deemed them unserious (e.g., computer generated answers). Out of the remaining 108 responses, one participant dropped out on the ‘Self explanation of choice’ page.

Moreover, six respondents that spent more than four minutes were still deemed unreliable and thus removed from the analysis. This was decided based on the quality of the responses given in the open-ended questions. Some examples of such ‘unserious’ answers are:

- One respondent (M:11270629093, from the third cohort) spent 18:44 minutes, but his/her answers on the ‘Self explanation of choice’ page was “none’, whereas for the ‘Words suggestions’ question the answers seemed to be generated by a computer: answer “the intimately transmitted from west to east” for life-aspect ‘harmony and equality’ (from Sec. 4.1.A); answer “Aesthetics is branch of philosophy that examines the nature” for life-aspect ‘aesthetics and arts’ (from Sec. 4.1.B); answer “the Alternatively,

---

10 Only one incomplete answer was removed, which was the response of the course tutor that wanted to scrutinize our survey. https://www.mturk.com/
11 A Master Worker is a top Worker of the mTurk marketplace that has been granted the Mechanical Turk Masters Qualification. These Workers have consistently demonstrated a high degree of success in performing a wide range of HITs across a large number of Requesters.
this is called first-order parametric continuity” for life-aspect ‘order and continuity’ (from Sec. 4.1.C).
• Another respondent (M:11270932685, third cohort) spent 11:00 minutes, but did not give any explanation on the ‘Self explanation of choice’ page, and for the ‘Words suggestions’ question wrote “just my opinion”, and for the Field of study/line of work question on the ‘Demographics’ page answered with “Letters of the words and Line”.

To the mTurk batch we also added the responses from the study run with SurveyMonkey respondents and with the respondents from the Culture studies.

In the case of the SurveyMonkey, we requested 50 respondents to be working in the field of Arts and Design. The country of origin of these participants was chosen to be Sweden. With the purpose of reaching the number of respondents having the required qualifications that we asked for, SurveyMonkey created two collectors:

(a) The first collector generated 39 qualified responses renumerated with SEK 8,982.8 in total, and SEK 230.32 per response.
(b) The second collector generated 11 qualified responses renumerated with NOK 2,332.77 in total, and NOK 212.07 per response.

Out of these 50 respondents we removed 15 that spent under 4 minutes on completing the survey. Out of the remaining 35 respondents we further removed 13 unserious answers. These answers were deemed as unserious based on the evaluation of the answers in the open-field questions, e.g.: the participant with the ID: S:11174871215 (from the ‘Arts and Culture’ cohort) answered “a e f fits my eye” on the ‘Self explanation of choice’ page and “now / now / nos”, “1 / 3 / 2”, “People love us / Color field tree / Lamp tide dog” on the ‘Words suggestions’ page. Of the remaining 22, two of the answers were incomplete, with one respondent that dropped out on the ‘Ranking life-aspects’ page and another on the ‘Self explanation of choice’ page.

7. Analysing the Data

All the examples of answers from participants are presented by us here in English, but many of them are translations from Norwegian (including grammar corrections; though not for the English ones, which are kept verbatim, including their grammatical errors).

7.1. Sensical vs. Nonsensical Answers

The participants’ explanations (i.e., their written texts) were analyzed qualitatively and coded into one of the three rationales from Section 3 (i.e., Section 3.1 ‘balance,’ Section 3.1I ‘shapes’, or Section 3.1II ‘algorithm’) depending on the fit between the text and the category. During the first full-scale study we found one answer (pID 38, first cohort) which triggered us to introduce another category or rationales, called ‘sounds’: the answer explained the choice of letter placement as “If you sing the alphabet in Norwegian then the best fit with the rhythm is to place ‘H’ to the left, because you have a small pause before singing ‘H’ after ‘G’.”.

There were still many answers that could not be categorized into the above, either because they did not make much sense, or the reason given was no reason at all. However, many of these answers were recurrent, transcending even the language differences, and this allowed us to group them in categories. Some of the more generic answers were so similar between English and Norwegian that we could regard them as ‘universal’.

• ‘Logical’: “I think it would be logical put the H in the right position” (pID M:11272137574, third cohort); or “Left seems right because it seems logical” (pID 53, first cohort); or “because it seemed most logical” (pID 12, first cohort).

• ‘Pattern’: “My choice was made by what I thought was a pattern” (pID M:11282013578, third cohort); or “because of the order of the previous ones.” (pID 47, first cohort); or “Due to previous placements above.” (pID 50, first cohort); or “The left seems to follow the pattern” (pID M:11270235127, third cohort); or “Because I think the pattern follows that path.” (pID 4, first cohort).

• ‘Random’: “Just chose something” (pID 33, first cohort); or “It seemed like the pattern of the letters would place the H on the right, but there isn’t enough information for me to decide, so it is kind of a guess.” (pID M:11270119183, third cohort).

• ‘Alphabet’: “…going in reverse alphabetical order.” (pID M:11272389655, third cohort); or “The letters are to be placed based on the alphabet song.” (pID M:11271323609, third cohort).

• ‘Handed’: “I’m right handed so I favor my right side and it just seemed like the ‘correct’ answer to me.” (pID M:11271930008, third cohort); or “Most people are right-handed, so dragging the letter to the right felt like an automatic default action. Dragging it to the left requires a more deliberate choice.” (pID M:11270365264, third cohort); or “I chose the right because in every day society its pretty common for the right side of thing to be accepted as good, such as right handed people, the right hand of god, etc etc. I also chose the right side because its ‘right’.” (pID M:11270469031, third cohort).

• ‘No reason’: “Because it looked most natural compared to what has already been done.” (pID 36, first cohort); or “it looked natural” (pID 7, first cohort);
or “It felt right” (pID S:11178992036, third cohort); or “Seems better” (pID S:11174871629, third cohort).

- ‘micro-balance’: “H on the right side follows the pattern of the EF on the left side, which are a pair.” (pID M:11272137574, third cohort); or “Because it makes sense to me that H and G are grouped together, since there is a grouping on the other side as well.” (pID T:11058678726, third cohort); or “In my opinion it looks nicer to have ‘H’ after ‘G’. It has a bit to do with how ‘E’ and ‘F’ are positioned.” (pID T:11058678669, third cohort).

To reduce the number of uninterpretable answers (i.e., explanations, rationales), starting with the second full-scale survey, we introduced the alternative answers that respondents could choose from; see Figure A.23 in the Appendix A. These alternatives were formulated based on the wordings that we encountered among the responses from the first study. Thus, the first study provided a type of ‘saturation’ of alternatives. As a result of coding and categorization we arrived at five alternative answers, as well as a sixth and seventh alternative: “I just chose something” and “I already gave an explanation”. We also used these to help us code the answers, i.e. when they did not give any explanation (it was not required) but instead chose from our list, we used that choice as the rationale. When they gave an explanation that did not make sense, but then also chose one of our example explanations, we again used the one that they chose, for our categorization.

There was also the case when their explanation somewhat seemed to contradict the choice that they picked. In this case, we still used the choice for the categorization. The following are a few explanations that made no sense, but an alternative was chosen: “The left side seems like the logical, correct side when compared with the letters that came before it.” (pID M:11270382691, third cohort) but then chose the alternative answer that sounded “Because of the appearance/form of the letters. On the left side they have straight lines, whereas on the right side are rounded.”; or “I choose left because I think it can be very good with random letters in the left.” (pID M:11270101280, third cohort) but then chose the alternative “The same number of letters on each side.”; or “There seems to be a pattern. Placing the letter on the left makes the most sense to continue that pattern.” (pID M:11271499180, third cohort) but then chose a pattern from the alternative that sounded “It creates a pattern of the type: 1-3, 2-4, 3-5, ... or 1-3-2, 1-3-2, ... or 1-3-2, 2-3-1, ...”.

We thus define as Sensical those answers that were interpretable and allowed for category inclusion in one of the three rationales from Section 3, and we define as Nonsensical the remainder of the answers, both those that made no sense at all, as well as those that could be coded as ‘sounds’ which were very few in number (i.e., one in the ‘Social sciences’ cohort, one in the ‘Natural sciences’ cohort, and five in the ‘Arts and Culture’ cohort, all of which did not involve our helper alternatives, but their explanations that described the reason as ‘sound’).

In the following we make two observations about our sensical vs. nonsensical perspective on the responses.

7.1.1. Helping with the Self-explanations

The first regards the level of help that the different cohorts received. We observe in the Figure 7 a significant increase of answers that allowed for interpretation when the respondents were offered the alternative explanation choices. The ‘Social sciences’ cohort were not helped and the percentage of sensical responses is only 58.49%. To all other respondents we allowed them to skip the ‘Self explanation of choice’ question and required that they at least chose one of the alternative explanations. The sensical answers increased significantly to 70.50% and 84.31% for the ‘Arts & Culture’ and the ‘Natural sciences’ cohorts, respectively. It is particularly noteworthy the increased level of interpretability that this choice in the design of our studies brought. We have counted 22 answers given by the participants in the ‘Arts and Culture’ cohort that were not understandable only by themselves, but could nevertheless be coded because of their choice of alternative explanation. This would have otherwise tilted the percentage to only 54% sensical answers. We also had 10 that chose to skip the self-explanation and only select one of the alternatives.

![Figure 7: The nonsensical answers are decreasing after improvements done to the survey after the first full-scale study.](image-url)

7.1.2. Programming Environment Confinement

The second observation regards the confinement as aspect that we described in Section 6.1. One can observe that as soon as the participants were confined their explanations became even more sensical. Here we look at the two cohorts that were helped (i.e., to whom alternative explanations were offered), and we notice the in-
crease from 70.50% to 84.31% in the case of the respondents from the ‘Natural sciences’ cohort who were confined to the classroom and course working-hours. This bears evidence that the transition from a non-professional towards a more professional programming environment (as explained in Section 6.1, where people are both ‘forced’ to focus appropriately on their task as well as being helped by resources or tools) would trigger the programmers to be more careful about their choices. This could also contribute to lowering the amount of bias. Indeed, we have observed that several participants tried to think in terms of games, since the task consisted of programming a game. Examples of such explanations are: “I choose left because it’s a game and I think according to the pattern gamer will choose right side psychologically. Thus he/she will lose.” (pID M:11274822275, third cohort); or “I feel the right side would be the most common choice so if the player was thinking creatively they would choose the left side to place the h” (pID M:11274883590, third cohort).

Another aspect of the confinement is that it triggers the System 2 thinking, which is known to result in a reduction of human biases. We have also observed instances of System 1 vs. System 2 thinking in our participants, i.e., ‘starting’ as a System 1 response, but then ‘self-apprehended’ and activated a System 2 response. One example is a participant who has chosen to answer the letter placement question with ‘Right’ but then when asked to explain the choice said “I chose right previously but actually left makes more sense. Balancing the sides; 4 letters on the left, 4 letters on the right.” (pID M:11272410463, third cohort). Another explanation applies this reasoning to the ‘players’, thus also thinking in terms of the game task at hand: “The reason why I would give the reward if the player place on the left is because of both dotted boxes are the correct answer. However, I feel that most would place it to the right since it is easier to recognize that H comes after G. So figured that would place it there without knowing they could be gotten the same correct. So I concluded that the reason why I choose the left is that fewer people would pick that.” (pID M:11272095027, third cohort).

7.1.3. Suggestions

Such observations should be further investigated using more controlled experiments. In any case, one piece of conclusive advice that we can offer is that it is useful for the outcome of the experiment if the respondents are given (i.e., as help) alternative choices of answers/explanations (or rationales in our case). These choices should be carefully made, preferably using answers that are observable in the target population (i.e., like we did ourselves, extracting answers from the first survey). A more controlled experiment should yield more sensical answers, e.g., by carrying out the experiment in a more strict ‘laboratory’ setting. It seems that only paying the participants, as we did through the two platforms SurveyMonkey and Amazon’s mTurk, does not increase the quality of the answers.

7.2. Demographics

We categorized the respondents into four age-groups, which we also named: (i) 18-25 – ‘younger students’, (ii) 25-30 – ‘older students’, (iii) 30-59 – ‘professionals’ working in their respective fields, and (iv) older than 59 – ‘approximation of the retirement age’. Relevant for our analysis was to look at which age group is representative for each cohort. As shown in the Figure 8 for ‘Social sciences’ and ‘Natural sciences’ cohorts, these are predominantly composed of respondents between the age of 18 and 25.

In the case of the ‘Arts and Culture’ cohort, the respondents, though they have the same educational/professional background, differ in location: (i) the group with respondents from mTurk are located in different countries, (ii) the group with respondents from SurveyMonkey are located in Sweden, while (iii) the group from ‘Cultural studies’ are located in Norway. In addition, the mTurk and SurveyMonkey respondents were recruited based on their professional affiliation, while the ‘Cultural studies’ respondents were recruited based on study line affiliation. From the chart in Figure 8, we can see that for the mTurk and SurveyMonkey respondents the predominant age group is 30-59, while for the ‘Cultural studies’ respondents is 18-25.

Based on these age groupings we can conclude that for the ‘Social sciences’, ‘Natural sciences’ and ‘Cultural studies’ cohorts the respondents are predominantly ‘young students’, of age between 18-25, while for the mTurk and SurveyMonkey, the respondents are predominantly ‘professionals’, of age between 30-59.
8. Control Questions

In the study we included several additional questions with the purpose to control for various aspects. As one can recall from Section 6.4, we have used the open-ended questions to identify robot/automated answers. Three questions were of particular importance, as they were meant to control, or to reinforce, three important assumptions that we have. These control questions are analysed in detail in the next subsections. Essentially, Section 8.1 reinforces our bias revealing test from Section 3 as a good instrument; Section 8.2 tests how well our priming metaphors from Section 4 worked, since such story-based metaphors may be revealed within listings of words/synonyms; whereas Section 8.3 reinforces our beliefs and categorization of the backgrounds of the three cohorts that we study, thus confirming that the categories/labels we provided in Section 6.1 are appropriate, and the bias transfer studies that we conduct, as described in Section 9.1, are well informed.

8.1. Left/Right Placement

On the ‘Programming task’ page of the survey, the respondents are asked to decide whether to reward the player for the placement of the letter ‘H’ on the left or right side of the vertical line on the game board. This is one of the three questions on this page, intended as a control question for the hypothesis that we made in Section 3, i.e., that choosing to place the letter to the ‘right’ should indicate a preference for the ‘algorithm’ rationale, while choosing ‘left’ a preference for the ‘balance’ or ‘shapes’ rationales.

An analysis of the ‘left/right’ placement wrt. each of the three rationales confirms this initial assumption, see Figure 9b for numbers. In particular, observe that in the case of the ‘algorithm’ rationale the choice of placement to the ‘right’ is overwhelming for each cohort; and similarly, ‘left’ is the preferred choice when answering with the ‘balance’ or ‘shapes’ rationales in all cohorts.

Moreover, the analysis of the ‘left/right’ placement overall inside each cohort, which we summarize in Figure 9a, confirms our earlier observation that the background of the participants is reflected in their preference for one choice of placement (and thus for one type of rationale). The ‘Social sciences’ cohort chose mostly ‘left’, associated with the ‘balance’ and ‘shapes’ rationale, in a proportion of 74.19%. The ‘Natural sciences’ cohort chose mostly ‘right’, associated with the ‘algorithm’ rationale, in a proportion of 53.49%. The ‘Arts and Culture’ cohort again chose mostly ‘left’ in a proportion of 54.08%.

Though there are more pattern combinations possible when the letter ‘H’ is placed on the right, there is also one pattern combination possible with the placement of the letter ‘H’ on the left. This is the sequence of 1-3-2-1 or a pattern of the type 1-3-2-1-2-3.

8.2. Words Suggestions vs. Priming Metaphor

In this section we analyse the words given by the respondents to the ‘Words suggestions’ question of the survey, which was meant as a control question for our priming metaphors, i.e., to reveal which of the primes worked and how well. This is relevant for our second research question: the transfer of contextual bias, which we analyse closer in Section 9.2. In particular, this control question was intended to check the suitability of the priming words that we have chosen. Moreover, in the future, one can extract from these suggested words more adequate priming words, for further research on alternative priming metaphors.

The words were cleaned up for spelling and uppercases. In addition, we changed the derived words to only one
syntactic form, i.e., adjective, verb, noun, or adverb. The precise syntactic form that was kept was decided based on the frequency of that form throughout all the responses. In the very few places where sentences were used instead of words, we kept only relevant individual or composed words (e.g., from “pleasant surroundings” we kept only the word “pleasant”; or from “being crafty” only “crafty”; or “appreciating beauty” was split in two words “appreciation” – changed into a noun, as this was the form most used in the responses – and “beauty”), and removed other syntactic forms, such as conjunctions, prepositions, and pronouns that did not have a meaning by themselves or their meaning was not relevant for explaining the respective ‘life-aspect’. We also created compounded words where this was possible (e.g., “looking good” was changed into “good-looking”). We strived to be minimal in such changes, and we especially did not do semantic changes.

The participants were given three pairs of words to suggest synonyms for, each containing two of the four priming words used in the ‘Philosopher story’, i.e.: “Could you suggest 1 to 4 individual words that for you have similar meanings as each of the three life-aspects: ‘harmony and equality’, ‘balance and fairness’, ‘order and continuity’.”

In analysing the responses for the ‘Words suggestions’ question, we looked for the occurrence of the other two words that were used in the ‘Philosopher story’ as primes (cf. Section 4 also), i.e.:

- For the case when the respondents were given ‘harmony and equality’ (which we will call ‘Aspect 1’) in the rest of the section as words to find synonyms for, we looked for the occurrence of ‘balance’ and ‘fairness’ in their answers.
- For ‘aesthetics and arts’ (‘Aspect 2’) we counted occurrences of ‘beauty’ and ‘forms’;
- For ‘order and continuity’ (‘Aspect 3’) we counted ‘structure’ and ‘linearity’.

Since we had Norwegian speaking respondents we prepared a Norwegian version of the ‘Philosopher story’ with the corresponding words being respectively: ‘harmoni’, ‘likhet’, ‘balanse’, ‘rettferdighet’; and ‘estetikk’, ‘kunstformer’, ‘skjønnhet’; and ‘struktur’, ‘kontinuitet’, ‘gjentakelse’, ‘linearitet’ (see translations in Table 2). In this case, the given words for the ‘Aspect 2’ were actually three of the four priming words because we gave the compound word ‘kunstformer’ which contains in translation both ‘forms’ and ‘arts’. As such, only one word of the four, i.e., ‘beauty’, remained to be counted. However, the responses contain one word that is very similar in meaning with ‘beauty’, namely ‘vakkert’ (beautiful). The degree of similarity was evaluated using the “meaning relation” of the Norwegian WordNet which gives the extremely close numbers 0.155 for ‘skjønnhet’ (beauty) and 0.153 for ‘vakkert’ (beautiful).

An overview of the first five most ‘suggested words’ for each cohort is given in Table 1: “The First Five Most ‘Suggested Words’”. The expected words for the ‘Aspect 1’ appear among the five most occurring words, sometimes at the top, in all three cohorts. For the ‘Aspect 2’, the word ‘beauty’ appears highly ranked, but not the word ‘forms’; again this is the case for all three cohorts. However, for ‘Aspect 3’ we only see the word ‘structure’ appearing among the top five only in the ‘Arts and Culture’ cohort, which had the English version. None of the two corresponding Norwegian words appear in the responses of the ‘Social sciences’ and ‘Natural sciences’ cohorts. The word ‘linearity’ did not appear at all.

We have used the Norwegian WordNet (most recent year count being 2013) frequency feature to check how frequent are the two words of the ‘Aspect 3’ in the Norwegian vocabulary. It turns out that these two words have...
a very low frequency: 0.0001857 for ‘gjentakelse’ (equivalent to ‘structure’) and 0.0000097 for ‘linearitet’ (equivalent to ‘linearity’). Compare these frequencies with those for the other words: 0.001938 for ‘balanse’, 0.001121 for ‘rettferdignett’, which are one order of magnitude higher, or 0.002912 for ‘vakkert’ or 0.001410 for ‘skjønn’ (short version of ‘skjønnhet’ which we used).

We conclude that the use of the word ‘structure’ was successful, and can continue to be used in further research. However, another more widely known word needs to be used as Norwegian equivalent. This concerns the ‘linearity’ word as well, for both Norwegian and English. We were not aware of the existence of WordNet at the time when we prepared our metaphors (in 2018-2019); but such tools can be valuable for choosing priming words. In our case we were not only interested in synonyms, but more in words that could prime towards the three different rationales for our programming test; e.g., the word ‘forms’ was important, as well as ‘linearity’.

Table 2 gives an overview of the numbers of occurrences (frequencies and percentages) of the priming words for each of the cohorts. The percentage numbers are calculated wrt. the number of respondents. In order to evaluate the priming effect of our metaphors we can compare the percentages from the primed group with the ones from the control group. The priming group considered is the one relevant for the respective aspect (highlighted in our tables), i.e., for the ‘Aspect 1’ we look at the numbers of the ‘Harmony and Equality’ group.

For the ‘Social Sciences’ cohort we see the effect of priming on both ‘Aspect 1’ and ‘Aspect 2’: 33.33% (representing the total from both ‘priming’ words) to 11.11% and respectively 50% to 11.11%. The ‘priming’ words for the ‘Aspect 3’ did not occur at all, thus, do not appear in the table. The table for the ‘Natural Sciences’ cohort shows again the influence of priming in the case of ‘Aspect 1’ and ‘Aspect 2’: 73.33% to 33.33%, and 33.33% to 25%. For the ‘Arts and Culture’ cohort we observe the effect of priming in the case of ‘Aspect 1’ and ‘Aspect 2’: 60% to 42.31%, and 21.43% to 7.69%. For the ‘Aspect 3’, only one of the two words occurred, hence only one was considered.

However, for the ‘Aspect 2’ we do not see a priming

<table>
<thead>
<tr>
<th>Group</th>
<th>Nr. Resp.</th>
<th>Balance</th>
<th>Fairness</th>
<th>Balance %</th>
<th>Fairness %</th>
<th>B + F %</th>
<th>Beauty</th>
<th>Skjønnhet</th>
<th>Vakkert</th>
<th>Skjønnhet %</th>
<th>Vakkert %</th>
<th>S + V %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harmony and Equality</td>
<td>15</td>
<td>5</td>
<td>6</td>
<td>33,33 %</td>
<td>40,00 %</td>
<td>73,33 %</td>
<td>3</td>
<td>2</td>
<td>20,00 %</td>
<td>13,33 %</td>
<td>33,33 %</td>
<td></td>
</tr>
<tr>
<td>Estetikk og Kunstformer</td>
<td>9</td>
<td>2</td>
<td>2</td>
<td>22,22 %</td>
<td>22,22 %</td>
<td>44,44 %</td>
<td>2</td>
<td>1</td>
<td>22,22 %</td>
<td>11,11 %</td>
<td>33,33 %</td>
<td></td>
</tr>
<tr>
<td>Struktur og Kontinuitet</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>26,83 %</td>
<td>33,33 %</td>
<td>50,00 %</td>
<td>7</td>
<td>4</td>
<td>16,67 %</td>
<td>8,33 %</td>
<td>25,00 %</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>41</td>
<td>9</td>
<td>11</td>
<td>21,95 %</td>
<td>26,83 %</td>
<td>48,78 %</td>
<td>7</td>
<td>4</td>
<td>17,07 %</td>
<td>9,76 %</td>
<td>26,83 %</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Occurrence of priming words.
effect, when compared with the control group. Quite the opposite, we can observe that the word ‘beauty’ appears many times, with a total number of 53 occurrences and its distribution between the groups relatively even: 50%, 40.63%, 35.71%, and 57.69%. The word ‘beauty’ seems to be present by default in the vocabulary of this cohort, irrespective of priming. This observation confirms that we did well by choosing a word that is largely used by this population.

This indicates two other factors that might have had influence on the priming effect. One of these factors is how familiar the respondents are with the priming words. If the words are very little known or not understood, they will not be primed by them, as it is the case of the ‘Aspect 3’ words. In addition, if the respondents have a large vocabulary to their disposal, which can be seen from comparing the ‘Nr. of unique words’ with the ‘Nr. of words with duplicates’ (see Table 1), the System 1 will be less inclined to use the priming words. Such observations can be made in the case of the ‘Social sciences’ students in comparison with the ‘Natural sciences’ students: 147 unique words compared to 95. We see how the priming was stronger in the latter cohort compared to the former (they strive to find similar words, and the availability heuristic retrieves the primes from the short term memory).

The words the participants choose the most can also be affected by other immediate contextual elements. In the case of the ‘Natural sciences’ cohort the survey was taken by the students as part of a course on logic. This made the word ‘logic’ occur the most for ‘Aspect 3’.

8.3. Life-aspects Ranking

The ‘Ranking life-aspects’ was meant as a control question for the way we identify the background in our cohorts. That is to say, we want to check whether there is a correlation between the self-ranking of the ‘life-aspects’ and what we have considered as the background of the respondents. Moreover, we want to also look at the coded answers from the ‘Self explanation of choice’ compared to the ‘Ranking life-aspects’ because if the correlation is similar to the one we have observed previously from the background, then this would reinforce our perception of background.

For creating the three types of cohorts we have considered the educational and professional backgrounds. However, these are only one part of a person’s background, arguably a large part, but yet a larger part is made of the society and culture that the respondents belong to. This is especially so for younger people, such as students. For the ‘Ranking life-aspects’ we observe influences that come from the socio-cultural as well as educational and professional backgrounds. How strong these are, and how much they relate to the bias transfer that we have observed before, is what we investigate in this section.

We summarize the three types of influences in the Figure 10, organized as a pyramid as we explain further.

Figure 10: Three sources of influence, correlated with the age when they are most strong, for the backgrounds observed for our cohorts, also indicating the age groups observed in the Section 7.2.

Figure 11: First ranked life-aspect in each cohort.
ranked is ‘harmony and equality’, followed by ‘aesthetics and arts’, and ‘order and continuity’ is only last. However, in the case of the ‘mixed location’ group from the Arts and Culture cohort most first ranked is ‘aesthetics and art’ whereas ‘harmony and equality’ is slightly less chosen as first ranked.

Our interpretations are based on the fact that we are knowledgeable when it comes to what characterizes the ‘Scandinavian’ type of culture and society, while having no knowledge about the ‘mixed location’ group. However, our observations for the ‘mixed location’ group are based on their professional background (as opposed to the educational background, in the case of the students’ cohorts).

The fact that ‘Scandinavian’ respondents across all cohorts value highest ‘harmony and equality’ can be motivated by socio-cultural influences. These respondents come from a culture that promotes social equality, with high taxes for social-welfare and strong disregard for social unrest. In the case of the ‘mixed location’ respondents, the view on life seems to be influenced strongest by their professional background, maybe also because these are “working age” respondents (as shown in the Section 7.2).

We also counted the choices of second ranked life-aspect (see Figure 12), looking inside each of the columns from the graph above. For the ‘Social sciences’ cohort the second ranking was ‘order and continuity’, being on the first rank in both columns with ‘harmony and equality’ and ‘aesthetics and arts’. Note that ‘harmony and equality’ can be seen both influenced by education and the socio-cultural background. However, in the column with ‘order and continuity’ the most second ranked is the ‘harmony and equality’ which corresponds with the educational background of the respondents. Thus we tend to conclude that the background of this cohort corresponds to the life-aspect ‘harmony and equality’ where both the socio-cultural and educational backgrounds contribute.

In the case of the ‘Natural sciences’ cohort, the second ranked was ‘order and continuity’, which is in accordance with the educational background.

When it comes to the ‘Arts and Culture’ cohort, the second ranked for the ‘Scandinavian’ group is ‘aesthetics and arts’, in both columns that do not correspond to their education i.e., ‘harmony and equality’ and ‘order and continuity’. This again tells that the educational background influences these respondents, albeit less than their socio-cultural background.

For the ‘mixed location’ group ‘order and continuity’ was ranked as second most in the ‘harmony and equality’ column, while ‘harmony and equality’ was ranked second most in the ‘aesthetics and arts’ column. In this case we tend to conclude that the ‘professional’ background influences the view on life of these respondents the most.

In conclusion, we think that the control question about ‘Ranking life-aspects’ confirms our assumptions about the backgrounds for our three cohorts and the fact that we have associated each of these cohorts with the life-aspect that is most predominant for those respondents. Therefore, we consider adequate the claims that we make throughout the paper where we correlate the background of a cohort with one specific life-aspect, and thus with one specific corresponding bias/rationale.

9. Results

The data is analysed both quantitatively and qualitatively. The qualitative analysis is done usually to detail the quantitative data, by analysing the participants’ responses to the open-ended questions.

Our study is exploratory and tentative, this is why we employ a combination of statistical and descriptive analysis. Statistical analyses were not possible in all situations because of the small number of respondents in those categories. Statistical analyses were possible when the results from all cohorts were put together (analysis across the three cohorts) or in the case of the ‘Arts and Culture’ cohort where we had 139 responses. Results from the descriptive analysis capture several systematic tendencies of the responses, which we detail below.
9.1. Influences from the Cultural Background

Students with a cultural background from social sciences differed significantly from students with a cultural background from computer science. ‘Social sciences’ students were significantly more prone than computer science students to describe their choices matching the rationale ‘balance’ from Section 3.1, whereas computer science students were significantly more prone than ‘Social sciences’ students to describe their choices matching the rationale ‘algorithm’ from Section 3.III: \( X^2(1, N = 71) = 8.1686, p < .05 \) (with calculated p-value of .004262). The results support the hypothesis that the cultural background influences people when they carry out programming tasks under conditions of uncertainty.

The statistical significance test, as well as the graph in the Figure 13a consider the total number of responses, from all four treatments. The same observations about the cultural background influence are confirmed also when looking only at the control group (see the graph in the Figure 13b), though a statistical test is not relevant in this case, given the small number of responses. For both graphs the percentages are calculated from the ‘sensical’ answers only.

When analysing the results from the ‘Arts and Culture’ cohort in comparison with the other two cohorts (see Figure 14), we see that the influence of their artistic background makes them choose much more the ‘shapes’ rationale from Section 3.II; i.e., 16% compared to 7% in the ‘Natural sciences’ cohort and 0% in the ‘Social sciences’ cohort.

Analysing further this cohort by itself, independently of the results obtained for the other cohorts, we see in the Figure 15 that the answers conforming to the ‘algorithm’ rationale are dominant; both when looking at all responses as well as only at the control group. This dominance could be explained by the fact that the respondents tried to comply with the nature and requirements of the exercise, i.e., a programming task where they were asked to assume the role of a programmer. One example of an answer from this cohort confirms this affirmation: “It was always drilled into my head in school, that when it came to math (which I assume is what most programming deals with) that the right side is always the right way... ‘right side right way’ that’s my reasoning here.” The respondent tries in this case to bring in to his/her help the math knowledge s/he has from the school, as s/he assumes that informatics “deals with” mathematics. Another example is “I can’t think of a better explanation but to involve mathematics in this game...”.

Moreover, when analysing qualitatively the answers to the ‘Self explanation of choice’ question we found a considerable number of respondents that brought the game aspect of the task into their reasoning (more than 30 out of 110 explanations of the ‘Arts and Culture’ cohort), i.e., they think in terms of programming a game. This is also an indication that these respondents focused on the task at hand, seeing the puzzle as part of this game programming task – as they have been asked to – and did not try to solve the puzzle per se. This increases our confidence in the fact that there was no debiasing happening, and that the respondents did not recognize that the task was in fact meant to reveal a background bias, let alone one of our three rationales or cohort backgrounds that we have assumed. Another aspect that could trigger debiasing is the fact that our puzzle does not have a ‘correct’ answer wrt. the letter placement. However, we have found only two responses that have identified this fact ( “ [...] because of both dotted boxes are the correct answer. However, I feel [...]” from pID M:11270469031, third cohort,
and “[…]there isn’t enough information for me to decide, so it is kind of a guess.” from pID M:11270119183, third cohort); therefore, we rule out this debiasing possibility as well.

9.2. Influences from the Priming Metaphors

The qualitative analysis of the respondents’ answers to the ‘Self explanation of choice’ shows three instances where the priming metaphors influence the answers of the participants. Here are three examples from the ‘Arts and Culture’ cohort that show how the respondents quote directly the primes from the ‘Philosopher story’ to help in arguing their reasoning behind the choice of letter placement. All three answers fall into the rational category related to the prime; and all invoke only some of the four words that we used for priming. Moreover, these words are taken both from the start and end of the story, which confirms our decision of using several words placed at different points inside the ‘Philosopher story’.

- pID M:11271203853, third cohort: “If this game is based on the philosopher’s tenet of balance and harmony, then […]”\(^\text{15}\)

- pID M:11270378880, third cohort: “The player should be rewarded when he/she places the letter on the right side because that is in keeping with the continuity and linear structure of the game.”\(^\text{16}\)

- pID 103, second cohort: “The philosopher thought balance and equality were important, and the player should therefore be rewarded for restoring the balance between the number of letters on the right and left sides.”\(^\text{17}\)

Since these examples are very few, they do not warrant a conclusion of conscious transfer of priming bias, which is exactly the point with priming techniques, i.e., that people who are influenced through priming generally do not realize it, and thus one does not normally see the priming expressed per se in the respondents arguments. Instead, the respondents being primed would make use of one or more of the heuristics that we mentioned in Sections 2 and 4, e.g., the availability heuristic is most often used when people make quick judgements; we have encountered one respondent in which this heuristic has obviously manifested, pID M:11272410463, third cohort: “I choosed right previously but actually left makes more sense. Balancing the sides; 4 letters on the left, 4 letters on the right.”.

Heuristics are also used substantially in situations of uncertainty, which is the case for our puzzle since we ask participants to find one ‘solution’ to this new puzzle, which at the same time does not have one single correct answer, as any argument would be acceptable. In cases of uncertainty two additional heuristics are usually employed, namely the representativeness heuristic and the anchoring heuristic. If the problem at hand is new, then the mind tries to find another previously encountered problem that, to some extent, has some similarities. This is the case with the puzzle that we devised, aiming to trigger associations with aspects from the cultural/educational background of the person, e.g., ‘Natural sciences’ respondents were expected to cling on to algorithms and the alphabet as an ordered source of indexing in mathematics, thus continuing along the line in our puzzle. The anchoring heuristic is even more important for priming since it is often employed when no useful information is readily available for the problem at hand, so the mind looks into the immediate context (e.g., physical, s.a., surroundings, or temporal, s.a., information received in the recent past, from the short-term memory) for clues. In our case the mind would anchor into the ‘Philosopher story’ metaphor, and maybe draw on the meaning of one of the four priming words.

In analyzing the explanations/responses we observed to some extent influences of our experimental manipulations, albeit not reaching statistical significance. Thus, since we can neither rule out a Type I error (i.e., failing to reject the research hypothesis, i.e., no priming effect) nor a Type II error (failing to reject the null hypothesis, i.e., there exists an effect, but we were not able to elicit it), the influence of contextual metaphors need to be further researched. In the rest of this section we present results from quantitative analyses of the priming effect and whether or not this transferred to the programs, i.e., found in the answers to the ‘Self explanation of choice’ and ‘Alternative explanations’ questions.

The graph in the Figure 16 shows the influence of the three groups of priming metaphors when the responses from all the cohorts are put together. This shows the priming effect irrespective of the participants’ background. We compare each group with the control group.

First of all, we observe that the ‘algorithm’ group gives
answers that cannot be readily seen as being influenced by priming. The same inconclusive observation is found also when looking inside each cohort, comparing the ‘algorithm’ group there with the respective control group. However, this is conforming with the observations made in Section 8.2, where the words used for priming are little known or maybe not understood by the participants, and thus cannot have an impact on their choice. However, one needs to take this conclusion with a grain of salt because the priming metaphor, depending on the anchoring heuristic, has a temporal flavour as it is stronger closer to the time of the priming; i.e., in our case the ‘Self explanation of choice’ question is very close to the priming metaphor, whereas the ‘Words suggestions’ question is farther away, maybe with a delay of a few minutes. This can mean that even if we do not see an effect of the metaphor in the ‘Words suggestions’ question one can still have some effect in the ‘Self explanation of choice’ question. Moreover, this can be compounded by other factors as well, such as for the ‘Words suggestions’ question we are looking only for two of the words whereas in the ‘Self explanation of choice’ question all our four priming words are in effect; or by the semantics of the words, which can have different meanings in different context, thus possibly causing one influence on the programming task and another influence on the synonyms generation task. Therefore, we focus in the rest of the section on the other two groups of priming.

Secondly, when we analyse the other two groups we clearly observe priming influences, albeit of different kinds as explained further. For the ‘balance’ group we see that the ‘balance’ rationale increases from 28.26% (for the control group) to 43.18% (for the ‘balance’ group), whereas in the case of the ‘shapes’ group the ‘shapes’ rationale increase from 8.70% to 15.79%; irrespective of the background.

Besides these observations about the general strength of the priming metaphors, we are to a greater extent interested in their interactions with the educational/professional background of the participants, as discussed in the previous subsection. Since we already established that the results related to the ‘algorithm’ metaphors are not reliable, we exclude them from our further investigation.

In the chart from the Figure 17a, related to the ‘Social sciences’ cohort, we observe that the background of the respondents is further strengthened by the priming metaphors, with an increase from 58.33% to 100%. Quite the opposite, in the case of the ‘Natural sciences’ cohort (see Figure 17b), we see a weakening effect of their background since the ‘algorithm’s rationale decreases from 71.43% to 58.85% in the group that was influenced with the ‘balance’ metaphor, in favor of the ‘balance’ rationale. For the ‘Arts and Culture’ cohort we again see (Figure 17c) that the ‘shapes’ metaphor strengthens their background since the ‘shapes’ rationale increases from 15% to 22.22% inside the group that was primed with the ‘shapes’ metaphor. For this cohort also the ‘balance’ metaphor has an influence (from 15% to 34.62%), due to the fact that this metaphor’s words were well chosen, as we have observed in Section 8.2.

We can conclude that the contextual metaphors that have been deemed as strong enough in the control question ‘Words suggestions’ are also found to have an effect in strengthening or weakening the influence from metaphors in the cultural background of the respondents. Contextual metaphors have a strengthening effect when the words are representative of the respective cultural background, e.g., the ‘balance’ rationale is strengthened by the priming metaphor words ‘harmony and balance; equality and fairness’, whereas the ‘shapes’ rationale by the ‘aesthetics and beauty; forms of arts’ metaphor words. On the other hand, when the contextual metaphor is well chosen, such as for the ‘balance’ group, it weakens the effect from the
cultural background of the respondents, as is the case with the ‘balance’ metaphor, which when applied in the ‘Natural sciences’ cohort, it increases the respective rationale.

10. Conclusions and Discussions

The aim of this study, as well as its implications, are manifold. The study can be categorized as both (i) a comparative/experimental study of how biases from cultural and contextual metaphors can be transferred from programmers to programs, and (ii) an exploratory study on how to develop ergonomically valid and reliable instruments, procedures and testing conditions to empirically study such biases transfer. As such, this paper is a foundation for future research endeavours to improve and diversify these instruments, procedures and testing conditions.

The strengths of this work reside in its exploratory nature in studying a hitherto not researched phenomenon, namely the transfer of human biases from the (not necessarily expert) programmer to the artefact that is developed (or configured). Concretely, we have exposed (in Section 9) interesting aspects of our main hypothesis, namely that machine bias may originate not only from biased data, but also from the programmer’s biases in terms of influences from the cultural background as well as contextual influences from the programming environment.

Interestingly, under conditions of uncertainty (e.g., in the absence of instructions or specifications, something which is often the case for ubiquitous systems programming carried out increasingly by non-experts), we observe that the programmers’ cultural background influences the choices they make and are subsequently transferred from the programmer to the program artifact. Thus, cultural metaphors in terms of irrelevant and inappropriate influences on the programming task at hand, represent instances of biases that are being transferred from humans to machines. This implies that human culture ‘transfers’ to machines through the humans that program these, thus representing a strong source of bias.

Equally interesting, attempts to moderate the strong influence from the cultural metaphors by means of experimentally introducing ‘hidden’ (i.e., not consciously detected) contextual metaphors, were only successful to a certain extent. When the priming metaphor was chosen well (as in the case of ‘philosopher story’ related to the ‘balance’ rationale; with words that were easy to understand and rather common in a standard vocabulary) we saw influences in both directions of strengthening the cultural background as well as moderating it, each time tipping the balance of answers in the direction of the metaphor. These findings are orthogonal to what traditional and current machine bias research suggests, i.e., that machine bias originates from data, and thus our findings provide new insights into the origins of bias in the wide spreading AI and decision-support systems.

We believe that the present study shows how various aspects regarding design, instruments, and procedures can be successfully explored and controlled, and consequently incorporated in future studies that could (i) extend the present study by exploring related causes and mechanisms that lead to the transfer of bias from programmers to programs, as well as (ii) improve the designs, instruments and procedures in order to undertake this expanded endeavour.

10.1. Discussions and Limitations

One potential limitation of our study is the relatively small numbers of explanations/responses that we obtained in the experimental groups, despite the fact that we have had quite a large number of participants, i.e., 242 completely sensical explanations/answers that we were able to interpret and categorize, plus 55 more that were not completely nonsensical, but nevertheless uninterpretable and uncategorizable. This limitation is explained, however, by our intention to perform a manifold exploratory study, investigating several related aspects and research questions, as well as providing a transparent foundation for further future improvements of design, instruments and procedures. As a consequence we arrived at a limited number of sensical and interpretable explanations ‘inside’ each of the respective cohorts. However, a major intention in our study is to allow future researchers to take our study as a point of departure and carry out more controlled studies of some of the single aspects that we identified. In order to achieve increased confidence in the results that we have presented here, it would be valuable to perform our survey with even more participants, maybe in the order of 1300+ (estimating, according to the observations from Section 7.1, with a 4-to-1 ratio of sensical vs. nonsensical results). Such a number would allow for all our groups to be populated with 30+ responses, something which would allow for an even more valuable use of statistical significance tests.

It proved difficult to achieve our number of 300 participants for this type of study; however, if one would want to restrict a similar study to only professionals, one of the many online forums or communities for programmers could be a good place to recruit participants (e.g., on freelancer.com orstackoverflow.com).

One interesting speculative observation that we would like to make out of our results regards a potential effect resulting from the difference between (i) interpreting information based on its structure and thus as something systemic that is ‘detached’ from having individual characteristics, versus (ii) interpreting information based on its content and thus as having individual characteristics. For example, subject-programmers that chose the rationales of ‘balance’ or ‘algorithm’ may view information (as the one coming from the ‘game board’ puzzle picture that we showed them) merely as representing structure and may thus have disregarded the notion that data could also have individual characteristics in addition to being part of an overall structure. Contrary to this, respondents that chose ‘shapes’ may in fact have acknowledged the notion that data do have individual characteristics and are thus
not ‘only’ part of an overall structure ‘outside’ the data’s individual characteristics. Interestingly, subjects in the arts & culture cohort provided explanations in terms of ‘shapes’ substantially more often than subjects in the ‘Social sciences’ cohort and the ‘Computer science’ cohort. This could indicate that people with a cultural background (judging from their education and/or profession) from arts and culture are more prone than others to view data as representation of individuals that have unique characteristics, rather than viewing data only as being part of an overall structure. In other words, people with a background in arts and culture may possibly exhibit a more ‘human’ interpretation of data, or at least they may be more prone than people from other cultural backgrounds to acknowledge data as ‘individual’ rather than ‘systemic’.

10.2. Possible Future Research Directions

Future studies are invited to investigate more deeply any of the aspects that we have explored and tentatively concluded from. One venue for future research would be to refine our study design’s ability to elicit cultural or contextual influences in an even more fine-grained manner, specifically by improving our instruments and procedures.

One possibility is to perform similar studies focused on specific categories of subjects that can be seen as programmers, e.g., one playfull possibility could be to study children as programmers – programming languages/environments specific for children abound, s.a. Google’s Blockly Trower & Gray (2015); Weintrop & Wilensky (2017) or MIT’s Scratch Resnick et al. (2009); Maloney et al. (2010); Armoni et al. (2015). Studies on biases in adults are more available Klaczynski et al. (1997); Klaczynski & Robinson (2000); Bruine de Bruin et al. (2007) whereas studies on biases in children are less Baron et al. (1993); Klaczynski (1997). One could argue that this is because children are not biased; others could claim that ethical considerations make such studies of children too difficult to carry out; yet others could argue that biases in children are distinctly different from biases in adults, given the differences in mental representations from children and adults. However, we think that it is important to test the age aspect in biases transferred to programs, given the ubiquity and pervasiveness of IoT-programming in everyday life for all age groups.

One useful refinement of our work could be to study professional programmers in a professional environment, both (i) classical programming environments, such as in a programming company guided by software development life cycle methods and tools, maybe focusing on current emerging programming cultures like Scrum or DevOps; and (ii) non-expert programming environments, such as complex configurations, DSLs, graphical programming, or curating of big data. It is possible to study different questions in this setting, for example:

- **On what avenues biases arrive into the programming environment?** We have assumed that it is a result of underspecified requirements; i.e., when some functionality is left open and the programmer does not have the resources to find more specific details. This is a most common form of uncertainty in programming; but there are others as well, and which of these give way to biases is important to know so that one can build debiasing techniques, maybe even incorporated in the tools of the programmers, like in IDEs (integrated development environments).

- **Are expert programmers, when under the scrutiny of their tools and methods, like testing suites, still transferring their biases into their programs?** Is this happening only when they are given choices, or also in other situations (e.g., even when fully specified requirements are given; or when working with big data)?

- **Are programmers immune to biases because of their education or because of their work (e.g., because at work their ‘mind-set’ is a “mathematical” one)?**

One good source of alternative investigations can be the study of specific biases in specific situations or social activities where software is paramount. One example can be biases related to privacy in the big data economy (sometimes called the ‘surveillance capitalism’ Zuboff (2019)), e.g.: Are privacy related concepts or views from the cultural background – which is specific to the programmer – transferred to the software – which is used on an international scale? One can imagine a programmer coming from a cultural background that always promotes the slogan “You have zero privacy; get over it!”, or another programmer from a background that “is entrenched by rules and regulations about who/how any form of private electronic data can be used”. Are such different cultural views transferred to the software built by these two different programmers? What is the global influence of such bias transfers? In this setting, one could alternatively study biases coming from the user of the software (not the programmer) to see whether the user biases (call them ‘wishes’ or ‘needs’) are transferred to the software through specifications elicitation, user stories, and other interaction design methods Rogers et al. (2011); Lazar et al. (2017) that are now a popular way of developing software systems.

We have studied two sources of biases, namely cultural biases and priming metaphors, that we consider situated at the two extremes on the vertical axis from Figure 1, which indicates the strength of the bias, and also a temporal aspect regarding the persistence of these biases (e.g., priming may not be as strong as the culture, and acts on a short time scale, usually minutes after the priming is applied). One could study other sources of influence that can generate biases through the metaphors that they induce in the human, which would lie on our vertical axis in between the ones studied in this paper. Examples of influencing methods relevant for our study can be propaganda (i.e., misinformation Mintz et al. (2012); Kumar
& Geethakumari (2014) and disinformation Graham & Metaxas (2003)), which may be done on limited but considerable stretches of time; or working cultures which can influence a programmer in different ways when changing jobs.

Acknowledgements

The authors would like to thank the student volunteers, their tutors and student administrative personnel from Bjørknes College, from Institute for Informatics at the University of Oslo, and from Oslo Art School (http://www.oslokulturskole.no) and Norwegian Academy of Music (https://nhm.no/) for their help with our studies. We would also like to thank John Søren Pettersson from the Karlstad University (Sweden) for allowing us to utilize their Usability laboratory and the eye tracking equipment for our pilot testing.

References

References


Maloney, J., Resnick, M., Rusk, N., Silverman, B., & Eastmond,


Appendix A. The Questionnaire

Survey on natural language in computer programming

Dear participant,

There is now a growing number of programmers using programming languages that do not require programming skills or education. We are therefore conducting research regarding programming done by non-experts and we invite you to participate in a brief exercise (that should take minimum 8 minutes) about the use of natural language to explain the function of computer programs.

Your participation is anonymous. It is not possible for anyone to identify you (you are anonymous also to us researchers). The results will be used for research purposes only and your individual answers will not be made public.

We do not test your skills or your intelligence in this study. It is not required to have any prior knowledge of computer programming.

The information in this study is confidential until the research is completed. It is therefore important that you complete the exercise on your own and do not discuss it with others.

Your contribution is very valuable to our research. Thank you for your participation.

Instructions:

The buck button is disabled. You will not be able to go back to a previous question, so we ask you to read each question carefully, because some depend on the previous ones. Please put effort into reading carefully everything on each page.

We will disqualify responses that spend less than 8 minutes or who we identify that answered randomly. Completing the survey never takes more than 15 min, and takes 10 min on average.

The questions not marked with star (*) can be skipped.

Figure A.18: Page 1 ‘Introductory text’.

Figure A.19: Page 2 ‘Instructions’.
Survey on natural language in computer programming

Please read the background information below:

A philosopher who lived a life filled with harmony and balance created the riddle used in the game we ask you to imagine that you program on the next page. Although the philosopher is nearly forgotten today, we know that the philosopher influenced many contemporary philosophers’ view of the world. The most prominent influence seem to have been the importance of maintaining equality and fairness in life and in society.

A philosopher who lived a life filled with aesthetics and beauty created the riddle used in the game we ask you to imagine that you program on the next page. Although the philosopher is nearly forgotten today, we know that the philosopher influenced many contemporary philosophers’ view of the world. The most prominent influence seem to have been the importance of maintaining various forms of arts in life and in society.

A philosopher who lived a life filled with order and structure created the riddle used in the game we ask you to imagine that you program on the next page. Although the philosopher is nearly forgotten today, we know that the philosopher influenced many contemporary philosophers’ view of the world. The most prominent influence seem to have been the importance of maintaining linearity and continuity in life and in society.

A philosopher created the riddle used in the game we ask you to imagine that you program on the next page. Although the philosopher is nearly forgotten today, we know that the philosopher influenced many contemporary philosophers’ view of the world. The most prominent influences seem to have been the importance of ethics in life and society.

* Do you think the philosopher was:

- a woman
- a man
Survey on natural language in computer programming

First, spend one minute imagining how you would be programming the simple task below. Then proceed to answering the following questions.

Imagine that you are a non-expert programmer who is developing a simple puzzle game. The game is based on a riddle made by the philosopher that you read about previously.

Imagine that you have already drawn the game board that you can see below:

```
A
B
C
D
E
F
G
H
```

Now you are going to program the player’s interaction with the game.

The player (NOT YOU, you are the programmer of the game) has to solve the puzzle by drag-and-dropping the letter H on to one of the two dotted boxes. The player is rewarded if the program accepts the placement of the letter H as the correct placement.

Please answer the following questions.

* What would you have drawn first?
  - [ ] the vertical line
  - [ ] the letters

* Would you give the reward when the player places H on the dotted box to the
  - [ ] left
  - [ ] right

* Originally, the philosopher used capital letters (as shown on the board). Would you
  - [ ] keep capital letters
  - [ ] use normal small letters
Survey on natural language in computer programming

Now, imagine that you have decided on which side (left or right) the player should be rewarded. Please tell us WHY you decided to instruct the program to reward the player for placing the letter on that particular side. The WHY means that you cannot simply state “left” or “right”, but that you need to provide an explanation.

Please try your best and write your explanation in the text box below. If you are not able to explain your choice, you can proceed to the next page and choose between some alternatives there.

A reminder of what the game looks like:

```
A  B  C  D
E  F  G
H...
```

Figure A.22: Page 5 ‘Self explanation of choice’.
Survey on natural language in computer programming

A reminder of what the game looks like:

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>F</td>
<td>G</td>
<td></td>
</tr>
</tbody>
</table>

* Alternatives to choose from in case you could not explain in text before.

- I just chose something and I am not able to explain why.
- The same number of letters on each side.
- It creates a pattern of the type: 1-3, 2-4, 3-5, ... or 1-3-2, 1-3-2, ... or 1-3-2, 2-3-1, ...
- The number of letters increases by one on both sides: 1, 2, 3, ... on the left and 3, 4, 5 on the right.
- Because of the appearance/form of the letters. On the left side they have straight lines, whereas on the right side are rounded.
- Because of how the letters sound when pronouncing them.
- I already gave an explanation in the text box before.

Figure A.23: Page 6 ‘Alternative explanations’.

Survey on natural language in computer programming

* Please rank the following three pairs of life-aspects in the way that best reflects how you view life yourself (where 1 is the highest while 3 is the lowest).

| Harmony and Equality | Aesthetics and Arts | Order and Continuity |

Figure A.24: Page 7 ‘Ranking life-aspects’.

262
Survey on natural language in computer programming

Could you suggest 1 to 4 individual words that for you have similar meanings as each of the three life-aspects (they are the same aspects that you ranked in the previous question):

<table>
<thead>
<tr>
<th>Harmony and Equality</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Aesthetics and Arts</td>
<td></td>
</tr>
<tr>
<td>Order and Continuity</td>
<td></td>
</tr>
</tbody>
</table>

Figure A.25: Page 8 ‘Words suggestions’.
Survey on natural language in computer programming

* Which field of study or/and line of work does your background so far consist mainly of

- Natural Sciences (physics, biology, chemistry, mathematics, technology, medicine, etc.)
- Social Sciences (anthropology, archaeology, communication studies, economics, history, musicology, human geography, jurisprudence, linguistics, political science, psychology, public health, and sociology, etc.)
- Visual Arts and Design (graphic design, illustration, clothing design, costume design, interior design, architecture, furniture design, etc.)
- Performing arts (music, dance, theatre)
- Other (please specify)

* What type of leisure activity (hobby) do you engage in, or, if you don’t have time to engage in leisure activities, what type would you have engaged in if you had the time? Please state any activity of your choice.

* Years of higher education (after high school) you have attended until now

* Your gender
- Female
- Male
- I prefer not to say
- Other

* Your age