Université de Montréal

A Multi-agent Nudge-based Approach for Disclosure Mitigation Online

par

Rim Ben Salem

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

en vue de l'obtention du grade de
en Informatique

22 Aout 2023

Université de Montréal
Faculté des arts et des sciences

A Multi-agent Nudge-based Approach for Disclosure Mitigation Online

Rim Ben Salem

Gilles Brassard
(président-rapporteur)

Esma Aïmeur
(directeur de recherche)

Louis Salvail
(membre du jury)

Résumé

En 1993, alors qu'Internet faisait ses premiers pas, le New York Times publie un dessin de presse désormais célèbre avec la légende "*Sur Internet, personne ne sait que tu es un chien*". C'était une façon amusante de montrer qu'Internet offre à ses usagers un espace sûr à l'abri de tout préjugé, sarcasme, ou poursuites judiciaires. C'était aussi une annonce aux internautes qu'ils sont libres de ne montrer de leurs vies privées que ce qu'ils veulent laisser voir. Les années se succèdent pour faire de cette légende une promesse caduque qui n'a pu survivre aux attraits irrésistibles d'aller en ligne. Les principales tentations sont l'anonymat et la possibilité de se créer une identité imaginée, distincte de celle de la réalité. Hélas, la propagation exponentielle des réseaux sociaux a fait chevaucher les identités réelles et fictives des gens. Les usagers ressentent un besoin d'engagement de plus en plus compulsif. L'*auto-divulgation* bat alors son plein à cause de l'ignorance du public des conséquences de certains comportements.

Pour s'attirer l'attention, les gens recourent au partage d'informations personnelles, d'appartenance de tous genres, de vœux, de désirs, etc. Par ailleurs, l'espoir et l'angoisse les incitent aussi à communiquer leurs inquiétudes concernant leurs états de santé et leurs expériences parfois traumatisantes au détriment de la *confidentialité* de leurs *vies privées*. L'ambition et l'envie de se distinguer incitent les gens à rendre publics leurs rituels, pratiques ou évènements festifs engageant souvent d'autres individus qui n'ont pas consenti explicitement à la publication du contenu. Des adolescents qui ont grandi à l'ère numérique ont exprimé leurs désapprobations quant à la façon dont leurs parents géraient leurs vies privées lorsqu'ils étaient enfants. Leurs réactions allaient d'une légère gêne à une action de poursuite en justice. La divulgation multipartite pose problème.

Les professionnels, les artistes ainsi que les activistes de tout horizon ont trouvé aux réseaux sociaux un outil incontournable et efficace pour promouvoir leurs secteurs. Le télétravail qui se propage très rapidement ces dernières années a offert aux employés le confort de travailler dans un environnement familier, ils ont alors tendance à négliger la *vigilance "du*

bureau" exposant ainsi les intérêts de leurs employeurs au danger. Ils peuvent aussi exprimer des opinions personnelles parfois inappropriées leur causant des répercussions néfastes.

L'accroissement de l'insécurité liée au manque de vigilance en ligne et à l'ignorance des usagers a mené les chercheurs a puiser dans les domaines de sociologie, des sciences de comportement et de l'économie de la vie privée pour étudier les raisons et les motivations de la divulgation. Le "*nudge*", comme approche d'intervention pour améliorer le bien-être d'un individu ou d'un groupe de personnes, fût une solution largement adoptée pour la préservation de la vie privée. Deux concepts ont émergé. Le *premier* a adopté une solution "*one-size-fits-all*" qui est commune à tous les utilisateurs. Quoique relativement simple à mettre en œuvre et d'une protection satisfaisante de la vie privée, elle était rigide et peu attentive aux conditions individuelles des utilisateurs. Le *second* a plutôt privilégié les préférences des usagers pour résoudre, même en partie, la question de personnalisation des "nudges". Ce qui a été motivant pour les utilisateurs mais nuisible à leurs confidentialités.

Dans cette thèse, l'idée principale est de profiter des mérites des deux concepts en les fusionnant. J'ai procédé à l'exploration de l'*économie de la vie privée*. Les acteurs de ce secteur sont, autres que le propriétaire de données lui-même, le courtier qui sert d'intermédiaire et l'utilisateur de ces données. Le mécanisme d'interaction entre eux est constitué par les échanges de données comme actifs et les compensations monétaires en retour. L'équilibre de cette relation est atteint par la satisfaction de ses parties prenantes. Pour faire de bons choix, l'équitabilité exige que le propriétaire de données ait les connaissances minimales nécessaires dans le domaine et qu'il soit conscient des contraintes qu'il subit éventuellement lors de la prise de décision.

A la recherche d'un utilisateur éclairé, j'ai conçu un cadre que j'ai nommé *Multipriv*. Il englobe les facteurs d'influence sur la perception des gens de la vie privée. J'ai ensuite proposé un *système multi-agents* basé sur le "nudge" pour l'atténuation de la divulgation en ligne. Son principal composant comprend trois agents. Le premier est l'agent objectif Aegis qui se réfère aux solutions généralisées axées sur la protection des données personnelles. Le second est un agent personnel qui considère le contexte dans lequel se trouve le propriétaire de données. Le dernier est un agent multipartite qui représente les personnes impliquées dans le contenu en copropriété.

Pour évaluer le système, une plateforme appelée Cognicy est implémentée et déployée. Elle imite de véritables plateformes de réseaux sociaux par l'offre de la possibilité de créer

un profil, publier des statuts, joindre des photos, établir des liens avec d'autres, etc. Sur une population de 150 utilisateurs, ma proposition s'est classée meilleure que l'approche de base non spécifique au contexte en termes de taux d'acceptation des "nudges". Les retours des participants à la fin de leurs sessions expriment une appréciation des explications fournies dans les "nudges" et des outils mis à leur disposition sur la plateforme.

Mots clés: Divulgation, économie de la vie privée, informations personnelles, nudge, Multipriv, système multi-agent, Aegis, agent personnel, agent multipartite, copropriété.

Abstract

When the internet was in its infancy in 1993, the New York Times published a now-famous cartoon with the caption "*On the Internet, nobody knows you're a dog.*". It was an amusing way to denote that the internet offers a safe space and a shelter for people to be free of assumptions and to only disclose what they want to be shown of their personal lives. The major appeal to go online was *anonymity* and the ability to create a whole new persona separate from real life. However, the rising popularity of social media made people's digital and physical existences collide. *Social Networking Sites* (SNS) feed the need for compulsive engagement and attention-seeking behaviour. This results in *self-disclosure*, which is the act of sharing personal information such as hopes, aspirations, fears, thoughts, etc. These platforms are fertile grounds for oversharing health information, traumatic experiences, casual partying habits, and co-owned posts that show or mention individuals other than the sharer. The latter practice is called *multiparty disclosure* and it is an issue especially when the other people involved do not explicitly consent to the shared content. Adolescents who grew up in the digital age expressed disapproval of how their parents handled their privacy as children. Their reactions ranged from slight embarrassment to pursuing legal action to regain a sense of control.

The repercussions of privacy disclosure extend to professional lives since many people work from home nowadays and tend to be more complacent about privacy in their familiar environment. This can be damaging to employees who lose the trust of their employers, which can result in the termination of their contracts. Even when individuals do not disclose information related to their company, their professional lives can suffer the consequences of sharing unseemly posts that should have remained private.

For the purpose of addressing the issue of oversharing, many researchers have studied and investigated the reasons and motivations behind it using multiple perspectives such as *economics*, *behavioural science*, and *sociology*. After the popularization of *nudging* as an intervention approach to improve the well-being of an individual or a group of people, there was an emerging interest in applying the concept to privacy preservation. After the initial

wave of non-user-specific *one-size-fits-all* propositions, the scope of research extended to *personalized solutions* that consider individual preferences. The former are privacy-focused and more straightforward to implement than their personalized counterparts but they tend to be more rigid and less considerate of individual situations. On the other hand, the latter has the potential to understand users but can end up reinforcing biases and underperforming in their privacy protection objective.

The main idea of my proposition is to merge the concepts introduced by the two waves to benefit from the merits of each. Because people exist within a larger ecosystem that governs their personal information, I start by exploring the economics of privacy in which the actors are presented as the data owner (individual), broker, and data user. I explain how they interact with one another through exchanges of data as assets and monetary compensation, in return. An equilibrium can be achieved where the user is satisfied with the level of anonymity they are afforded. However, in order to achieve this, the person whose information is used as a commodity needs to be aware and make the best choices for themselves. This is not always the case because users can lack knowledge to do so or they can be susceptible to contextual biases that warp their decision-making faculty. For this reason, my next objective was to design a framework called *Multipriv*, which encompasses the factors that influence people's perception of privacy.

Then, I propose a *multi-agent nudge-based approach for disclosure mitigation online*. Its core component includes an objective agent *Aegis* that is inspired by privacy-focused one-size-fits-all solutions. Furthermore, a *personal agent* represents the user's *context-specific* perception, which is different from simply relying on preferences. Finally, a *multiparty agent* serves to give the other people involved in the co-owned content a voice.

To evaluate the system, a platform called *Cognicy* is implemented and deployed. It mimics real social media platforms by offering the option of creating a profile, posting status updates, attaching photos, making connections with others, etc. Based on an evaluation using 150 users, my proposition proved superior to the baseline non-context-specific approach in terms of the nudge acceptance rate. Moreover, the feedback submitted by the participants at the end of their session expressed an appreciation of the explanations provided in the nudges, the visual charts, and the tools at their disposition on the platform.

Keywords: disclosure, self-disclosure, multiparty disclosure, economics of privacy, personal information, nudge, decision-making Multipriv, nudge-based multi-agent system, personal agent, Aegis, multiparty agent.

Contents

List of tables

List of figures

List of acronyms and abbreviations

SNS	Social Networking Sites
SPT	Social Penetration Theory
SPD	Sensitive Personal Data
PII	Personally Identifiable Information
MAS	Multi-Agent System
AI	Artificial Intelligence
ML	Machine Learning
ML	System Usability Scale
IT	Internet Technology
EU	European Union
PIPEDA	Personal Information Protection and Electronic Documents Act

GDPR	General Data Protection Regulation
ICCPR	International Covenant on Civil and Political Rights
ECFR	Charter of Fundamental Rights and Freedoms
OPC	Office of the Privacy Commissioner
NLP	Natural Language Processing
NER	Named Entity Recognition

Acknowledgement

I would like to thank many people whose undying support paved the way for my work.

I want to express my profound gratitude to Professor Esma Aïmeur who has been my research director and mentor for the past few years. She spared no effort in guiding me and I am proud to say that I have learned so much from her that I will carry with me for the rest of my life. Beyond her renowned expertise, she has offered me a genuine human connection and kindness that eased my troubles.

I am grateful to Professor Hicham Hage who has collaborated with me and contributed to the advance of my thesis. He always offered an enlightening perspective. The brainstorming sessions that we had made me more passionate about my research subject and the potential I can reach.

I would like to thank Professor Gilles Brassard with whom I had many constructive discussions during my research. His invaluable input and insightful feedback allowed me to improve the quality of my research.

I am extending my gratitude to the members of the jury, Professor Gilles Brassard, Professor Louis Salvail, Professor Esma Aïmeur, and Professor Julita Vassileva for the time and effort they dedicated to judging my work.

My most special thank you goes to my father Rabah Ben Salem and mother Najah Saïdi without whom I would not be where I am now. There was a time before even stepping foot in Canada when I thought this dream was out of reach. They never did. They always believed in this and most of all in me. They showered me in love and support fueling me to go further and persevere to reach my goal. I am blessed by their presence in my life.

My heartfelt appreciation goes to my older brother Farouk who was by my side during the hardest time of my life. He knew when to actively offer help and when to watch over me as I pursue my endeavors. The past year brought us closer together more than ever before, which I'm very grateful for. I am deeply thankful to my younger brother Abdennasser who

checks up on me all the time and with whom conversations are always a breath of fresh air. My mood instantly brightens when we talk about our shared hobbies and discuss the latest chapter of our favourite work.

Finally, to everyone at the lab whose company I cherished a lot throughout the years, I wish you all the best. In particular, I would like to thank Muxue Guo for kindly sending me her feedback on a section of this dissertation.

Introduction

Social Networking Sites (SNS), also known as social media platforms, are proliferating exponentially and have undoubtedly become an intrinsic part of our daily lives. There are numerous benefits that have been reaped from them such as building and maintaining contact with communities, empowering minorities by supporting social causes, and offering an opportunity for small business owners to grow a customer base.

However, the advantages of this thriving digital environment are often eclipsed by the looming concerns that seem to be getting more serious as time goes by. One of them is that SNS foster the spread of *fake news*, warp users' opinions and beliefs, and cause physical harm during times of crisis such as the anti-vaxxer propaganda throughout the Covid-19 pandemic [**1**]. Another major issue that shrouds these platforms in uncertainty and risk is *privacy*. With the rise of *user-generated* content like videos on Youtube and TikTok, a hub for *disclosure* has been established and reinforced. *Oversharing* details about one's private life has become the norm due to the effect of the *Social Influence Theory* (SIT), which states that from a psychological perspective, people are likely to adopt the commonly observed behaviour [**2, 3**]. This is particularly the case when the action of sharing a photo or private post is met with praise and *gratification* in the form of likes, comments, re-shares, gaining followers, etc [**4**]. Multiple *behavioural biases* act as catalysts for disclosure online such as *reciprocity*, which explains that individuals are more likely to open up after the other party shares their own secret. The findings of a study by Acquisti *et al.* [**5**] show that participants tend to confess their past illegal behaviours when other users had already acknowledged their own misdeeds. Lee *et al.* [**6**] proved through their research that this is applicable even when the other party is a *chatbot*, which is a computer program that simulates human conversation.

Even if privacy-aware individuals decide not to create an account on any platform and simply navigate other websites to read the news, check the weather, or make a purchase, they do not escape the clutches of SNS tracking. All websites that contain a Facebook "like" or "share" button, for example, can lead to the creation of *shadow profiles* that describe

non-users' activity for the purpose of monetization [**7**]. An entire economy has been created around the process of selling, exchanging, processing, and transforming personal data. The current situation is nothing short of the commodification of private life.

Facing these growing issues and concerns, various approaches have been proposed such as regulating the economy of data and working on legislations to counter malicious acts that infringe on privacy such as *doxing*[4]. While extremely important, these solutions do not attend to one of the core issues, which is addressing individuals with the goal of changing their behaviour and mitigating disclosure. Hence, there was a need for user-specific approaches that focus on the human in the loop. Thus, *frameworks of influence* have emerged.

The **first** wave of which was spearheaded by *classic paternalism*, which describes interventions in which an individual's will, liberty, or autonomy is compromised for the sake of their well-being or best interest. They are built on the idea that someone with more knowledge and authority (boss, expert, father, etc.) is more qualified to determine what course of action is the best for the person they are in charge of. These interventions faced a lot of criticism due to how they treat adults as infants who need to be put in a child harness to prevent them from harming themselves. The **second** wave comes in the form of *choice architectures* or *libertarian paternalism* as coined by Richard Thaler and Cass Sunstein [**8**]. The authors also popularized the term *nudges* to mean the same thing as the aforementioned terms. The principle of this is that it is possible to affect human behaviour and steer decisions without limiting freedoms. Since then, nudging for privacy preservation gained momentum and continues to show promising potential.

0.1. Problem

One-size-fits-all nudge-based approaches are simple to implement since no user modelling is required and everyone is served identical interventions. They have the advantage of being privacy-focused and not catering to individual preferences. However, that can also be their downfall because people want to be addressed individually[**9**], and in the absence of all forms of tailoring, privacy nudges can feel like any popup that they see on web browsers aiming to deter them from accessing dangerous content. It is unrealistic in the age of SNS to expect users to stop any and all forms of self-disclosure and multiparty disclosure because these habits have put down their roots in individuals' daily lives. As for the other types of approaches that offer personalized nudges, they tend to be based on a *preference elicitation* process during which users self-report what they like to share and what they can refrain

[4]Doxing is the act of revealing identifying information about someone online, such as their real name, home address, workplace, phone, financial, and other personal information. This is usually done with malicious intent.

from disclosing if prompted to do so. Such systems understand the users better and tend to be better received by users thanks to their flexibility. But, these advantages come at the expense of privacy preservation. If Alice expresses that she likes to post her location publicly, for example, a typical personalized nudge-based system would not react when such a situation occurs because it aligns with her preferences. Let us go through the problems hindering progress in this field, one by one.

First: It is difficult to understand who has access to our data online. Understanding how our data exists in the digital world is paramount to raising awareness of privacy. One of the reasons why people disclose their data without much care is because they think that it is already out there so there is no harm in re-sharing it. People will not value something and seek its protection if it is perceived as common knowledge and unfortunately, privacy is deemed by many as a runaway train.

Second: There are no comprehensive user-centric frameworks that illustrate the factors leading to self-disclosure and multiparty disclosure. Many scholars have proposed models for self-disclosure with the aim of pinpointing the factors or catalysts leading to it [**10, 11, 12, 13**]. However, multiparty disclosure has not garnered the same interest. Researchers have been focusing on proposing mechanisms to resolve conflicts that may arise between users [**14, 15, 16, 17, 18**], but they have not tackled the origin of the issue in a similar manner to self-disclosure.

Third: There are no approaches that aim to find a middle ground between one-size-fits-all and personalized nudges. There is a need to find a middle ground between what the aforementioned two approaches offer. Balancing privacy and personalization is one of the major challenges that will continue to face researchers and scholars as more technological advances are introduced to customers. In today's world where people have gotten used to drawing benefits from intelligent systems and platforms, which are capable of predicting their needs before even expressing them, safeguarding privacy is not only a growing concern but also an urgent need.

Fourth: There are no solutions that consider the context. Although many scholars have acknowledged the impact of the context on decision-making [**19, 20, 21, 22, 23**], the existing personalized privacy-preserving approaches do not consider this factor in their respective proposed solutions. Context-aware nudging remains an unexplored field despite the decade-old calls for leveraging it as a support tool for better decision-making [**24, 25**].

Fifth: The potential of using the framing effect for personalized nudges is yet to be explored. While the *framing effect* has been well documented in the literature [**26, 27**], it is yet to be used in personalized approaches. It refers to how the wording and presentation of an item to users has the potential to sway their decision. The framing effect occurs when people react differently to something depending on whether it is presented as positive, negative, or even neutral. For example, people tend to buy discounted products that clearly display the original price because they think they are getting a great deal. If the store simply shows the new price, they are less likely to make the purchase despite the current price being the same either way. There is an emerging interest in using the framing effect for the purpose of privacy preservation [**28**], but even then, it was integrated as part of one-size-fits-all nudges. As a result, all users end up receiving the same phrasing regardless of their situation or preferences.

0.2. Objectives and contributions

This dissertation contributes to advancing the research on privacy preservation on social media, specifically disclosure mitigation. The overarching goal is to design context-aware personalized nudges capable of resolving self-disclosure and multiparty disclosure issues and serving as a middle ground between the existing one-size-fits-all and the preference-based solutions. To achieve this, I outline four major objectives that can also be perceived as my main contributions.

0.2.1. Designing a multi-agent system to understand the economics of privacy

In order to address users for the purpose of disclosure mitigation, we must first examine who can infringe on their privacy and how much freedom they truly have over their own personal data. While most laymen would agree that some party is collecting data about them for one reason or another such as targeted advertisement, personalized services, or governmental oversight for security reasons, they do not know the extent of data mining, its profitability for others, nor how vulnerable it renders them as data owners. I tackle this objective by proposing the following:

- Scenario and rules that govern the interactions between the parties involved: data owners, brokers, and data users. Each one of them is represented through an agent, hence the appellation: "multi-agent system".
- Utility functions that mathematically represent each agent's objective.
- The strategies that each agent should follow to reach an equilibrium.
- A simulation of the proposed system, which aligns with the literature on the subject.

0.2.2. Proposing MULTIPRIV: A framework for MULTIfaceted PRIVacy decisions

After tackling the economics of privacy as a whole, I focus on the data owner, often referred to simply as the user whose privacy is in jeopardy. Understanding the factors leading to disclosure both on the personal and multiparty levels is the keystone of enhancing personalized approaches if we are to move beyond preference-based systems. As such, I propose *Multipriv* to cover the multifaceted issue of privacy decision-making. It encompasses the following:

- An environmental layer that sits on top of the framework and includes the judicial, cultural, and economic factors.
- A user-specific layer is in the middle and centers around the sharer, their connection with multiparty members in the case of multiparty disclosure, and the tools readily available to them that can sway their decision to disclose data.
- A context layer, which is the most specific to the user's situation and includes contextual biases, audiences, etc.

0.2.3. Proposing a nudge-based system to mitigate disclosure

Centred around the user and the factors leading to disclosure, Multipriv offers the basis for the proposed nudge-based system designed for disclosure mitigation online. The aim is to reduce instances of oversharing personal data about oneself and others while acknowledging that users have a drive for disclosure. Considering these two seemingly contradictory concepts in tandem can allow us to reach a user-specific balance. Multiple components and design characteristics collaborate for this purpose:

- Domain knowledge encompassing all the general information that the system needs to push the nudges.
- Novel context-aware user model based on an adapted version of the *Rasch model*.
- A new way to represent privacy protection versus the drive for disclosure using multiple agents.
- *Mediation* between all the parties involved in both self-disclosure and multiparty disclosure issues.
- *Framed nudged* according to the user model with the purpose of increasing the acceptance rate of these interventions.
- A proof of concept platform called *Cognicy*, through which it was possible to compare my context-aware personalized nudges with their non-context-aware counterpart, corroborates the need for such a system, and highlights its potential in the real world.

0.3. Overview of the dissertation

This dissertation is broken down into eight chapters starting with the background research.

Chapter 1:*Background Research.*

This chapter sets the foundation of my work by diving into concepts such as behavioural biases, the privacy calculus, and the privacy paradox. It explains the notions that are used and referred to throughout the dissertation.

Chapter 2:*Related Work.*

In this chapter, I present the existing research on the subject including how other scholars have approached the topic of privacy preservation and the rise of nudge-based solutions. I also compare them and point out some of their advantages and limitations, which I aim to remedy.

My original work is detailed between chapters three and six as shown in Figure 0.1.

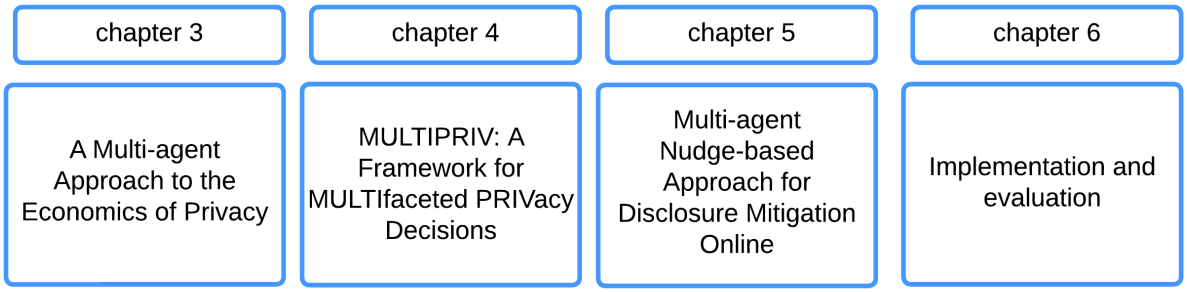

Fig. 0.1. Structure of the dissertation

Chapter 3:*A Multi-agent Approach to the Economics of Privacy.*

A multi-agent approach is proposed to offer an understanding of the economics of privacy. Each party is represented by an automated agent aiming to maximize its utility to reflect the real-life interactions between data owners, brokers, and data users.

Chapter 4:*MULTIPRIV: A Framework for MULTIfaceted PRIVacy Decisions.*

Multipriv is introduced as a novel user-centric framework capable of explaining the factors impacting privacy decisions on a personal and multiparty disclosure level. Going from least to most specific, its layers detail the elements encouraging or deterring the user from sharing personal content.

Chapter 5:*Multi-agent Nudge-based Approach for Disclosure Mitigation Online.*

Chapter 5 capitalizes on the findings in Chapter 4 and proposes a multi-agent nudge-based

system to tackle the issue of disclosure mitigation. It uses contextual parameters like motivation, biases, and the audience to customize the interventions.

Chapter 6:*Implementation and evaluation.*

A platform called Cognicy is proposed to evaluate the aforementioned system. It is designed after real SNS and offers the same functions with the advantage of being a controlled environment in which the efficiency of the context-aware personalized nudges can be fairly tested.

Chapter 7:*Conclusion.*

The conclusion summarizes the objectives that have been achieved in this dissertation and the challenges encountered along the way.

Chapter 8:*Future perspectives.*

Finally, the future perspectives chapter offers insight into what the next steps of my research will be based on what has been achieved and documented in this dissertation. It expands the horizon beyond the typical sense of privacy awareness to include studying the long-term impact of nudging and how my research can contribute to the field of fake news mitigation.

Chapter 1

Background Research

Social Networking Sites have long since surpassed the constraints of the virtual universe and stepped into the real world. Decisions that users make online have implications for their day-to-day lives. There is no shortage of instances where Ivy League universities such as Harvard rescinded offers to students because of past controversial Tweets [**29**]. Another example is that of a newly promoted employee who was discussing her salary increase on TikTok and got fired because her sharing behaviour caused her employer to distrust her [**30**]. The consequences of these decisions impact the individual's personal, professional, and social life alike.

However, making the correct choice and behaving in the least problematic way is becoming increasingly difficult. This is due to multiple reasons, the *first* of which is the ever-changing landscape of *Internet Technology* (IT) in general and social media specifically. Let us take TikTok as a case in point, by the time it became popular, users started complaining after learning that the company used their clips to promote the app on other platforms without their permission [**31**]. It was revealed that this is indeed stated in the user guidelines that TikTok is allowed to re-post the submitted clips and profit from the likeability of the content creators in any way that is deemed fit. The surprise amongst users stems from a lack of knowledge concerning their rights and who owns the uploaded videos.

This is connected to the *second* factor that makes decision-making hard, which is that people often do not have a sense of foresight and cannot properly estimate the repercussions of their choices. By the time they regret having posted something, it is already too late since the Internet does not truly forget. Minaei *et al.* [**32**] explain that sometimes deleting a post in retrospect makes users more vulnerable because malicious actors specifically signal this action as an intent to hide something that is damaging to the owner.

The *third* point is that individuals are susceptible to *cognitive biases* that can alter their perception and cause them to act rashly and without a thorough assessment. It is important

to consider this because even an intricate level of digital literacy and a rational tradeoff between gains and losses can be overridden in a specific context.

This chapter starts by examining the current situation in which privacy is more at risk than ever before. Then, it details the relevant recurring notions in this thesis such as self-disclosure and multiparty disclosure. Following this, I explain the concepts that are intertwined with privacy decision-making, namely the *privacy paradox*, behavioural biases, the economics of privacy, and the existing laws and regulations.

1.1. The current state of privacy

Information privacy is defined as the ability to control information about oneself and decide when and how much personal information can be collected and used by others. Although the digital age has sparked discussions of privacy, its roots go back as far as ancient Greek philosophies. Most notably, Aristotle makes a distinction between the public sphere of political activity and the private sphere associated with family and domestic life [**33**]. In modern times, the development of privacy protection can be traced to the 1890s, specifically, the famous essay by Samuel Warren and Louis Brandeis, which served as the first advocacy for *"The Right to Privacy"*.

With the technological development of the 20th century, discussions of privacy became more prominent. They centred around monitoring individuals and the fear of losing one's privacy. George Orwell's classic dystopian novel *1984* [**34**], which was published in 1949, epitomizes the growing unease and distress of a futuristic totalitarian state in which mass surveillance is the norm. The oppressive iron-handed party limits the freedom of speech and robs individuals of their right to privacy.

In 1990, the scope of the fifth principle of the *United Nations Guidelines for the Regulation of Computerized Personal Data Files* was focalised on the principle of non-discrimination. Personal data, according to this definition, included any "*information on racial or ethnic origin, colour, sex life, political opinions, religious, philosophical and other beliefs as well as membership of an association or trade union, should not be compiled.*". In today's world, this definition remains valid as one of two approaches to privacy.

The *first* of which considers it a means to an end and not the ultimate goal. In other words, by protecting it, we can eliminate or at least reduce discrimination, which is the real objective. The European Union's (EU) *General Data Protection Regulation* (GDPR) states that sensitive data must be regulated to avoid the risk of discrimination against vulnerable groups and individuals.

The *second* approach, according to Quinn *et al.* [**35**], views the protection of personal information as the true and ultimate aim since it is a fundamental right. On a global level, Article 17 of the *International Covenant on Civil and Political Rights* (ICCPR) protects everyone from arbitrary or unlawful interferences with their "*privacy, family, home, or correspondence*". On a regional level, the EU's *Charter of Fundamental Rights* (CFR) identifies data protection as a fundamental right. In Canada, in 2019, the *Office of the Privacy Commissioner* (OPC) declared that "*privacy is a precondition for citizens' other freedoms as well as a keystone right for democracy*"

Before the popularization of social media, the Internet was uncharted territory. The concept of privacy was not geared towards end users but rather countries, the military, organizations, and global businesses. The focus was not on educating the individual and providing them with protective tools except for antivirus. At the time, this approach made sense because hackers were thought to be honing their skills to target banks, energy companies, the healthcare sector, governments, etc. To be clear, such attacks are still taking place in the present as proven by various incidents. In Canada alone, companies are still paying nearly $7 million in data breach costs [**36**]. However, solely focusing on strengthening the software and hardware is hardly enough, especially with the proliferation of social media platforms. The definition of privacy as "*the right and ability to choose what to share, when to share it, and with whom*" is not that straightforward anymore. Prior to Facebook, Twitter, and Instagram, it was virtually impossible to have a complete picture of one's life unless the person chose to disclose it to close friends and family. Even then, it was mostly through verbal communication leaving no traceable footprint. Long gone are those days due to the growing number of websites and apps, in general, and social media platforms, in specific. Accessing them has never been as easy with the convenience of mobile devices.

This brings us to the situation today, in which privacy is in jeopardy. The rise of shadow profiles is one of the indicators of this aggravating crisis. The term describes the situation when users' and/or non-users' information is collected to be monetized without their consent. Using cookies, Facebook can track individuals who do not even have a social media account across all websites on which a Facebook "Like" or "Share" button appears, even if the user does not actively click on them [**7**]. It is not mere speculation since Mark Zuckerberg admitted that this is common practice in a congressional hearing in 2018. Somehow, people have grown accustomed to the various ways in which SNS infringe on their privacy and that of those around them. Another example of this is the continuous contact upload feature that is activated by default upon creating a new account. This means that Facebook can read the user's contacts and gain access to the photos on their phone, which have never been uploaded to the platform. While the company assures users that this is safe and that their

information is safeguarded, that can never be truly guaranteed, as proven by past breaches. Simply by signing into the platform and not taking the proper steps to deactivate the default features, a person can be jeopardizing multiple individuals. This includes themselves and their social circles. This has further ramifications that touch upon ethical issues as well.

On social media, a product seller can identify how much different users and non-users are willing to pay and subsequently, offer personalized prices based on their profiles [**37**]. Facebook is far from the only platform to adopt such dubious practices. In a report by CBS News, a senior technology reporter showed how TikTok tracks its users' likes, dislikes, and personal information, including email addresses, phone numbers, and WiFi networks [**38**]. In today's world, one cannot discuss privacy online without bringing up TikTok and its controversies. An investigative report by Forbes [**39**] found that ByteDance, the company that owns this platform, planned to use it to monitor the locations of specific American citizens.

The threats to privacy are magnified by poor digital literacy. The term is often mixed with *Information Communication Technologies* (ICT) competence [**40**], but there is a difference between the two. The latter term refers to a range of basic tasks like turning on a PC as well as extremely time-consuming and demanding ones like editing a film. Whereas digital literacy means "*the interest, attitude, and ability of individuals to appropriately use digital technology and communication tools to access*, *manage*, *integrate*, *analyze and evaluate information*, *construct new knowledge*, *create and communicate with others*" [**41**]. Hence, inadequate digital literacy results in poor awareness of privacy issues. The next section goes back to the basic categories of personal information.

1.2. Categories of personal information

Throughout this document, "personal data", "private data", and "sensitive data" are used interchangeably. The main question is: what makes a piece of information personal? Formally, the *Canadian Privacy Act* [**42**] defines personal data as any recorded information about an identifiable individual including:

- Education, medical, criminal, or employment history of an individual or information about financial transactions.
- Any assigned identifying number or symbol.
- Address, fingerprints, or blood type.
- Personal opinions or views except where they are about another individual or about a proposal for a grant, an award or a prize to be made to another individual by a government institution.
- Private or confidential correspondence sent to a government institution.

- The views or opinions of another individual about the individual.
- The name of the individual where it appears with other related personal information or where the disclosure of the name itself would reveal information about the individual.

The regulations of the European Union identify personal data as any identifiable information such as name, identification number, location data, online identifier, physiological, genetic, mental, economic, cultural, or social identity [**43**]. Since the focus of this dissertation is the user's privacy and providing support for better decision-making, it is relevant to discuss the impact of disclosure on the individual. Regret as the consequence of sharing personal data is well-documented. The findings of Wang *et al.* [**44**] link the highest level of regret with the following pieces of data: Personal and family issues, work and company, religious issues, religion and politics, alcohol consumption and illegal drug use, sexual content, strong negative emotions, attacks on individuals, attacks on collectives, lies, and secrets. Chapter 2 further discusses the different approaches to measuring the sensitivity of the shared data. In the following, I move on from the data being shared to the forms of disclosure starting with self-disclosure.

1.3. Self-disclosure

It has become the norm for individuals to share their private information, often publicly, to enjoy some perceived benefit like personalized services [**45**]. Self-disclosure is the act of revealing data about oneself to others such as information about one's family, friends, colleagues, feelings, private opinions, aspirations, goals and aspirations, deep fears, and failures [**46**]. Potential consequences of oversharing include the risk of cyberbullying and other forms of victimization [**47, 48**]. The security implications of ubiquitous social media include cyberbullying and other forms of victimization. Sharing can become intensified during adolescence since increased reliance on peer relationships is associated with an increase in the rewarding aspects of self-disclosure [**49**]. With increased opportunities for extensive self-disclosure on social media, a culture of oversharing has emerged, in which users often post accurate identifying information about themselves with little regard to consequences [**50**]. Shared information on social media has the potential to be a source of privacy attacks for users. Ghazinour *et al.* [**51**] investigate and confirm the potential of an enemy gaining access to private information because of a single photo or video. People are familiar with the saying: "*If you are not paying for it, you are not the customer*; *you are the product being sold*.". Indeed, the data of social media users is highly lucrative for brokers and the e-commerce sector, in general, which is discussed throughout this chapter. Next, the different types of self-disclosure are detailed starting with prompted disclosure.

1.3.1. Prompted disclosure

I call this category "prompted disclosure" to illustrate any disclosure that was instigated by someone other than the sharer. This could happen if Alice receives an innocuous message from Bob asking for a piece of private information, which she might be inclined to share due to their close relationship. Another side of prompted disclosure is more nefarious and usually requires the use of persuasive techniques and *social engineering*. It is known as a type of manipulation that pushes people into taking a certain action or revealing specific information that the hacker can benefit from. Identity and personal data theft are often achieved through *phishing attacks*. They are more vicious and effective than ever, managing to bypass email filters, pull the wool over the users' eyes, and disguise their attacks as legitimate trustworthy actions.

In relation to social media, phishing is prominent, especially on Facebook, Instagram, and TikTok marketplaces. Typically, these attacks happen when users are about to make a purchase. They consist in emulating a genuine website such as a bank and asking the user to provide their private information to log in as seen in Figure 1.1, on the right. The hackers get the data through the fake website and in this example, they can gain access to the user's email or phone number and password associated with their PayPal account, which is connected to their bank. One of the telltale signs of the scam is the different URL, but it is quite subtle, especially to a user who is not paying attention. This issue is bound to become more problematic as according to a new report from *Accenture*, social commerce[1] is estimated to become a $1.2 trillion global market by 2025, accounting for 16.7% of total e-commerce spend [**52**]. A survey found that 49% of TikTok users have reported making a purchase after learning about the product or service on the platform [**53**].

Aside from phishing, personal data theft, and prompted disclosure, users' actions without any prompt tend to disclose a lot of private information. This will be tackled next.

1.3.2. Unprompted disclosure

Contrarily to the social engineering attacks, these threats are not incited by an external factor but that does not make them any less dangerous. In this subsection, the focus is on actions and practices that are detrimental to peoples' personal, social, and professional lives. One of these forms of unprompted disclosures that has become very popular on social media over the past decade is geolocation tagging. The alarming part about this practice goes beyond unaware laymen who cannot fathom the consequences of their actions, as in fact, many people do it despite being aware of the potential repercussions, sometimes out of habit as if it were muscle memory. Some of them do it to share that they are in a luxury resort,

[1] e-commerce on social media platforms such as Instagram, Facebook, TikTok, etc.

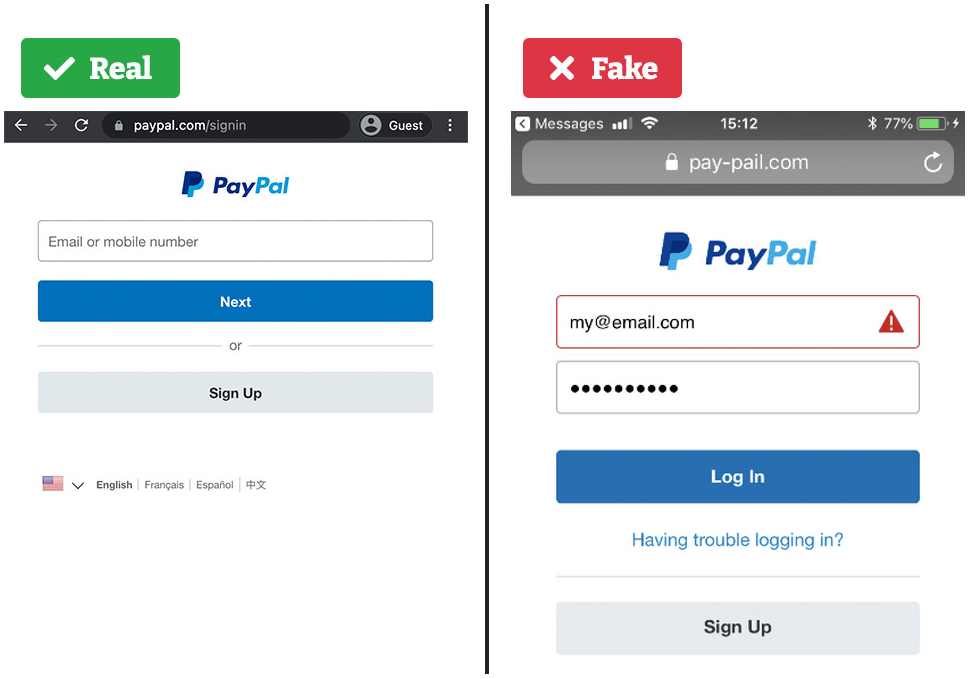

Fig. 1.1. Example of a phishing attack[2]

for example, and it reflects positively on their online social life and following. While others just adopt an apathetic negligent attitude about it assuming that no one would go through the trouble of harming them since they are not celebrities or politicians. An example of inattentive oversharing happened when a host of a TV program tagged a photo of his car without disabling the *Global Positioning System* (GPS) function and in the tweet, he said: "*Now it's off to work*" [**54**]. A geotag was embedded into the image pinpointing his home address. He, not only posted his location but also included that he would not be at home. As such, any thief can know the location and that he is away offering them the perfect opportunity to break in. He admitted that he was fully aware of the demerits of geotagging, but he did not think he would be a target.

In extreme cases, unprompted disclosure especially in photo and video forms can result in being used as a subject for deep fakes. Siwei Lyu, a researcher who is working with the defense department, said that anyone who posts photos on social networking sites like Instagram is at risk of being deep faked [**55**]. In fact, he specified that it only takes about 10 seconds of video to create a realistic deep fake. While most of the concern about deep fakes has focused on the spread of misinformation and the defamation of public figures, private citizens can also be targeted. Of course, this should not strike fear in everyone and lead them to never post anything on social media. Public figures aside, people can use some basic

[2]Source: `https://www.verified.org/articles/scams/craigslist-paypal-scam`

directives to limit the spread of their data such as sharing photos only with friends, not the general public.

Moreover, unprompted disclosure is not only discouraged because of malicious people but also the platforms themselves. In 2019, the US's *Federal Trade Commission* (FTC) filed multiple claims against Facebook stating that the social media platform uses deceptive practices that violate a prior commission order (United States v. Facebook, 2019) [**56**]. The argument was that the company continued to misrepresent the ability of users to control their privacy, as well as how much data it made available to third-party advertisers. Moreover, the FTC added that the platform's claims about collecting users' phone numbers to improve the security protocols are misleading. The numbers are allegedly also used for targeted advertising without making that clear to the users whose data is being sold. The issue extends beyond self-disclosure due to the social and interactive nature of SNS, which brings me to *multiparty privacy* issues.

1.4. Multiparty disclosure

While the decision to disclose information is up to a single individual who presses the share button, the content often includes people other than the user. Whether it is a photo taken on vacation with their family, a video recorded at their favourite coffee shop with their children, or a text post in which they tag their friend, the content often involves people around the sharer. What if Alice took a video while drunk at a party with Bob and shared it online without knowing that this might impact an upcoming promotion that he is eligible for? Alternatively, what if Bob re-shares Alice's post, which she originally set with the audience "friends", with his own social circles that Alice is unfamiliar with?

Having explained the main types of disclosure, there is an equal need to understand the decision-making process that users go through prior to the disclosure and this is where the *privacy calculus* comes in.

1.5. Privacy calculus

Laufer *et al.* [**57**] coined the term "*calculus of behaviour*" in reference to the cognitive process that underlies people's disclosure decisions. Simply stated, the individual asks themselves: "*If I am seen engaging in this behaviour or that behaviour or am seen with this person or that person, what are the consequences*?"[**57**]. The privacy theory states that individuals make decisions about self-disclosure by weighing the risks and benefits [**58, 59, 60**]. The outcome of this evaluation is based on the subtraction of the *perceived risks* from the *anticipated benefits*. If positive, the decision maker proceeds to disclose personal information disclosure and if negative, they exhibit less willingness to do so [**61**].

The perceived benefit factor is tied to the concept of perceived usefulness, which is a pillar of the field of *Communication Privacy Management* (CPM) called the *Technology Acceptance Model* (TAM) [**62**]. TAM states that there are two factors that determine whether a computer system will be accepted by its potential users: (1) perceived usefulness, and (2) perceived ease of use. Originally, TAM was designed for the workplace and has since been extended. When applied to the internet, a broader concept was established, which is the perceived playfulness, enjoyment, and pleasure [**63**].

In contrast, the construct of risk is defined as the perceived damage and the negative repercussions associated with the disclosure. The most direct way of calculating the risk is in monetary terms such as the expected financial loss from identity theft or from losing one's job. Other risks are not as quantifiable such as the degradation of a familial relationship or the public shame following the revelation of private data especially in certain cultures [**64**]. The privacy calculus has been used to explain users' behaviour on social media such as befriending strangers. The same affordances that facilitate relationship building can also create new privacy risks [**65**].

One of the main criticisms of the privacy calculus is that, even if we assume that the calculation is valid at all times (ignoring the impact of the context), the risks are hard to assess. One might have an easier time estimating their expected gratification by calculating the number of likes or re-tweets they got, but doing the same for the other factor is difficult [**66**]. Moreover, disclosure decision-making processes are mostly non-conscious [**67**] based on habits [**68**]. This includes employees who end up compromising their company's cybersecurity by performing actions that have become routine to them such as forwarding an email to a co-worker without double-checking the right address. Not only does the individual suffer the consequences in their professional and personal life, but the company also ends up reporting damages that can be irreparable. Direct financial loss from successful phishing increased by 76% in 2022 [**69**]. This behaviour also ties in with cognitive phenomena such as the anchoring bias, which will be further discussed in Section 5.2.3.

Another criticism of the privacy calculus is that history matters: someone who has been a victim of identity fraught for example is more likely to put emphasis on the cost than they did before the incident. On the other side of the equation, if a positive outcome is observed initially (getting a financial return on disclosing data, for example), the behaviour is likely to persist. Hence, the values assigned to risks and expected gains fluctuate. Thibaut *et al.* [**70**] highlighted this while arguing that social exchange influences the instigation of a relationship and addressed that "If good outcomes are experienced in initial contacts or if these contacts lead the persons to anticipate good outcomes in the future, the interaction

is likely to be repeated." Interestingly, the privacy calculus stands in stark contrast to the privacy paradox, which I will go through in the next section.

1.6. Privacy paradox

1.6.1. Arguing for the existence of the paradox

The privacy paradox indicates that online privacy concerns are not enough to explain the behaviours on SNS. Before this phenomenon got its appellation, researchers noted some discrepancies between privacy attitudes (or concerns) and peoples' disclosure behaviour. A study in 2002 led by Spiekermann *et al.* [**71**] was designed to investigate the inconsistency between the two in an electronic commerce environment. Their findings seemed counterintuitive at the time as participants who seemed protective over their privacy ended up forgetting about their concerns and indulging in oversharing. Once "inside the web", most of them did not live up to their self-reported privacy preferences. In fact, they displayed a surprising readiness to reveal private and even highly sensitive information. Wilson and Valacich [**72**] also focus on the irrational aspect of decision-making that cannot be fully explained using the aforementioned privacy calculus.

Then, the term privacy paradox was first used in an essay by Barnes in 2006 [**73**], after which, many researchers studied the phenomenon in connection with SNS. While the majority of participants voiced their heightened fear and anxiety about privacy infringements [**74, 75**], SNS use is ever-increasing and more private data is being voluntarily disclosed. [**76**]. The research by Barth *et al.* [**77**] investigated whether the existence of the paradox is solely due to the users' lack of technical knowledge or if it is a more complex phenomenon. The authors confirmed that despite the users' technical backgrounds and possessing a higher-than-average understanding of privacy intrusion possibilities, they remain unwilling to invest either the time and effort or the money necessary to protect their privacy.

However, despite the reported evidence supporting the existence of the privacy paradox, other researchers are either on the fence or completely reject its foundation.

1.6.2. The myth of the privacy paradox

Some scholars deny the existence of the privacy paradox [**78, 79, 80**]. Solove even calls it a myth and completely rejects the notion [**81**]. Furthermore, The findings of Krasnova *et al.* [**82**] suggest that privacy concerns do, in fact, translate into the amount of self-disclosure on social media platforms. This is corroborated by the work of Baruh *et al.* [**83**], which shows, through a meta-analysis of data, that users with higher perceived risk had a lower intention

to disclose private data (r = -0.18[3]) and ended up sharing less personal information (r = -0.14). Higher expected rewards have also been associated with an increase in self-disclosure, which supports the privacy calculus and stands against the paradox [**80**].

Moreover, most of the studies that support its existence are based on self-reported data that may involve biases or inaccurate estimations [**84**]. So, if the concerns are subject to doubt, how can we be sure that the discrepancy exists, to begin with?
Maybe the concerns and behaviour are actually in alignment, unbeknownst to the researchers. Solove states that it is not a paradox at all because it hinges on the misalignment of the individual's preferences with their actions, and this in itself is a wrong perspective. The argument here is that most scholars compare context-specific actions with concerns that were expressed in a general sense (out of context) and as a result, one cannot conclude that there is a discrepancy, to begin with. Reducing privacy issues to a matter of "disclosing versus not disclosing" marginalizes the actual issue. It equates sharing with not caring about one's privacy regardless of the circumstances, whereas, in reality, it is a more complex issue. It cannot be studied in a comprehensive manner without posing questions such as: In what context do people tend to share personal data? With whom do they do it and for what reason? Is it out of their own volition or is it due to compulsory factors?

Furthermore, privacy decisions are intertwined with trust. An individual can have concerns but choose to share data because they assume and believe that it is not going to be leaked intentionally or unintentionally. For example, in Canada, 89% of people used online banking as of last year [**85**], but that does not mean that they are freely giving up their data to corporations. In a report prepared for the Office of the *Privacy Commissioner of Canada* (OPCC), the vast majority of Canadians (81%) have expressed a fair amount of trust that banks will protect their personal information [**86**]. So, it would be disingenuous to use the fact that individuals behave this way after expressing their concerns as a supporting argument for the existence of the paradox.

To conclude, there are two different perspectives on the subject, each of which is supported by renowned researchers and experts in the field. Beyond proving or disproving this notion, the belief that *behavioural biases* impact privacy decision-making is more widely held.

1.7. Behavioural biases

Behavioural biases are irrational beliefs that can unconsciously guide and influence peoples' actions. They are often at play in the decision-making process leading to self-disclosure. These biases are generally considered to be split into two subtypes: *cognitive biases* and

[3]r is the pairwise correlation coefficient

emotional biases. Even scholars who are proponents of the privacy calculus over the privacy paradox acknowledge the major role of emotions in the decision-making process [**87, 88, 89**].

Being susceptible to cognitive biases is the result of having strong confidence in concepts and practices that may or may not be accurate. If a user mistakenly believes that the data they are disclosing is not private and makes the decision to share it, this is a manifestation of a cognitive bias. On the other hand, emotional biases typically occur spontaneously on the spur of the moment fueled by feelings and the current context such as the specific time, location, and circumstances. They are not usually based on reasoning, preconceived ideas, or deeply held beliefs. An example of this would be if Alice receives a frantic call from her distraught friend Bob asking her to transfer money for an urgent situation. Alice might be privacy-conscious under normal circumstances, but her friend's extreme distress can make her feel equally anxious and in certain situations even reveal private data. Hence, this is not a matter of lacking knowledge and is instead a contextual lapse of judgement.

The importance of examining these biases in connection to my research subject is twofold. First, people's innate biases impact their privacy decisions and as such, social engineering tactics can exploit them to incite a specific action. This is not a dystopian future as we have all already witnessed the case of Cambridge Analytica. Their use of psychological targeting gained infamy during the 2016 US presidential election in which they mined the profiles of millions of Facebook users to target them with psychologically tailored advertising [**90**]. Second, these same biases can prove to be a force for good if wielded for the right reason and with users' consent, which is something that I explore in Chapter 5 as part of my approach.

1.7.1. Emotional biases

In 1980, Zajonc [**91**] argued that individuals' emotional responses to stimuli are more potent than their cognitive biases when making decisions. This has been corroborated by decades-long extensive experimental research in the field [**92, 93, 94, 95, 96**]. The individual's positive or negative emotions have been linked to their perception of benefit and risk and as a result, their inner privacy calculus [**87**]. The former type of emotion increases the perceived reward and diminishes the concerns, which ends up favouring disclosure over withholding private information. One of the contextual biases that contribute to this effect is the *availability bias*. It illustrates the human tendency to rely on information that comes readily to mind when evaluating situations or making decisions. If the user finds themselves in a situation in which they take a selfie, for example, then, they recall how they felt good about getting many likes for a similar post recently, they might do it again. Experiments [**97**] have shown that a reminder of their privacy concerns has the potential of pulling the user out of this effect and making them warier of the disclosure.

Another type of emotional bias is the *present bias*. It manifests itself in the lack of forethought and tunnel vision on the present that a person can experience leading them to make hasty decisions. There is a tendency to discount future risk in favour of immediate rewards [**4, 66**]. *Default bias* or *status quo bias* refers to the tendency to gravitate towards the highlighted default choice because it is often perceived as the best normatively approved recourse [**98**]. Moreover, deviating from such an option requires knowledge, effort, and responsibility on the user's side. When people make dozens of choices on a daily basis, it is unlikely that they can and would afford to give each one of them appropriate reflection time.

To summarize, emotional biases have been documented since before the internet as a factor in decision-making in general, and in the domain of privacy, in particular. However, this is not the sole type as cognitive biases tend to play a major role in the process.

1.7.2. Cognitive biases

Social engineering attacks are notorious for targeting biases to manipulate their victims. *Cialdini* [**99**] argues that all influence schemes fall into one of six categories: *scarcity*, *reciprocation/reciprocity*, *consistency/commitment*, *authority*, *social validation*, and *friendship/liking*. First, scarcity is grounded in the psychological principle of *loss aversion*, which states that we assign greater value to things that are in limited supply [**100**]. One of the main ways through which scarcity is used to promote self-disclosure behaviour is when companies launch limited edition products, which play on the *Fear Of Missing Out*(FOMO) frenzy that many people are susceptible to. Combining this with social engineering pushes users to disclose their private information to fraudulent parties while seeking the gratification that comes with exclusivity. The scarcity-based tactics are not solely commerce-specific, but also, target individuals seeking employment, especially in times of economic recession. Following the increasing layoffs and bankruptcy filings due to the COVID-19 pandemic, these scams were on the rise [**101**]. As it turns out, while most people are automatically suspicious of job pitches via unsolicited emails, they are less cautious of posts on job sites like LinkedIn, or even Facebook. When users do not feel singled out and personally targeted, they tend to lower their guard and fall victim to the false veneer of legitimacy that is carefully crafted by malicious individuals and organized groups. When times are tough, people are more likely to ignore what might otherwise seem like obvious red flags, said AJ Nash, vice president of intelligence for the cybersecurity company ZeroFox [**102**].

Second, one of the most widespread and universal norms in human behaviour is reciprocation [**103**]. In a general sense, people feel obliged to pay back what they receive from others in the form of actions, monetary values, gifts, or services. The most intriguing part

is that this can result in an unequal exchange favouring the person who initiated the process. Most people do not feel at ease when indebted to someone else and may make haste to pay them back with a substantially larger favour than the one they received, to begin with. In a social media setting, if Bob opens up to Alice about hardships in his life, whether they are personal, familial, or professional, she feels the need to reciprocate by disclosing her own experiences. Altman and Taylor's *Social Penetration Theory* (SPT) [**104**] draws a similar conclusion, which is that reciprocity is necessary for relationships to prosper. This is a theory that makes predictions about relationship development using a cost-reward model. SPT compares people to a multilayered onion as seen in Figure 1.2. As the interpersonal relationship blooms and individuals get to know each other, they shed the layers leaving their core vulnerable.

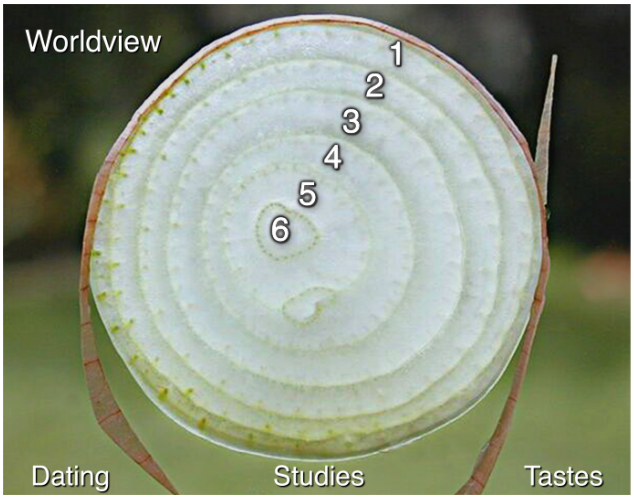

Fig. 1.2. Layered architecture of the social penetration[4]

Third, consistency manifests itself as follows: Once an individual commits to a course of action, they are unlikely to change their mind. There is a tendency to continue pursuing an endeavour, which gets stronger the further the person is down that road. It is often referred to as the *sunk cost fallacy* and is generally associated with money investment. On SNS, if Bob is used to sharing his political opinions publicly on Facebook and posting controversial content, he is unlikely to deviate from this track unless forced by an external factor such as being banned on the platform or experiencing a major life event.

Fourth comes authority. People want to make the right choice and in general, tend to trust experts in the field as they hold authority over the subject. For a very long time, news segments have used the term "experts say" followed by a statement to encourage the general public to adopt a specific behaviour. On social media, this has been used countless times in a deceptive manner, which was the case with the international company called *ILikeAd*

media [**105**]. It ran an advertisement on Facebook claiming to have experts on board who can provide advertising and marketing services to small businesses on the platform. What actually happened is that they made use of the false authority they portrayed to infect victims' accounts with malware. Once installed, the company would collect Facebook login credentials from the victims' browsers and access their accounts. This is yet another case highlighting the need to educate the user because, at the moment, the platform is not capable of perfectly eliminating these schemes. In fact, the company used a cloaking method to hide the landing page of the advertisement rendering the automatic detection and flagging of the website highly improbable.

The Fifth concept is social validation. Basically, people's privacy decisions are not completely driven by rationality and are instead influenced by numerous heuristics, such as knowing others' privacy decisions [**106**]. This is a psychological phenomenon by which people tend to conform to the actions of others within a group. It is a cognitive bias because there is a strong underlying assumption that cannot be verified, which is that the majority knows best. If Alice receives a message stating that "90% of her friends have applied for this online service and have rated it 5 stars", she might be a lot more willing to respond and fill out the form sent to her. This example highlights how even the false resemblance of a majority ruling (maybe none of her friends even know of the service) can persuade a person one way or the other.

The sixth and final principle of persuasion according to Cialdini is friendship/liking. It stipulates that, when you have someone with whom you share a close relationship such as childhood friends, you are more likely to comply with their request. In general, this applies to almost all areas of our day-to-day lives. Close connections are the gateway to asking favours and making requests that one would never ask of a stranger. If I apply this to privacy, hackers impersonate someone close to their potential victim when asking for personal information. They are leveraging the importance of the history between the victim and the impersonated user and the former's aversion to upset or anger the latter.

Aside from Cialdini's principles, other cognitive biases are prevalent in the privacy decision-making process such as the *anchoring bias*. It describes people's tendency to rely too heavily on the first piece of information (anchor) and base their judgement on it. Regardless of the accuracy of that piece of information, the anchor is used as a reference point as if it is an established fact. This disproportionate reliance on previously available information has been documented by many researchers. Ariely *et al.* [**107**] recruited participants and proceeded to ask them 2 successive questions: first, what the last two digits of their

[4]Source: `https://commons.wikimedia.org/wiki/File%3AOnion_slice.jpg`

Social Security Numbers (SSN) are and second, what is the estimated price of some products? Although seemingly completely unrelated, people made estimates close to the last two digits of their NAS. In the context of self-disclosure, Chang *et al.* [**108**] investigated the impact of this bias on sharing decisions by using mature selfies as an anchor. Indeed, the answers showed a strong connection between being exposed to these photos and disclosing more private information.

Finally, I would like to discuss another type of category adjacent to cognitive biases, which is logical fallacies. I included it in this subsection as it is not an emotional bias and tends to come from an erroneous belief similar to cognitive biases. The most used phrase in this context is "*I have nothing to hide*" or "*Why should I be worried as long as I have done nothing wrong*?". In response to this, Solove argues that this fallacy has an underlying assumption, which is that *privacy is about hiding bad things* [**109**]. Hence, people who believe this would equate not doing anything bad with not needing privacy protection. If every minute detail of everyone's lives is known by everyone else or even to a single person or government department, what becomes of basic rights then? This raises the curtains on an era of scrutinized and closely surveyed human behaviour [**110**] that can, in the worst-case scenario result in forced obedience and puppeteering. This poses the questions that the next section will tackle: Who is the puppet, who are the puppeteers, and what do they stand to gain?

1.8. Economics of privacy

Up to this point, I have discussed privacy from various perspectives and how it is perceived depending on the person, context (and the biases it presents), and disclosure circumstances. This subsection is dedicated to explaining the role data plays as a valuable commodity in the digital age. I will tackle this subject from the perspective of the *data owner* whose private information is being commercialized, the *data brokers* (or collector hubs), and the *data users* (or 3rd parties) who acquire said data for various reasons. It is worth mentioning that another core part of the economics of privacy paradigm is the laws and regulations, however, they will be discussed separately afterwards.

First, starting with the data owners, they are often referred to as consumers because, in general, their information is collected while they are seeking a service or buying goods. Individuals value privacy both as an *intermediate good* and as a *final good* [**111**]. In economics, an intermediate good is a partially finished product that is used as an input or raw material in the manufacturing of other commodities that become final goods [**112**]. If we focus on privacy, this is an example in which the same piece of data can be both final and intermediate in that order:

1) If a customer's home address is valuable in itself (used to push advertisements) then, we can consider it a final good that third parties would seek to acquire.
2) If the same address is used as part of pattern detection for the purpose of behaviour prediction such as concluding the following: "people who purchased a specific product tend to live in the Zip Code 'H3T 1J4'", then, the data is intermediate. It is a piece of a bigger puzzle.

Second, let us move on to the data brokers or collectors. The most profitable aspect of their business is its nonrivalry [**113**]. This means that once consumers' data like preferences or addresses are shared with firm A, the data is still valuable to the other firms. One batch of information can theoretically be sold to an unlimited number of companies that want to maximize their sales. The same cannot be said about tangible goods. It is not an exaggeration to assume that data brokers know almost everything about people to an insensitive degree, at times. This was the case when OfficeMax, an American office supplies retailer ended up sending mail in which the name of the recipient read: "*Mike Seay*, *Daughter Killed In Car Crash*" [**114**]. OfficeMax got this information from a broker and proceeded to send out mail in bulk without any manual revision. This shows how much these collectors know about users and how little they care about disclosing the data. The major role they play is not entirely understood by the public. In fact, many people assume that medical information, in general, is protected under laws such as the *Personal Information Protection and Electronic Documents Act* (PIPEDA) in Canada and the *Health Insurance Portability and Accountability Act* (HIPAA) in the US. However, that is a misleading statement because while the communication between a patient and their doctor is indeed protected under such laws, other acts of self-disclosure are not. For example, if Bob searches for "*sugar-free candy*" and days later looks up "*Is insulin covered by RAMQ*[5]*?*", the websites that he consulted are under no obligation to refrain from collecting his data and eventually sell it for profit under the category "*diabetic*".

Cases of data brokers getting their hands dirty are far from scarce. *Epsilon*, one of the largest collectors and resellers of data in the world was sued and agreed to pay $150 million for facilitating elder fraud schemes in 2019 [**115**]. Epsilon admitted that, from July 2008 through July 2017, employees in its *Direct To Consumer* (DTC) unit knowingly sold lists containing the information of more than 30 million consumers to clients engaged in fraud. *Life360*, a popular tracking app, found itself in a similar position due to its nefarious practices. The company promoted its services that allow family members to keep track of one another especially parents concerned for their very young children. Ironically, their slogan

[5]The Régie de l'Assurance Maladie du Québec (RAMQ) is the government health insurance board in the province of Quebec, Canada.

is: "*the world's leading family safety service*". As it turns out, Life360 was selling data to roughly a dozen data brokers, some of which have sold data to US government contractors and basically anyone who is willing to pay the price [**116**]. It seems like quite the oxymoron putting together "family safety" and "selling precise locations" in the same context. To their credit, this service and many others claim to de-identify the data to make it anonymous. But in reality, it is very easy to re-identify the individuals with enough data, which is my next point.

Researchers from European universities have published their findings showing how they managed to re-identify 99.98% of individuals in anonymized data sets with just 15 demographic attributes [**117**]. One of the authors of the paper had previously worked on a study, which focused on credit card metadata and concluded that using just four random pieces of information is enough to re-identify 90% of the shoppers as unique individuals [**118**]. If this is coupled with the locations such as places where the users tend to shop or get a cup of coffee, one could infer not only who the individual is, but also their day-to-day lives and the places they frequent. Having access to location-based data, in particular, is very alarming because it makes de-anonymization much easier as proven in a report by CBS News [**119**]. The journalists contacted data brokers without revealing their real identities and managed to buy batches of information. No names or phone numbers were tied to the data, but it was not difficult to figure out to whom each phone belongs, based on where they spend their nights. The information they could extract went beyond identifying the address of the person, to include even tracking people as they move from room to room in their own homes. Someone's phone pinged their location 231 times as they left home for an hour to run an errand, thus, clearly showing their daily routine to anyone who acquires the dataset. This might come as a surprise, but when the journalists contacted someone they could identify to inform them of what information is being leaked about them, the person did not care. He verbatim said: "I'm not worried about it" and here lies one of the major issues in the field and the motivation behind this thesis: the sheer lack of awareness of the human in the loop.

Moreover, even when data brokers attempt to implement some standards to protect data (with dubious efficiency), there is no guarantee that this will be upheld by their partners. This was the case of *X-mode*, a collector that forbids the resale of raw data and only permits aggregated insights such as a group of devices belonging to sports fans. X-mode sold data to *NybSys* and the latter ended up reselling it to a firm called *LocalBlox*, which was already banned from X-Mode's platform in April 2020. A lawsuit was filed against NybSys stating that people's exact location data was sold through a chain of industry players, rather than an analysis of that information [**120**]. Some brokers are not even thought of as such as they have integrated their service into people's daily lives. This includes all the big credit monitoring

companies known in North America like *Equifax*, *Experian*, and *TransUnion*. Experian, for example, sold *Alteryx* access to a de-identified data set containing 248 attributes per household for 120 million Americans [**121**]. That is a lot more than the 15 attributes needed to re-identify individuals.
Next, I move on to the final party in the economics of privacy.

Third, the data users buy information from brokers and collectors. The main purpose is to personalize their services for existing customers or get in contact with potential ones. But who are these third-party data users?
Since they get access to data that can potentially re-identify individuals, one might think that they undergo rigorous background checks, but that is not the case. In an experiment run by a researcher at the Stanford School of Public Policy, 11 data brokers agreed to sell information that identified students by mental health issues including depression, anxiety, and bipolar disorder. The researcher did not even make a purchase and was offered free samples to see what they could expect. It was reported that one out of the 11 brokers made no demands on the information should be handled by the buyer and promised that it could offer names and addresses of people with "*depression*, *bipolar disorder*, *anxiety issues*, *panic disorder*, *cancer*, *post-traumatic stress disorder*, *obsessive-compulsive disorder*, and *personality disorder*, as well as individuals who have had *strokes* and data on their *races* and *ethnicities*" [**122**].

From a classical economic welfare perspective, it is generally considered beneficial and profitable for all of these three market parties to have access to the same information. *Information asymmetry* (the opposite of having the same information) can prevent everyone from moving forward, which is the stance of those who promote sharing personal information. In an article published in 1981 in *The American Economic Review* [**123**], Posner argues that consumers and firms should trade information to achieve a market outcome with free access to personal information. The author states that if information about Alice is forbidden to the merchant, substitute and even inferior sources will be utilized by them to infer her likes, preferences, etc. Following this reasoning, it is in Alice's (data owner) best interest as much as it is in the seller's (who can be seen as the data user) to make the data available. Posner stated that "*exhaustive information costs more than it is worth*", which should serve as some reassurance to people who are concerned about their privacy. However, that statement was made in pre-Internet time. As explained before, batches of data can be sold and resold and are definitely worth more than they cost. In today's world, some calls for a free data marketplace are still echoing promising mutual benefits for everyone. However, most of these proponents have a stake in the game and are benefiting from the lax restrictions they are afforded. This brings us to the laws and regulations.

1.9. Laws and regulations

The regulatory framework operates in fundamentally different ways between Canada, the US, and Europe. According to Stephan Grynwajc, a Canadian privacy and data protection attorney, each of them has a different perception of privacy that is evidently reflected in their laws [**124**]. In Canada, privacy protection is focused on individual autonomy and controlling one's information. In the US, privacy protection is essentially about protecting liberty. For Europeans, privacy protects the dignity or the public image of the person. I will leave the assessment of the Canadian laws until the end of this section because it is a middle ground between the US' and the EU's approaches.

In the US, concerns of "*big brother*"[6] are still resonating with people, and privacy infringements are mostly perceived as the actions of governments.
Currently, the US does not have a single law that covers the privacy of all types of data across all states. "*Historically, in the US, [they] have a bunch of disparate federal [and state] laws*" said Amie Stepanovich, executive director at the *Silicon Flatirons Center at Colorado Law* [**125**]. There is a mix of laws specific to one sector like health care or finance that go by acronyms like *HIPAA* (Health Insurance Portability and Accountability Act), *FCRA* (Fair Credit Reporting Act), *FERPA* (Family Educational Rights and Privacy Act), and *ECPA* (Electronic Communications Privacy Act). Even these laws that seem to fully cover one field can be outdated and lacking such as HIPAA, which only includes communication between the user and "covered entities", which include doctors, hospitals, pharmacies, and insurers. Thus, using a Fitbit[7] is not covered, for example. The only states that have comprehensive consumer privacy laws are California (*California Consumer Privacy Act* known as *CCPA* and its amendment is *CPRA*), Virginia (Virginia Consumer Data Protection Act known as *VCDPA*), and Colorado (Colorado Privacy Act known as *ColoPA*). Some initiatives have been taken in this direction, but they never saw the light of day.

In 2014, the *Federal Trade Commission* (FTC) released "*Data Brokers: A Call for Transparency and Accountability*" [**126**]. This document was published based on a study of nine brokers and it highlighted the lack of transparency in the industry. It was a call for Congress to pass legislation to limit the reach of brokers and offer users some control over their data. In 2016, the US passed a landmark privacy law that forced service providers to tell consumers what data they collect and why, as well as to take steps to notify them of data breaches [**76**]. Under the *Federal Communications Commission*'s (FCC) new rules, consumers may

[6]Big Brother is a fictional character and symbol in George Orwell's dystopian 1949 novel Nineteen Eighty-Four. He is ostensibly the leader of Oceania, a totalitarian state wherein the ruling party, Ingsoc, wields total power "for its own sake" over the inhabitants.
[7]A Fitbit is a fitness tracker that uses an accelerometer to measure users' movements.

take control of their own privacy and forbid Internet providers from sharing any personal information they deem sensitive. This includes apps, browsing histories, and location data. However, less than a year later, these regulations were rolled back before they even went into effect, granting brokers absolute freedom to collect, store, share, and sell consumers' data [**127**].

Europe's approach is on the opposite side of the US' sectoral division. The *Council of Europe*, established in 1949, developed a comprehensive, principle-based approach to privacy. This was particularly critical at the time since it followed the tragic aftermath of World War II in which data was misused to target specific groups of people. Following the expansion of the internet and the deployment of numerous social media platforms, more up-to-date regulations had to be created to adapt to the situation. On the 25th of May 2018, the GDPR went into effect and widened the scope of personal data to include online identifiers, such as IP addresses and biometric data. Moreover, it gives individuals the right to control their own information, which despite seeming evident, was and still is quite the revolutionary insurance for internet users. This made GDPR the strongest privacy and security law in the world [**128**]. The aforementioned rights are the users' *right to access their personal data*, the *right to have their data corrected or erased* (*right to be forgotten*), and the right to *data portability*[8]. To avoid any claims of plausible deniability following a potential privacy breach, all companies that collect or process personal data in the EU are required to appoint a *Data Protection Officer* (DPO) to ensure compliance with the GDPR.

Furthermore, there is a transparency requirement with data owners whose information is being handled since organizations need to report data breaches to the relevant authorities within 72 hours and to directly inform the affected individuals if the incident poses a high risk to their rights and freedoms. Less severe non-compliance cases can result in fines of up to 2% of global annual revenue or €10 million (whichever is greater). The more serious violations can result in a fine of up to 4% of a firm's annual revenue or €20 million, depending on what is higher. These regulations clearly put restrictions on data brokers and the potential misuse of private data.

In Canada, on a federal level, privacy protection can be divided into two categories: *commercial* and *governmental*. PIPEDA applies to the collection, use, and disclosure of personal information in the course of commercial activities. For protection from federal government institutions, the Privacy Act was put in place to regulate the collection and use of individuals' data. The laws around private sector organizations tend to take the same sectoral route as

[8]The right to data portability allows individuals to obtain and reuse their personal data for their own purposes across different services. It allows them to move, copy or transfer personal data easily from one IT environment to another in a safe and secure way, without affecting its usability.

the US. This includes the *Quebec Act* and Alberta's *Personal Information Protection Act* (PIPA). There are industry and sector-specific regulations such as the *Personal Health Information Protection Act* (PHIPA) and the *Bank Act*. The *Office of the Privacy Commissioner* (OPC) investigates complaints under PIPEDA and has the authority to conduct an audit.

The GDPR, US sectoral laws, and PIPEDA are aligned in numerous respects. All of them impose obligations, covering, in particular, the most sensitive fields such as health and finance. Moreover, both GDPR and PIPEDA define personal data in very similar terms. However, there are a few differences. According to GDPR, the most severe incidents are fined at least €20 million, while Canadian privacy laws generally provide for fines of up to $100,000 per violation. Furthermore, while non-profit organizations and political parties in Canada do not have to abide by PIPEDA[**129**], GDPR does not exempt any firm or group that collects data even for non-commercial use [**130**].

Despite the existing regulations, Canada, the US, and a few European countries are on the list of the top 10 countries ranked by the number of breaches per million people [**131**]. One might wonder, are the laws obsolete because of the rapid proliferation of social media platforms and the ever-changing landscape of the virtual universe?
While that definitely does play a role in some cases, there is a clear reluctance to implement solutions because some of these issues date back to almost two decades ago. The re-identification of de-identified data explained in Section 1.8 was first reported in 2006. The incident occurred when, *AOL Research*, now merged with *Yahoo*, released a text file on one of its websites containing twenty million search queries for over 650,000 users over a 3-month period [**132**]. Amongst them, user no. 4417749 was identified as Thelma Arnold, a 62-year-old widow who lives in Lilburn [**133**]. Thus, the issue is by no means novel, it has simply been ignored for a long time.

Just because the law exists and there are formal bodies to enforce it, there are no guarantees that it will be upheld. In an investigation of the privacy policy landscape after the GDPR, it was revealed that there are many European websites do collect visitors' personal data without consent even when the individual has explicitly opted out [**134**]. They deploy deceitful nudges to push people to share by making it the highlighted easy option [**135**].
What makes the situation worse is the conflict of interest between data brokers and policymakers especially in the US. In 2020, collectively, data brokers' spending on lobbying rivaled the spending of big tech firms like Facebook and Google [**136**]. On top of that, politicians rely on data acquired from brokers for their campaigns. According to a report by the Washington Post, the *Republican National Committee* (RNC) took pride in the fact that it has more than 3,000 data points on every voter. The judiciary system is also entangled in this

web of intertwined interests. The government typically needs a warrant to collect data about a person without their consent, but if information was acquired through collectors without using force, there are no legal issues. Agencies from the *Federal Bureau of Investigation* (FBI) to *Immigration and Customs Enforcement* (ICE) obtain data from brokers without notice, public disclosure, or any oversight in order to carry out everything from criminal investigations to deportations [**137**].

1.10. Conclusion

"*Man is by nature a sociable animal, an individual who is unsocial naturally and not accidentally is either beneath our notice or more than human*", said Aristotle, the renowned Greek philosopher. There is no denying the need for socialization and sharing personal experiences, preferences, etc. This is becoming a reality with the growing use of social media and its accessibility to the younger generations. As such, quitting these platforms is not a viable means of privacy preservation. Only through mitigation and awareness raising can this issue be approached.

This chapter started by discussing a brief history leading to the situation today. After that, it divided the action of sharing private data into two categories: self-disclosure and multiparty disclosure. Both of these rely on a decision-making process, which is the privacy calculus, according to many scholars. However, facing the discrepancies between the user's concerns and actions, one school of thought adopted the notion of the privacy paradox. It was used to explain the seemingly irrational behaviour of humans that appears to contradict their best interests. Another school of thought refutes the notion of the paradox and criticizes the research used as the foundation for its existence. One of the voices supporting this position is Solove who wrote the article "The myth of the privacy paradox" in 2020 [**81**] denying the need to understand the phenomenon as there is nothing to understand if it is a misconception. I provide the supporting arguments for both parties, after which I move on to the behavioural biases that affect decision-making, regardless of whether the paradox truly exists or not.
The economics of privacy section focuses on the role each party plays from the individual to the data brokers and ends with third-party data users. I conclude this chapter with the laws and regulations meant to highlight the positives such as the strict GDPR that prioritises people over companies. I also shed light on the negative side that is yet to be addressed and some of the reasons why we have not reached a better situation in today's world. The next chapter focuses on disclosure mitigation.

Chapter 2

Related Work: Disclosure Mitigation in the Digital Age

The *National Institute of Standards and Technology* (NIST) of the US [**138**] and the European Union's General Data Protection Regulation [**139**] defined any physically unique, psychologically expressive, cultural, social, biometric, genetic and health data as *Sensitive Personal Data* (SPD) or *Personally Identifiable Information* (PII). There is a growing body of research focusing on protecting PII and pointing out the risks associated with mishandling such information intentionally or unintentionally such as stalking, identity theft, price discrimination, or blackmailing [**76, 140**].

To counter this, two main approaches garnered the interest of researchers. *One-size-fits-all solutions* are simple to implement and can be effective in some cases such as encouraging all users to create strong passwords and to refrain from visiting unsecured websites. However, in more complex situations, they can be less productive, especially on SNS platforms. Detecting disclosure, assessing risks, and convincing individuals to adopt a specific behaviour is more challenging. Thus, *personalized solutions* emerged to customize the user experience by considering their preferences. This is not without its disadvantages as the gain in terms of user satisfaction comes at the cost of potential *privacy leakage*. The term refers to the accidental disclosure of supposedly insignificant pieces of information, which, when grouped together can reveal very sensitive data. Since they are more lenient than the one-size-fits-all approaches, privacy issues are more likely to result from the former than the latter. Neither is definitively better than the other, however, the focus on personalization seems to be growing, which leads us to one of its pillars: preference elicitation.

2.1. Privacy preference elicitation

Understanding privacy preferences is of great importance to politicians [**141**], healthcare providers [**142, 143**], advertisers [**144**], economists [**145**], and business managers [**146**] for a variety of reasons. Navigating this topic while considering privacy and the ethical considerations that arise from collecting information about users is an ongoing field of research [**147**].

A key point of providing personalized support for privacy decisions, which this chapter will elaborate on, is eliciting the individual's preferences [**148**]. Based on this, the system can proceed to *user modeling*, which is a subdivision of human–computer interaction focused on describing the process of building up and modifying a conceptual understanding of the user. In the literature, there are various uses for user preferences as seen in Figure 2.1 [**149**]. This includes the automatic completion of online profiles and the prediction of preferences on a social network, for example. The external visualization is the most in tune with my work as it relates to self-awareness, control of profiles, and privacy visualization, all of which my approach touches on.

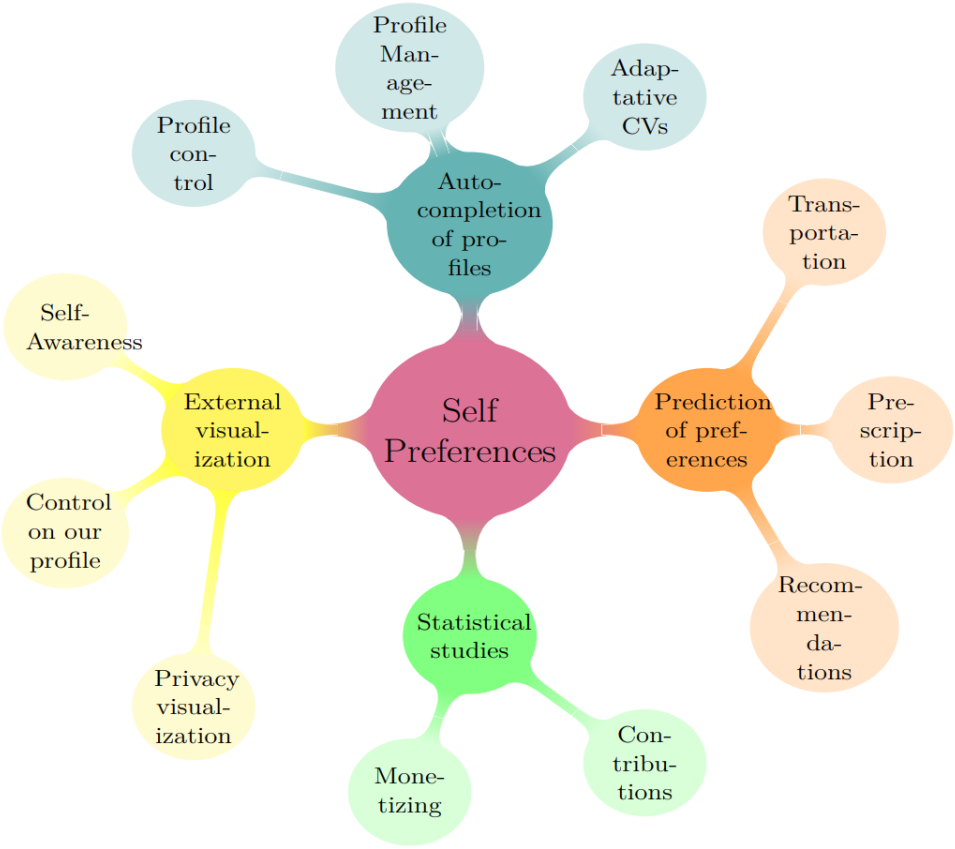

Fig. 2.1. A Non-exhaustive list of uses of self-preferences [**149**]

Personalization and privacy can appear to be opposing themes: personalization is based on personal information gathered through mining user data for purposes such as targeted advertising. However, they can converge in privacy-aware or privacy-preserving recommender systems, which prioritize safeguarding the collected data from unauthorized third parties. One of which is *Alambic*, which was designed by Aïmeur *et al.* [**150**] as a means of protecting both customers' and vendors' best interests on e-commerce platforms. The focus of this thesis is not achieving privacy preservation "despite the need for personalization", it is using the latter in favour of the former. This idea will become clearer in the next chapters.

Generally, preference elicitation can be an offline static process or an online adaptive task. In the former case, the user can either answer a questionnaire and self-report their valuation [**151**] or grant access to their history for an implicit elicitation. I summarize this in Figure 2.2. The study by Carbone *et al.* [**152**] is an example of direct preference elicitation in which 552 participants filled out a survey designed to investigate their psychological motivations for sharing data. The recruited people responded to questions about whether they remembered a situation in which they were very eager to share with others and if that motivation led to the actual disclosure.

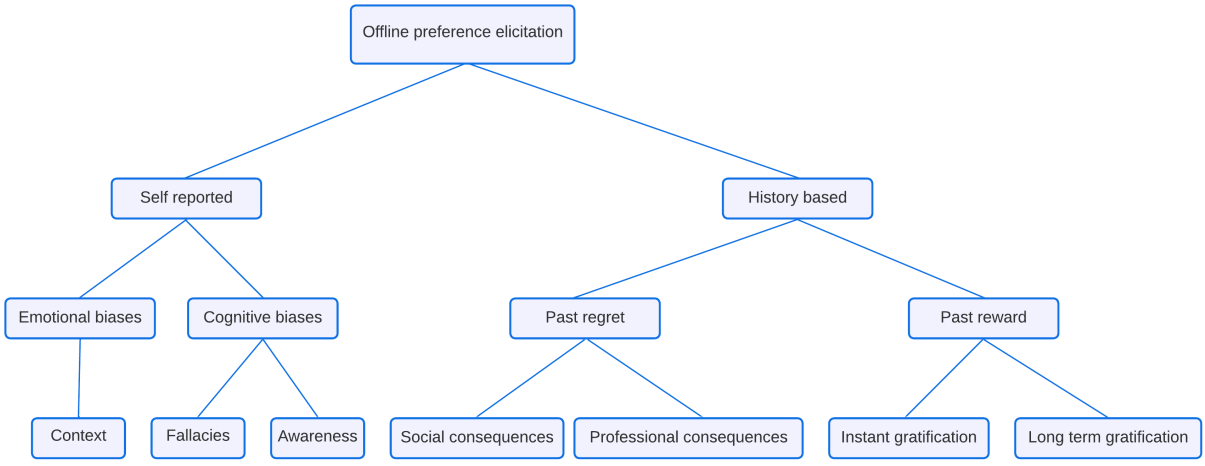

Fig. 2.2. Offline preference elicitation

For the self-reported approach to be valid, the questionnaire needs to be realistic and leave as little room as possible for users to provide answers influenced by cognitive biases. The scenarios should put the participant in the mindset of someone who is experiencing that situation [**153**]. The classic research of Westin [**154**], measures privacy with a one-time survey including questions to determine how much accessibility the respondents think is important and how much they value the ownership of data and their control over it.

The case of implicit elicitation falls under the umbrella of the *revealed preference theory*, pioneered by economist Paul Anthony Samuelson in 1938 [**155**]. At the time, the theory was specific to shopping behaviour and it suggested that the preferences of consumers can be revealed by their purchasing habits. This approach to preference elicitation assumes that the observed individual is both rational and consistent. The former means that the person would not deviate from the choice that maximizes their benefit or perceived utility. The latter refers to not deviating from an established pattern, which can be demonstrated in the following example:
1) Alice thinks that her home address is more sensitive than her phone number.
2) She also thinks that her phone number is more sensitive than her email address.
The consistent conclusion based on 1) and 2) is that the order of sensitivity is:
home address > phone number > email.
There should be no situation in which Alice would face a decision to share either her share home address or her email address and end up disclosing the former.

The final preference elicitation approach is done online and it is based on monitoring and analyzing the user's current interactions with the platform. It is very similar to its implicit counterpart because the user cannot keep providing direct feedback continuously in real-time. The difference between the two is that, in this case, it is a stream instead of a static bundle of data. Here is an example of this: Alice uses Amazon to buy a coffee machine, and then, the system captures this information and recommends her ground coffee. Had it been an offline process, she would have only seen the recommendation after logging in next time and not in the same session.

Some scholars have focused on identifying the predictors of user preferences. Taylor *et al.* [**156**] studied the willingness to share data amongst people from different cultures and concluded that trust leading to disclosure depends on their upbringing and background. The main core of the research on this topic has been geared towards preference modelling on mobile apps, specifically. This comes after a growing trend since 2017, which shows that users tend to access the internet using apps on their phones more often than on their computers [**157**]. As the popularity of these GPS-equipped devices grew, so did the focus on location preferences in particular.

The main challenge of preference elicitation is that human tendencies are time-dependent [**158**], circumstance-specific [**159, 160**], and driven by the conscious and the subconscious [**161**]. This ties in with the discussion of the privacy calculus, paradox, and biases in Chapter 1. Incorporating some of these aspects in a timely manner to respond to a critical issue

like privacy preservation can be a complex subject [**162**]. To tackle this, let us first address privacy modelling before moving on to the existing approaches.

2.2. Privacy modelling

SNS users are threatened by numerous issues such as having their data de-anonymized [**163**], their profiles autocompleted across multiple platforms [**164**], and their sentiment analyzed for potentially iniquitous objectives [**165**]. In this section, the terms *threat level*, *privacy risk*, *privacy score*, and *information leakage* are all considered to be part of privacy modelling. In general, privacy leakage refers to an intentional or accidental disclosure of information that exposes sensitive details about one's identity. The word "accidental", in this context, does not simply refer to a mishap in which a person sends the information to the wrong recipient, for example. If Alice re-shares Bob's post and it ends up reaching an unintended audience, this is a form of privacy leakage. Many privacy issues on social media are caused by having an unaware entourage.

Tackling the subject of privacy scoring often starts with processing the user's input and detecting the disclosure. This requires handling various types of data that fall into one of two categories: structured and unstructured. Structured data is highly specific and is stored in a predefined format such as age or date. This makes it simpler to handle. Unstructured data, on the other hand, is more complex to work with as it represents a combination of many types of data that are stored in their native formats like comments, tweets, and posts. Table 2.1 highlights some of the other approaches to privacy score calculation in the literature.

Tableau 2.1. Existing research on privacy score calculation

Paper	**Description**	**Metric valuation**
Hassanpour *et al.* [**166**]	The privacy leakage (PR) score is based on three components: sensitivity, visibility, and linkage. The sensitivity is classified on the scale: sensitive, semisensitive, and non-sensitive. Visibility denotes the audience allowed to see the content. The linkage parameter is calculated based on the pieces of shared data that can be grouped together to reveal a bigger picture.	$\boldsymbol{PR = F\ (sensitivity, linkage, visibility)}$ $= sensitivity * linkage * visibility$
Senarath *et al.* [**167**]	The authors build a model to measure the perceived privacy risk from the existing domain knowledge. The privacy risk P_{C} of a data element D_{i} in an application context C_{j} is in a monotonically decremental relationship with the relatedness of the data element D_{i} to the application context C_{j}.	$\boldsymbol{P_{(\mathrm{i,j})} = (S_i^a * V_{(i,j)}^b)/R_{(i,j)}^c}$

Li *et al.* [**168**]	This approach relies on capturing the user's attitude through a questionnaire to measure the proposed metric. The larger the numerical value is, the more sensitive the attribute is, and the more worried the user is about privacy leakage.	$\boldsymbol{Sensitivity\ \theta = (0.5 * L_3 + L_4 + 1.5 * L_5)/1.5}$ where: each L_i value corresponds to the number of users who expressed their concern as level "i" on this 5-point Likert scale: L_1: not worried at all; L_2: not worried; L_3: not clear; L_4: worried; L_5: very worried.
Oukemeni *et al.* [**169**]	The main metrics used are privacy impact and visibility. The privacy impact score $\boldsymbol{Imp_{\mathbf{Priv}}}$ reflects the impact of the assessment question on privacy, whereas the visibility score determines how accessible the data discussed in the question is.	$\boldsymbol{Imp_{\mathbf{Priv}} = Imp_{\mathbf{Pred}} + Imp_{\mathbf{Manage}} + Imp_{\mathbf{Diss}}}$ where: $\boldsymbol{Imp_{\mathbf{Pred}}}$: measures the impact of the question on predictability. $\boldsymbol{Imp_{\mathbf{Manage}}}$: measures the impact of the question on manageability. $\boldsymbol{Imp_{\mathbf{Diss}}}$: measures the impact of the question on dissassociability.

		Visibility=AV* Diff Accessibility Value (AV) measures the permissions given to share information with others. Data extraction difficulty (Diff) refers to the difficulty of extracting private information from the formats of data discussed in the assessment question.
Fu *et al.* [**170**]	The authors use visibility and granularity measures to calculate the overall privacy risk of a user. The former defines who the individual wants to share their profile content with. The notion of granularity, in this context, is similar to sensitivity and this work uses the following scale: four granularity levels for user profile items: 0 (no information shared by the users), 1 (low granularity), 2 (medium granularity), and 3 (high granularity).	$\boldsymbol{PS(i,j) = \sum_{j=1}^{m} \phi_{ij}}$ $\boldsymbol{\phi_{ij} = \sum \beta_{jk}\delta_{ijk} + \gamma_{ij}}$ $\boldsymbol{\phi_{ij}}$ is the privacy risk score for a given user v_i and a given profile item a_j.

Calculating the privacy score/ leakage/ risk is needed to assess the potential damage SNS users can incur on these platforms that they often frequent. This is the first step or prerequisite for privacy preservation. The next and main objective is to provide the appropriate mechanisms to protect users.

2.3. Privacy-preserving mechanisms

Privacy concerns have accompanied the Internet since its birth and deployment to the public. One of the first notable efforts to guide users online was the *Platform for Privacy Preferences* (P3P) project, which was developed by the *World Wide Web Consortium* (W3C) [**171**]. The mechanism's goal was to empower people to take their privacy into their own

hands by drawing their attention to potentially privacy-violating websites and giving them the option to make an informed decision.

With the progress in machine learning, novel approaches to privacy preservation have been proposed such as fighting fire with fire. Chakraborty *et al.* [**172**] proposed the *FORGE* repository, which uses deception to counter deception. To prevent malicious people from stealing real classified documents, the researchers thought of a solution to generate a similar but fake copy for each file. They conducted experiments with 20 people and their work proved to generate believable fakes capable of fooling potential attackers.
There is a trend of corporations deploying cunning techniques as preventive measures. Tushar Kothari, the CEO of a cybersecurity defence organization reveals a popular technique to trick attackers into revealing themselves [**173**]. Then, the defenders strengthen the network to cover any existing blind spots.

While this is promising, the most fortified secure system in the world can become useless if it is handled by someone who lacks the proper knowledge and awareness to operate it without creating openings for attackers. This brings us to the human at the centre of all of this. The well-established approaches addressing the user include dedicated official curriculums, training, and campaigns. Numerous governments have dedicated considerable resources to this area. The government of Canada, for example, is currently following a budget plan it made in 2018, which totals more than $500 million dollars over five years to foster innovation in the field [**174**]. In a similar initiative, in the US, a bipartisan law was passed in 2022, which will use $1 Billion dollars in funding for the first-ever state and local cybersecurity grant program [**175**].

The main criticism of traditional cybersecurity training and certifications lies in the approach itself, which is centred around task completion instead of the retention of applicable information. It is similar to a student who prepared for an upcoming exam and learned the required chapters by heart without diving deeper into the subject. Said individual would not benefit from the knowledge and skills in the long run. Hence, the need for novel approaches, one of which is games.
Table 2.2 shows a myriad of examples that have been released in the past 6 years addressing many issues. I widened the scope to include cybersecurity as well because the two concepts are intertwined. The main difference between the two is that cybersecurity tends to focus more on preventing breaches rather than on how user data is being handled. These games are referred to in the literature as *serious games* or *persuasive games*, sometimes interchangeably. There is, however, a difference between the two.

Serious games are games used for purposes other than entertainment [**176**] such as education [**177, 178, 179**], nutrition [**180**], cybersecurity training [**181**], and disaster management [**182, 183**][**170**]. Clark C. Abt coined the term "serious games" in 1970 and defined their characteristics as having an "*explicit and carefully thought-out educational purpose, not intended to be played primarily for amusement*" [**184**]. As long as they are designed for a goal beyond leisure and enjoyment, the games can be under this category. Persuasive games take it one step further as they are serious games whose goal is not just to educate people but also to change their behaviour. This can be connected to the concepts of the privacy paradox, which is detailed in Chapter 1. Serious games can inform people and heighten their privacy concerns and risk awareness (concerns), but persuasive games can have an impact on their actions (behaviour). Having the required knowledge is a necessary yet insufficient condition to make the best privacy-aware decision.

Tableau 2.2. Examples of privacy and cybersecurity games since 2017

Game	**Privacy and cybersecurity issues**	**Target users**	**Year of release**
CyberAwareness Challenge	General cybersecurity, insider threats, social networking, teleworking security	Young adults and older	2022
CyberSprinters	General cybersecurity	Children	2021
Cybersecurity Lab	General cybersecurity, data breaches, phishing, malware, general cybersecurity, password security	All	2021
Hotspot (living security)	Phishing, good security practices	Young adults and older	2021
Officeware Inc.	Phishing and malware	All	2021
Privacy Pirates	Privacy	Children (7 to 9)	2021
Infosec Deep Space Danger	Social engineering and malware	Young adults and older	2021
Hacker Bot	Passwords	Young adults and older	2021
Cybersecurity Game Spoofy	Online safety and privacy	Children	2021

The Missing Link	Phishing and smishing	all	2020
Cybersecurity Ops	General cybersecurity	Young adults and older	2020
Microsoft Security Adventure	Phishing, malware, identity theft, hackers	Young adults and older	2020
CyberLand	Routers, security laws, passwords, firewall	All	2020
Band Runner	Online safety	Children (8 to 10)	2020
Cyber Challenge	Cybersecurity	Young adults and older	2020
Education Arcade: Brute Force	Passwords	All	2020
Cyberhunters – Ghost in the net	Identity theft	Young adults and older	2019
Cybersecurity Circus	Identity theft, passwords, malware, general cybersecurity, e-safety	all	2019
J'accepte (UFC)	Privacy and online safety	All	2019
Centigrade Black Belt Cybersecurity Training	Privacy, cybersecurity, spam Defense, phishing, social engineering	Employees	2019
Hacking Hero – Cyber Adventure Clicker	Hacking	Young adults	2019
Enter	IT security, phishing, cybersecurity	All	2018
AggieLife	Online scams, passwords, e-safety, spyware	all	2018
Cryptris	Asymmetric cryptography	Young adults and older	2018
Conectado	Cyberbullying	Teenagers	2018
HackTale3D	Hacking, security Key, database encryption, SQL, cyberdefense	Cybersecurity Students or people familiar with security	2018

SOS FBI - Safe Online Surfing	Online Safety and Privacy	Children	2018
Keep tradition secured	Spoofing, privacy, internet security, phishing	all	2017
CyberJulie	Privacy, cybersecurity practices	Children	2017
Interland	Privacy	Children	2017
Trend Micro (Data Center Attacks)	Privacy, data leakage	Young adults and older	2017
Blue Team: A firewall Setup game	Firewall setup	Cybersecurity specialists	2017

One of the biggest and most well-known privacy and data security training platforms is *TeachPrivacy* [**185**]. It is deployed by the company of the same name whose chief executive officer is Daniel J. Solove, a John Marshall Harlan research professor of law at George Washington University Law School. TeachPrivacy offers topics like GDPR, privacy awareness, HIPAA, Family Educational Rights and Privacy Act (FERPA), etc. It is more customizable and engaging than traditional approaches.

Another instance of user-centric solutions is the work by Charu Singh *et al.* [**186**], which identifies fake websites designed to phish users. The authors proposed *Cuckoo Search-Support Vector Machine* (CS-SVM), which is a machine learning-based algorithm that gives 99.52% accuracy by using a *Radial Bias Function* (RBF). Along the same line comes the privacy assistant *PACMAN* [**187**], which exists as a social media guide and a reminder to carefully choose the audience of each post. The proposed system takes as input the user's information to utilize the existing yet often ignored or forgotten options to mitigate self-disclosure such as limiting the audience of certain posts. In fact, Facebook users, for example, have always had the option to share either with the public, friends, or a custom audience, but rarely do people analyze and decide to whom every post of theirs should be visible. That is the case, especially for users who are very active and engage with others on a plethora of pages and groups.

As for the research on resolving privacy issues in multiparty disclosure, the focus is on conflict resolution when the sharer's preferences do not align with the other party members. Wishart *et al.* [**188**] use suggestions to remind the user that the current post might be a point of contention and that they need to consult the other people involved. Moreover, the authors' collaborative privacy policy hinges on every co-owner's self-reported "strong" and "weak" privacy preferences. The desired outcome is for the different parties to come

to an agreement through manual non-automated negotiation. This is becoming more and more tedious as social media platforms keep growing and people build more connections. On Facebook alone, individuals are uploading more content than ever and have an average of 338 connections as of 2023 [**189**].

The idea of using automated negotiation for multi-agent issues was theorized since the 1980s and 1990s, as this time period saw the emergence of systems like the *persuader* [**190**] and *Oz* [**191, 192**]. Their approaches achieve privacy resolution through specific bargaining protocols. In general, privacy negotiations can be divided into *horizontal*, where users negotiate with each other, and *vertical*, where the negotiation is conducted between a user and a service provider [**193**].
The most commonly used one is protocol is alternating offers [**193**] in which each agent (representing a real person) takes turns to propose a way to settle the conflict. It is up to the other party to accept or counteroffer. They negotiate in a time-bound scenario and if the deadline is reached without an agreement, whether through consensus or majority ruling or other, they either get their reservation utility or in dire cases, nothing at all. The latter condition forces agents to cooperate since they want a minimum gain. The reservation utility is agreed upon before entering the negotiation. Table 2.3 shows some of the existing mechanisms of multiparty privacy conflict resolution.

Tableau 2.3. Examples of multiparty conflict resolution methods

Technique	**Description**
Manual [**194**]	The originators of content can specify a privacy policy for each of the content items they own and they can add/remove friends on the service. This approach assumes that the parties involved want to maintain a good relationship. Hence, once a conflict is detected, the co-owners get to either concede to the sharer or modify the content so as to no longer be impacted by it. An example of this could be cropping oneself out of a photo.
Ontology-based [**195**]	The OSN users are modelled as agents that use semantic rules to represent their human counterparts. They negotiate using the assumptions from their ontology, which can be enriched by new information, should the agent make the request.

Auction-based [**16**]	This work proposes a personal assistant to help end-users with managing conflict of co-owned data. When such an issue arises, an auction is triggered to regulate the privacy of the content. This is achieved through a series of bids that the agents are capable of issuing after learning the individual's preferences and how to represent their point of view.
Federated learning [**17**]	A multi-attribute reverse auction model is proposed to be used for user selection as well as payment calculation for participation in federated learning. The model uses a combination of economic and non-economic attributes in the negotiation resolution process.
Contract theory-based data trading [**18**]	The system provides a set of optimal contracts detailing various privacy-preserving levels and data trading prices. The final agreement is reached by using a group-weighted *maximum likelihood estimation* method.
Game theory-based [**196**]	The proposed *Online Information-Sharing Assistance* (OISA) is an interactive information-sharing trade-off solution. In the "game", the players are the users, and the pay-off function uses both the benefits and costs of the information disclosure.

The main criticism of these approaches is the strong assumption that everyone involved can and will express their preferences to the system. It is unrealistic especially when the user shares numerous posts involving many people. Additionally, the manual approaches would not be successful when the sharer is seeking immediate gratification. It is unlikely that in such cases, the person would individually reach out to each multiparty member, wait for their response, get their opinion on the subject, and make adjustments accordingly.
Receiving a prompt intervention when it matters is crucial, which is one of the foundations of nudges.

2.4. The nudge theory

The term was popularized by University of Chicago economist and Nobel prize winner Richard Thaler and Harvard law school professor Cass Sunstein in their book "*Nudge: Improving Decisions about Health, Wealth and Happiness*" [**8**]. Upon its publishing, it was

considered to be "*Best Book of the Year, 2008*" by both *The Economist* and *The Financial Times*. A nudge, according to the authors is any form of choice architecture that alters people's behaviour in an expected way without restricting any options or significantly changing their economic incentives. Thaler and Sunstein coined the term "*Libertarian paternalism*" in connection with nudges to denote the idea that it is possible to change one's actions without infringing upon their freedom of choice. The intervention must be easy and cheap to avoid. Nudges are not taxes, fines, subsidies, bans, or mandates. If a supermarket puts fruits at the forefront of the store and baked goods at the back to promote healthy choices, this is a nudge [**197**]. If taxes are imposed on junk food to discourage people from consuming it, this does not qualify as a nudge. On the subject of taxation, there is evidence corroborating the efficacy of nudges over fines and that citizens showed more willingness to comply with the former over the latter [**198, 199, 200**]. In Chapter 1, I tackled the notion of rational behaviour mainly in the privacy calculus and how introducing biases counters that principle. Theories of rational choice do not take into consideration the discrepancy between the observed actions and the optimal outcomes. The nudging theory has similar aspects as it acknowledges and aims to address the fact that individuals do not always take the most favourable course of action about their own welfare. This diverges from the traditional theories of rational economic behaviour. On a governmental level, in 2009, the US recruited Cass Sunstein to bring this new perspective and head the *Office of Information and Regulatory Affairs* (OIRA) to streamline new regulations [**201**]. In 2010, the United Kingdom joined this movement by creating the *nudge unit*, a *Behavioural Insights Team* (BIT) in the Cabinet Office. According to BIT's report, it achieved 22 times more savings than its running cost between 2011 and 2012.

In more recent years, governments are increasingly adopting behavioural science-based methods to meet their objectives [**202**]. The application of nudges has not lost momentum and it continues to be a trend in the public sector as reported by *Deloitte* [**203**]. In the US, since 2009, the *National Institutes of Health* (NIH) made "*The Science of Behaviour Change*" a priority by designating it as a Roadmap Initiative. A total of 20 million dollars was budgeted for "Behavioural Economics for Nudging the Implementation of Comparative Effectiveness Research". Now, about 14 years later, the potential of nudging is still promising not only for the US as the *World Health Organization* (WHO) estimated that by investing in the most cost-effective interventions, low and middle-income countries can expect a sevenfold return by 2030 [**204**]. A systematic review by Ledderer *et al.* [**205**] examined the existing research that uses nudge-based approaches to positively introduce healthy behavioural changes. Their findings point out that 42 out of 66 studies reported a positive effect aligning with the authors' objective. One was excluded for using financial incentives, which deviates from Thaler and Sunstein's definition of nudges.

Digital nudging refers to the use of elements from the user interface to apply this choice architecture and help technology users make a more informed decision [**206**]. The work of Congiu *et al.* [**207**] breaks down nudges into mainly three categories: *pro-self*, *pro-social*, and *marketing*.
The goal of the first type of nudges is the benefit of the nudged person. The term comes from the work of Hagman *et al.* [**199**], in which it is defined as a means of "*helping individuals steer away from irrational behaviour*" and "*decrease their long-term well-being*".
The second category, pro-social nudges, extends the benefit beyond a single individual to reach society. These nudges are equally called pro-others or simply social nudges. In real life, they are applied to various cases and for many objectives such as sustainable tourism [**208**], reducing energy consumption [**209**], blood donation [**210**], charity donations [**211**], and encouraging social distancing during the COVID-19 pandemic [**212**].
The third type is marketing nudges. Billboard advertising is a form of this as the placement, product, and slogan are all meant to increase sales. They do not force a choice on consumers, but they do utilize behavioural science to attract people and steer them towards a specific purchase. Amazon does this by showing messages like "*Want it delivered by today, 6pm–10pm*?" or using the "*Amazon choice label*" [**213**]. There are different opinions on whether this last type qualifies as a nudge to begin with, a point that I will further discuss in Section 2.7, which focuses on the ethics of nudging. Before that, a distinction needs to be made to separate nudges from recommendations.

2.5. Nudges versus recommendations

Recommender systems are a subclass of *information filtering systems* that suggest to the user the items that they are likely to be interested in [**214**]. The foundation of recommender systems is based on providing personalized content and services in accordance with individual preferences and interests. Nudges, on the other hand, are not a response to the user's immediate goal and subsequently, are not at the forefront of their mind. If Alice scrolls through Amazon looking for clothes and the website shows her personalized results for that request, this is a recommendation. However, if she is on Facebook, a social media platform not primarily used for e-commerce, and ends up getting an advertisement for clothes, this is a nudge.
Alice's aim was not to purchase clothes in the second scenario. It was instead the nudge's goal (and by proxy the advertiser's goal).

One might wonder why I am opting for the use of the term "nudge" in this dissertation over "recommendation". Well, while my goal is to raise people's awareness and mitigate disclosure, that is not the primary reason for which they are using social media. Paradoxically,

if one was on social media for the purpose of not disclosing data, they would never share any piece of persona, which defies the purpose of these platforms. My target users' goal is to socialize in any shape or form: share written posts, life events, selfies, videos, geotags, etc. The nudge's goal is privacy preservation. The next subsection dives deeper into this topic.

2.6. Disclosure mitigating nudges

Acquisti proposed that nudges have the potential to influence users' privacy decision-making and decrease their regret [**215**]. Thus, this section focuses on the two main approaches to nudging: a) non-user-specific *one-size-fits-all* and b) *tailored personalized nudges*.

2.6.1. One-size-fits-all nudges

In my article entitled "*Towards Enhanced Privacy-preserving Nudges*" [**216**], I provided arguments for the existence of one-size-fits-all nudges as well as tailored nudges. In this subsection, the focus is on the first category, which in the article, is also referred to as "objective nudges" since they are pushed in the same way to everyone in that situation regardless of what their preferences might be. One of the foundations of the objective approach is that it focuses on the privacy preservation objective over appeasing the user. Scenario 1 shows an example of this.

Scenario 1: "Bob and his friends Sam and Alex are about to graduate and will be looking for employment soon. Bob often shares social media posts with his friends and enjoys feeling closer to them through these interactions. One of which is instigating controversial debates between the three of them involving politics, religion, human rights, etc."

The same nudge would be pushed to Bob, Sam, and Alex were they to share their personal views on religion and politics. It does not matter whether Bob is more driven to share such content (based on his history, for example) than his friends, they would all receive it in the same manner. Although these nudges are not personalized, they can incorporate some effects and cues that are known to draw most users' attention. Turland *et al.* [**217**] make use of different colours to highlight the intensity of the privacy risk and draw the individual's attention. Their target users are people who are connecting to unknown public WIFI networks. Although it is generally recommended to never use such networks, individuals might find themselves in situations in which they need to choose one of the available ones. In those cases, receiving guidance via nudges can be the most accepted form of risk mitigation especially when accompanied by explanations and examples.
If we go back to the aforementioned scenario, instead of telling Bob that his post contains sensitive information, which is his religious and political opinions, further details are added.

He might receive the following nudge:
"*You are about to share sensitive content, which is your political and religious opinions. You are advised not to pursue this disclosure because it can have dire consequences on your personal, familial, and professional life. Did you know that 65% of Canadian companies in a recent survey said they use social media to screen job applicants? Even after becoming an official employee, 63% of the companies said that they would fire the person who damages the company's reputation.*
Read more on the subject by clicking the following link: [**218**]".

Moreover, one-size-fits-all nudges can make use of visual elements to influence the user by showing them how others responded in the same situation [**219**]. This plays on the general public's tendency to be susceptible to social influence [**220**]. Instead of the previous nudge, Bob might receive the following:
"*Y*ou are about to share sensitive content, which is your political and religious opinions. You are advised not to pursue this disclosure. In fact, 80% of users have chosen not to share similar content when they received this nudge. 70% of those who ignored the nudge expressed their regret later".

A supporting argument for the conception of objective nudges is that it eliminates the need for user modelling, which in itself is privacy-preserving. There is no need to collect and process user history and preferences and no risk of that information being leaked at any point.
However, the main criticism of this approach is that the generic non-customizable experience does not appeal to users and as such, this might defy the purpose of the nudge if it is not accepted. Hence, personalized nudges emerged in an effort to better understand and address the user.

2.6.2. Personalized nudges

The concept of personalized or tailored privacy solutions is not foreign to online platforms. Some SNS such as Facebook offer a native form of personalized settings through which the user gets to set privacy rules for their account. They continue to be applied in the future unless the user makes adjustments to them. These solutions do not often yield much success due to the general sense of apathy that a lot of individuals experience online. They do not make the effort, not because they are unaware of the repercussions, but because they simply are "*privacy fatigued*" [**221**]. The term encompasses emotional exhaustion, apathy, and cynicism, which take a toll on people causing them to reduce or completely shut down their decision-making faculty. To tackle this issue, there is an ongoing effort towards making the process more tailored yet less engaging on the user's side lest they feel burdened and

overwhelmed. One of the main consequences of this fatigue, which tailored nudges have the potential to address, is *digital mindlessness* [**222**]. As the term suggests, it refers to a fugue-like state in which people can find themselves when spending hours on hours scrolling through SNS, replying to comments and messages that they get without any second thought, etc.

Tailoring nudges requires knowledge of the user's preferences, and/or history, and/or context. The underlying assumption here is that users want to be addressed individually and have their preferences considered by the nudging system [**9**]. They tend to complain when they feel like the nudges are not specific enough to their case. In the literature, the terms tailored nudges, personalized nudges, and smart nudges are used interchangeably. Karlsen *et al.* [**223**] use smart nudging and they explain the process of designing them in Figure 2.3.

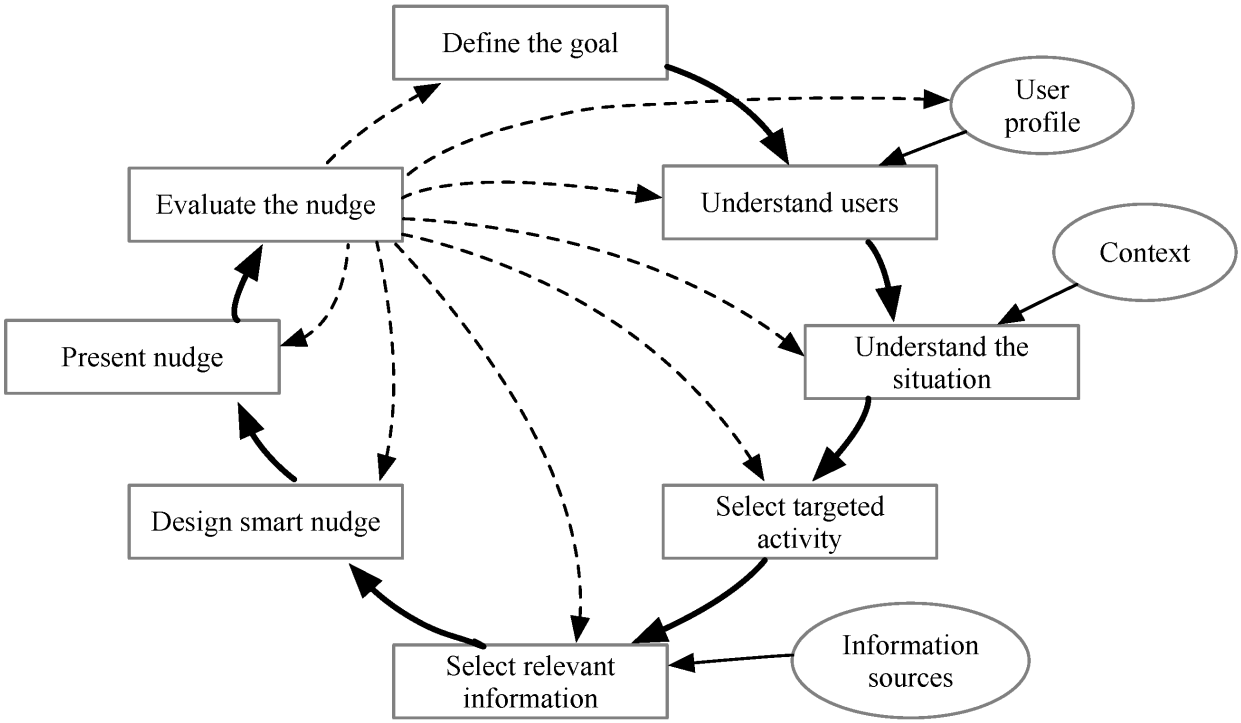

Fig. 2.3. Designing a smart nudge according to Karlsen *et al.* [**223**]

Wang *et al.* [**9**] proposed a social media-specific system to mitigate disclosure by showing the sharer randomly selected examples of who they are about to share the post with. Following this, the user has some time to reflect on whether they still want to disclose the content or not. Their findings showed a decrease in shared information. emphPrivacy Wedges [**224**], through visual means, also aimed to control the audience of a post, thereby reducing the privacy risk. The user's friends are arranged in groups around the centre, the closer they are

to it, the more trustworthy they are assumed to be. The calculation of the distance is based on the depth of their interpersonal distance with the sharer. The inner circle of friends is highlighted using a yellow border to draw the user's attention.

Moving on from audience-based to content-based solutions, Botti-Cebriá *et al.* [**225**] use *Natural Language Processing* (NLP) techniques to detect sensitive data in a written post. Their work focuses on the following categories: Location, health, drugs/alcohol, emotions, personal attacks, personal details (postal code, identification card, etc.), and neutral information (none of the previous categories). Although these tailored nudges hold the potential of reminding users of their privacy concerns when they are subjected to biases, what about cases in which the user is unaware, to begin with? Let us consider scenario 2.

Scenario 2: "Alice's friends often share the glamorous side of their life such as their luxury high fashion clothes and accessories, which makes her feel comfortable doing the same thing. She makes sure to post about her extravagant vacations and the frequent lavish dining experiences that she splurges most of her income on. She recently decided to apply for a mortgage and become a homeowner".

If asked to answer a questionnaire to tailor her future nudges according to her preferences, Alice, who is used to seeing others overshare without any immediate consequences and has done so herself, is likely to answer that she is being careful. She perceives the situation to be safe and her sharing behaviour to be far from detrimental. She lacks the knowledge to properly identify the issue, especially with regard to her goal of becoming a homeowner. According to Simon Conn, an overseas mortgage expert with 35 years of experience in the financial services industry, one of the red flags based on which a mortgage application can be rejected is a "*boast of excessive lifestyle habit*" [**226**]. Alice, unaware of this, would report that what she is constantly sharing is acceptable. Hence, a solely preference-based nudge would not be pushed to deter her or mitigate her disclosure. One could say that it even further perpetuates this behaviour and makes Alice more at ease with it since this "protector" nudge-based system allowed it.

The main advantage of tailored approaches is their potential of aligning users' preferences with their actions [**227**], something that one-size-fits-all nudges, by definition, cannot do. However, both types of nudges, more so the tailored category, do not elude ethical concerns.

2.7. The ethics of nudging

Even though this thesis contends that nudging for privacy holds great potential for behavioural alteration, this position is not unanimously held. Nudging can serve as a corrective

measure to adjust or at least mitigate the outcomes of irrational behaviour, which can be perceived as a form of limiting individuals' freedom of choice [**228**]. Some opposers of nudges argue that this mechanism can subvert the freedom of choice making them not as "*liberty-preserving*" or "*easily resistible*" as supporters claim. In the provocatively titled paper: "*Old wine in new casks: libertarian paternalism still violates liberal principles*" [**229**], Grüne-Yanoff presents many arguments against this choice architecture, amongst which is the ambiguity of "welfare". Nudges, by definition, should aim for the public welfare, but that notion is not as clear as one might assume especially when governments decide what is good for the public.

Moreover, a popular method of nudging is through presenting the desired alternative as the default option. If there are two choices one of which is the designed nudge, it might be highlighted while the other is greyed out. So, if this is done because users are susceptible to the default bias, is it not further instilling this bias? What happens when, in the future, the "highlighted default alternative" is not the "good alternative"? Does this not reinforce digital mindlessness [**222**] in people by offering them an easy choice that they rely on?
Haan *et al.* [**230**] studied the phenomenon and compared two groups of participants: 1) those who always faced random defaults throughout the entire experiment with 2) those who faced the "good default" in the first half of the experiment followed by random defaults in the second half. Their findings revealed that the second group chose the default option significantly more than the first. In other words, people get complacent when they find a shortcut or a readily available option that requires no further thinking on their part.

However, in my opinion, this poses a more fundamental question "*Are choices ever completely free of external factors, incentives, and limitations*?". In a way, our own body nudges us to eat when we are hungry and to sleep when we are tired. Let us go back to the earlier example of arranging food in a supermarket by putting healthier food at eye level or processed snacks that do not provide much nutritional value at the back. There is no neutral nudge-free manner to arrange the products and offer everything the same visibility as soon as a customer enters the store. Regardless of the intention, the end result is that some of them will be more accessible than others. If we agree on this point and the inevitability of nudges, then it is just a matter of "*In which direction should we nudge*?" and "*Where does the line start to blur between nudging and manipulation*?" rather than "*Should nudges exist*?" or "*Should we be influencing people's decision-making process*?".
Nudges are inescapable and are omnipresent in every aspect of our lives. Cohen [**231**] argues that not only does nudging not damage autonomy but it can enhance it. The key point of nudges, contrary to directives, edicts or any coercive approach, is that they influence choice

without imposing restrictions or impeding the liberty of choice [**232**]. If a nudge robs an individual of their autonomy, it is no longer a nudge.

Furthermore, my response to the criticism of the nudge being a reinforcement of the default bias and perpetuating mindlessness is adding a layer of explainability. I will explain this in the following chapters, but at this point, it is worth pointing out that my proposition aims to slowly but surely build up the user's awareness. It does not stop at highlighting the better alternative because that is not promising in the long run and what good does behaviour altering do if it is short-lived?
Another ethical concern is around marketing nudges. It can be said that calling them nudges is controversial to some degree. According to Thaler and Sustein's definition, this type mostly fulfils the criteria of not being restrictive nor significantly changing the economic situation of users. However, it is hard to determine where the line lies exactly. If a carton of milk has a promotional offer deducting 1$ from the original price, this does not seem "significant" enough to change the buyer's situation. Then, what about billboards advertising a chance to win a dream car or fancy house with an eligible purchase at a specific supermarket? If the first nudge is acceptable but the second is not, where does the limit lie?

To summarize, the ethical concerns around nudging are not to be taken lightly, but when applied to social media decision-making assistance, the issue tends to be more philosophical. Whether they compel people in a way that limits their freedom or not, the reality is that they already exist all around us. How I chose to circumvent the issue to some degree is through transparency. Informing people before using the system of its goal and the data that it collects from them gives them the freedom to approve or reject the guidance. However, ethics are not the sole consideration when tackling such a problem as the existing approaches have other limitations that the next section details.

2.8. Limitations of the existing approaches

While the topic of privacy nudges has garnered a lot of attention, especially in the past few years, the existing approaches have a few drawbacks. One of them is that most approaches lack transparency because they aim to make the intervention brief and direct. It is important to inform the user of the reason for which they are being nudged and to offer explainable pop-ups, messages, warnings, etc. It is very important to avoid low-hanging fruit lest the users end up laying their critical thinking to sleep. We do not want them to grow accustomed to accepting the quickest solution or else when hitting "accept all cookies" is the easy choice to access a website, they would not hesitate to click on it.

As for the preference elicitation process that precedes the mitigation process, most scholars consider the personal preferences of everyone involved to be known. If Alice tags Bob and Sam in a photo, most of the solutions detect the conflict based on a comparison between the three individuals' preferences. However, they do not dive into how such information can be acquired especially if the others are not users of the system. If they did not undergo a direct preference elicitation through a questionnaire or by granting access to their history, what is to be done then? If the only point of access to the group is the sharer and maybe few pieces of the other parties' public profiles, it would be interesting to figure out how to bypass the cold start problem and gain insight into the multiparty members. I aim to address these points through my own proposition in Chapter 5.

Table 2.4 includes some of the existing research papers on the subject. Various metrics are used to evaluate privacy such as visibility, which is almost universally defined as the number of people whom the sharer decided to disclose the content with.

Tableau 2.4. Limitations of the existing research on privacy preservation on social media

Paper	**Description**	**Dataset**	**Limitation**
Alemany *et al.* [**233**]	The authors use visual audience and text messages with a degree of privacy risk.	Data from an experiment with 42 teenagers.	No consideration for: • multiparty disclosure. • Data sensitivity. • Drive for disclosure. • Disclosure goal. • Previous behaviour and/ or preferences.
Meier *et al.* [**234**]	This work relies on fear appeals and social norms to raise participants' awareness.	Online experiment with 304 participants.	No consideration for: • multiparty disclosure. • Drive for disclosure. • Disclosure goal. • Contextual parameters influencing decisions.

Rudnicka *et al.* [**235**]	This research assesses the likelihood of sharing personal data after being primed by a motivational message that emphasised 'Learning' opportunities compared to other messages.	Data from 331 participants who take part in evaluating the research.	No consideration for: • multiparty disclosure. • Drive for disclosure. • Disclosure goal. • Creating disclosure-mitigating intervention.
Smith *et al.* [**236**]	The authors propose and test a Facebook privacy training program.	Dataset based on 204 adult Facebook users in the US.	No consideration for: • multiparty disclosure. • Drive for disclosure. • Disclosure goal. • Platforms other than Facebook.

Craciun *et al.* [**237**]	This work uses cognitive bias, namely social compliance, by informing participants of what their peers choose to share online. The goal is to reduce the disclosure of personal information.	Data from 455 United States residents recruited from Amazon Mechanical Turk.	No consideration for: • multiparty disclosure. • Drive for disclosure. • Disclosure goal. • How using cognitive bias can further reinforce it without adequate *explainability*.
Hirschprung *et al.* [**196**]	The proposed platform aims to solve the information-sharing trade-off problem, which occurs in multiparty conflicts.	Data belonging to 157 participants.	No consideration for: • Self disclosure. • Context. • Drive for disclosure. • Preferences over time. • Different social circles

Akkuzu *et al.* [**238**]	The system uses dynamic trust values for weighting co-owner opinions and reaching a consensus.	Data from 316 participants.	No consideration for: • Self disclosure. • Drive for disclosure. • Different social circles. • Context. • The fact that co-owners' preferences might not be known. • The drawbacks due to the co-owners needing to provide a high volume of data to make their preferences known.
Mosca *et al.* [**14**]	The authors propose an agent that supports image-based multiparty privacy using abductive reasoning.	Data from 321 participants.	No consideration for: • Self-disclosure. • Data other than images. • Drive for disclosure. • Preferences over time. • Different social circles. • Context. • Offline users whose preferences are unknown.

To conclude, the existing approaches tend to address either self-disclosure or multiparty disclosure, but not both. Tackling the two at once is not a mere issue of combining the solutions that address each individually. This is because they are intertwined. Only by considering both types of disclosure can a system perform comprehensive user modelling. The preference elicitation for offline co-owners needs to be taken into account if we are to approach the subject realistically.

2.9. Conclusion

This chapter details the existing work addressing the issue of disclosure. I start by explaining the preference elicitation process. Then, I move on to nudges, both one-size-fits-all and tailored interventions. Following this, there is a discussion about the ethical concerns that arise with these approaches.

Today's online users are surrounded by threats that may differ in their techniques and motivations, but most of them share one common point: they are increasingly focused on exploiting the individual's vulnerabilities. The existing solutions are hardly enough to subdue all of these issues, especially considering users' cognitive biases and their sense of privacy fatigue. The main criticism of the existing solutions is that, overall, privacy awareness is mostly being handled as another educational course when in reality, the issue goes beyond that. While these platforms provide knowledge in an abstract or simulated sense, the vulnerability that people exude is situational and contextual [**239**]. Even the novel approaches do not handle both self-disclosure and multiparty disclosure. In addition, they are based on strong assumptions such as having an omniscient system capable of eliciting the privacy preferences of every individual.
This thesis aims to promote responsible socializing online by, first, proposing a framework to understand the economics of privacy.

Chapter 3

A Multi-agent Approach to the Economics of Privacy

The purpose of this chapter is to offer an understanding of the parties, transactions, and dynamics in the economics of privacy using automated *Multi-Agent Systems* (MAS). As detailed in Chapter 1, there are three main players that act upon personal data one way or another: the individual/data owner, the data brokers/hubs/collectors, and the data users. While this thesis focuses on the protection of the data owner, this individual does not exist in an isolated bubble from the other parties and as such, it is important to observe and model all the interactions in this ecosystem.
The work in this chapter was achieved in collaboration with Laila Lamrabet.

3.1. Problem definition

This chapter assumes that every party involved is rational at all times, however, as discussed in Chapter 1, behavioural biases do play a role in the decision-making process. They are not considered at this point, by design, because the purpose is to study what ought to be if the user is knowledgeable and capable of governing their personal data. The reasoning for this lies in the outline of the contribution chapters: I start with what should be, move on to what is (Chapter 4), and end with what can be done about it (Chapter 5).

The data owner decides to sell a "quantity" of their personal if two conditions are met: the monetary compensation is satisfactory, and the privacy protection offered by the data broker is in alignment with the user's own concerns. The brokers collect, process, bundle, and anonymize (most commonly using K-anonymity) data before selling it to data users.
A dataset is k-anonymous if *quasi-identifiers* for each person in the dataset are identical to at least $k-1$ other people in the same dataset. A quasi-identifier is a piece of data that does

not identify an individual on its own but can become uniquely identifying in combination with other quasi-identifiers.

To explain k-anonymity, let us consider the example of 3-anonymity (k=3) and the quasi-identifiers are postal code and phone number. Medical condition is the sensitive attribute. In order for the dataset to be compliant with the principle of 3-anonymized, if we consider the user Alice, the records must contain at least two other records for each value combination of postal code and phone number. Figure 3.1 shows how *suppression* is applied to Alice's postal code and phone number by replacing parts of the data with an asterisk "*". This allows the collector to obtain the medical condition data with a general idea of the postal code and the phone number of the data owner Alice. Including hers', there are exactly 3 identical records for the combination (Postal code, Phone number)=(H3S ***, 514 827 ***). Hence, this is indeed compliant with 3-anonymity.

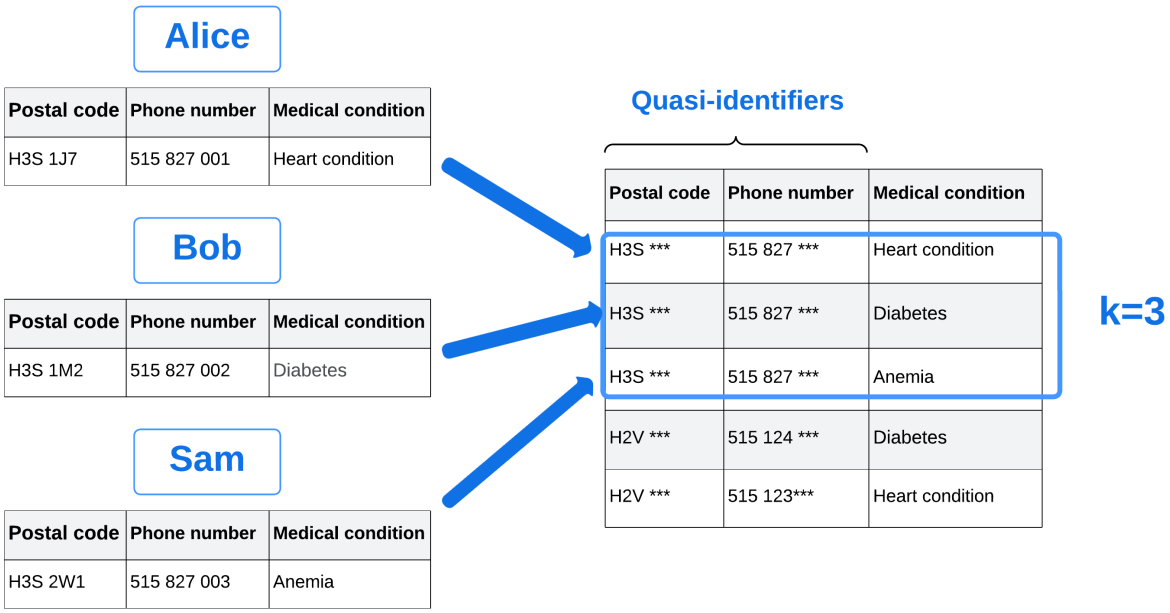

Fig. 3.1. An example of 3-anonymity

Evidently, there are conflicting interests between the three parties. The data users prefer a lower degree of anonymization because of the loss of specification, and ultimately quality or utility of the data. Knowing every piece of information about data owners is more profitable than partial disclosure. If we consider the example in 3.1, a data user who is a pharmaceutical company would rather know Alice's phone number and send her targeted advertisements. On the other hand, Alice, the data owner, prefers a higher degree of anonymization to protect her privacy and minimize the loss she might suffer if the data reaches the wrong hands.

As for data brokers, they would rather buy data from Alice at the lowest price and sell it for the greatest profit to data users. If each party is represented as a rational agent within a MAS, as seen in Figure 3.2, they all seek the maximization of their utility. The complexity of the issue resides in the constraints that apply to these transactions, namely, how the quality of data is impacted by the degree of anonymization, and in return it impacts the pricing of personal information.

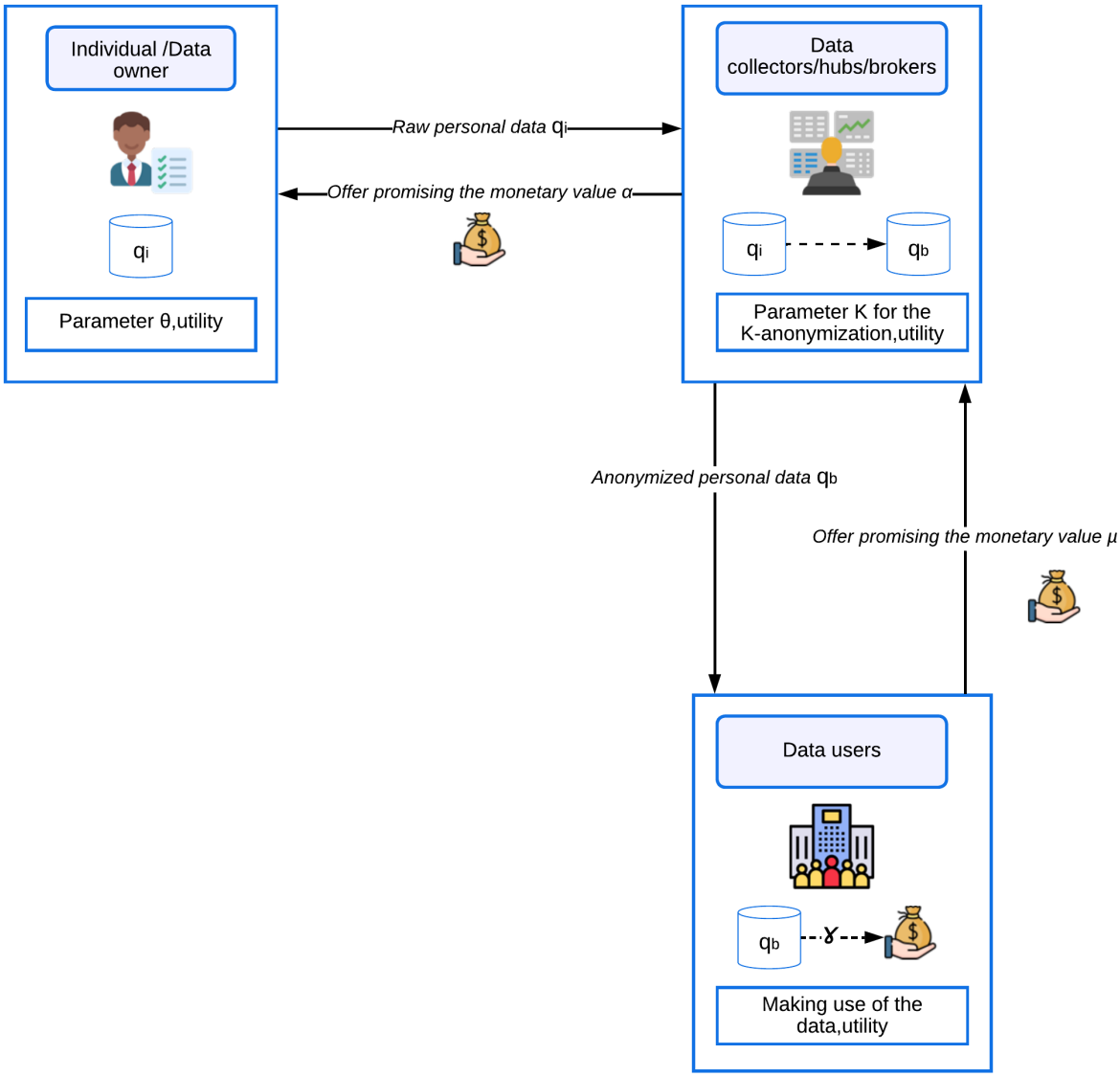

Fig. 3.2. A multi-agent system representation of the economics of privacy

Table 3.1 details all the parameters used in Figure 3.2. The anonymization affects both the quality and the price of data. The higher the degree of anonymization, the lower the quality/utility of data and the less profit brokers can make by selling it to the data users.

The lower, the more profit brokers can make but the less willing individuals are to sell their data.

Tableau 3.1. Parameters of the multi-agent system.

Parameter	**Description**
θ	Individual/data owner's preference in terms of confidentiality.
α	Monetary value of a single piece of data (or single record) to the data owner.
μ	Monetary value of a single piece of data (or single record) to the data broker.
γ	Monetary value of a single piece of data (or single record) to the data user.
k	Level of protection of confidentiality. It is the parameter in k-anonymization.
q_{i}	Utility of a single piece of raw data.
$q_{\mathbf{b}}$	Utility of a single piece of anonymized data.
U_{i}	Utility function of the individual/data owner.
U_{b}	Utility function of the data collector/hub/broker.
U_{u}	Utility function of the data user.
$off_{1,\mathrm{b}}$	Negotiation offer from the broker to the individual. It is the monetary compensation that the data user is prepared to pay the broker α.
$off_{2,\mathrm{u}}$	Negotiation offer from the data user to the broker. The offer consists of two parameters: the degree of k-anonymity k and the monetary compensation that the data user is prepared to pay the broker μ.

The parameter $\theta \in [0,1]$ stands for the individual's preferences in terms of confidentiality. A risk-averse person would choose a high value (closer to 1). Spiekermann *et al.* [**240**] developed a categorization that was originally proposed by Ackerman *et al.* [**241**]. Table 3.2 shows the concern level depending on the variation of θ.

Moreover, the higher the preferences in terms of confidentiality (increasing θ) the lower the value of $q_{\mathbf{c}}$ and ultimately the monetary values α, μ, *and* γ.

Tableau 3.2. Concern level depending on the user's confidentiality preference.

Parameter θ	**Concern level**	**Description**
$\theta \approx 0$	Little to no concern.	Privacy and confidentiality issues are not of concern to the individual.
$\theta \approx 0.5$	Averse to profiling.	Selective concern about their salary, postal address, and medical records.
$\theta \approx 1$	High concern	Very concerned about the totality of their private data.

3.2. Scenario and rules

The scenario detailed in this section is broken down into two steps since there are two *bilateral consecutive negotiations*: the first is between the data owner and the broker and the second involves the broker and the data user. It can be seen as a sequential game in which players (agents) take turns to propose an offer to reach the end of the negotiation. They know their own preferences as the past actions of the other agent (history). It is like a game of chess in which each player's current situation does not solely depend on them, but also on the other player.

3.2.1. First step: Negotiation between the data owner and the broker

The broker starts by proposing an offer to the data owner in the form of monetary compensation for the personal data: $off_{1,\mathrm{b}} = \alpha$. The data owner accepts or rejects it. In the case of a rejection, the game ends without any party accumulating any gain. In the other case, the data owner chooses the confidentiality value θ that aligns with their preferences. This has implications on the utility value $q_{\mathbf{i}}$.
Figure 3.3 shows the explicit representation of the players, offers, and utility functions of each player. In game theory, this is called an *extensive-form game*.
Throughout this chapter, whenever the asterisk **"*"** is used after a parameter, it denotes its optimal value.

3.2.2. Second step: Negotiation between the broker and data user

The data user presents an offer $off_{2,\mathrm{u}} = (k, \mu)$ to the data broker. This is based on the parameter k, which stands for the level of confidentiality that the broker applies to raw data

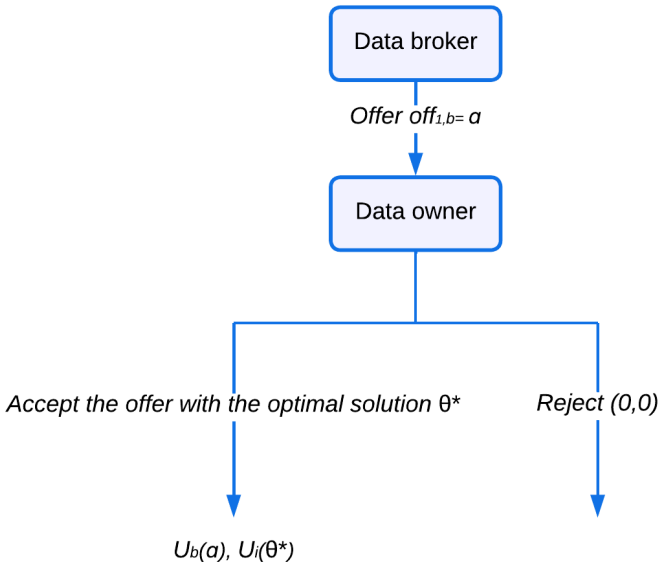

Fig. 3.3. Negotiation scenario between the data broker and the data owner

to align with the user's preferences. This exchange is shown in Figure 3.4. The data broker can reject it and end the negotiation or choose to accept it. In the latter case, the broker attributes the monetary compensation α to the data owner.

The proposed mechanism for both steps of the negotiation is based on a game theory approach, which models negotiations as a sequential dynamic game with complete information. Aliprantis [**242**] proposes a precise mathematical description of the *backward induction method*. It proves that every n-person (or agent in this case) sequential game with perfect information has a subgame perfect equilibrium. Thus, the solution to the economics of privacy issue, as presented in this chapter, does indeed exist. In addition, every subgame perfect equilibrium is a *Nash equilibrium* [**243**]. This term refers to a decision-making theorem within game theory, which states that a player can achieve the desired outcome by not deviating from their initial strategy. In other words, each player's strategy is optimal when considering the decisions of other players. Hence, it is a win-win situation since every party involved gets the outcome that they desire.

3.3. Utility functions

In general, a utility function is a representation of an individual's preferences for goods or services, which in the case of the economics of privacy, revolves around data and monetary compensations. Rational agents seek the maximization of their utility functions. However,

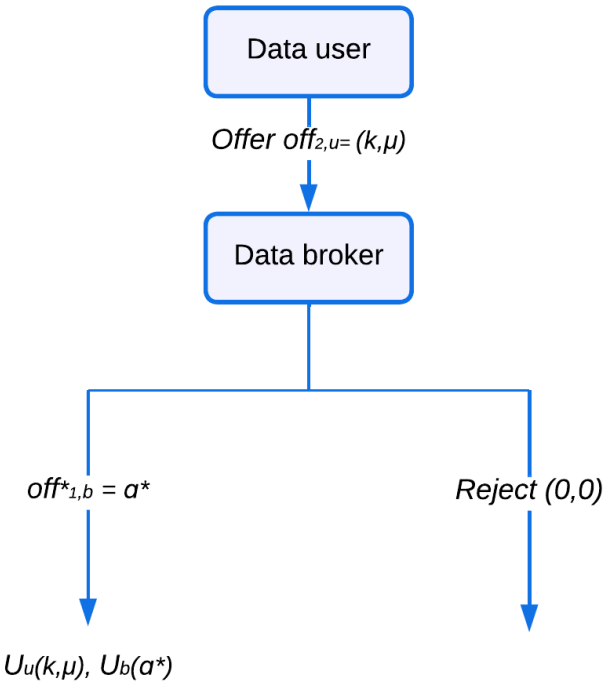

Fig. 3.4. Negotiation scenario between the data user and the data broker

reaching a consensus is paramount or else each agent has nil gain at the end. The following subsections detail the utility of the three agents.

3.3.1. Utility function of the data owner

The price at which the data owner sells their personal information depends on the utility q_{i}. To simplify the expression of the price, let us suppose that it is proportional to q_{i} whose value depends on the user preferences θ. The revenue $Revenue_i$ generated by the owner can be written as follows:

$$Revenue_i = \alpha * q_i(\theta) \tag{3.1}$$

Where $\alpha > 0$ is the monetary value of a single piece of data and q_{i} depends on θ. On the other hand, the data owner risks losing the confidentiality of their personal data if it is revealed. The assumption is that the monetary value of this risk $Risk_i$ is:

$$Risk_i = (1 - \theta) * q_i(\theta) \tag{3.2}$$

For example, an individual Alice who prioritizes the minimization of the risk would choose a confidentiality value $\theta \approx 1$. Applying this to Equation (3.2), the calculation shows that the $Risk_{\text{Alice}}$ is approximately nil as demonstrated below:

$$Risk_{\text{Alice}} = (1-\theta) * q_i(\theta) = (1-1) * q_i(\theta) = 0$$

The probability of re-identifying a k-anonymized sample or record of personal data is $\frac{1}{k}$ [**244**]. As a result, $\delta = 1 - \frac{1}{k}$ is the probability of an individual's confidentiality being protected against re-identification. The **total loss in terms of privacy/confidentiality** L_{i} can be calculated as the **expected value of this risk** $\mathbb{E}[Risk_i]$ and is expressed as follows:

$$\begin{aligned} Loss_i &= \mathbb{E}_\delta[Risk_i] \\ &= \mathbb{E}_\delta[(1-\theta) * q_i(\theta)] \\ &= (1 - \frac{1}{k}) * 0 + (\frac{1}{k}) * ((1-\theta) * q_i(\theta)) \\ &= \frac{(1-\theta) * q_i(\theta)}{k} \end{aligned} \tag{3.3}$$

By using Equations 3.1 and 3.3, the utility of the data owner U_{i} can be written as follows:

$$\begin{aligned} U_i &= Revenue_i - Loss_i \\ &= (\alpha - \frac{1-\theta}{k}) * q_i(\theta) \end{aligned} \tag{3.4}$$

Each couple $(\theta, q_i(\theta))$ defines a strategy that the "player" data owner can follow.

3.3.2. Utility function of the data broker

The data broker chooses to buy a quantity of personal data, which, when anonymized, has the utility q_b. The price of this data is increased if the quality or quantity increases. Supposing that the price is proportional to q_b, the revenue $Revenue_b$ of the data collector or broker is written as follows:

$$Revenue_b = \mu * q_b \tag{3.5}$$

Where $\mu > 0$ is the monetary value of a single piece of anonymized record or data. The loss of the broker called $Loss_b$ corresponds first to the revenue of the data owner $Revenue_i$ since the broker pays it to them. The second parameter contributing to the loss is the added fees of processing, anonymizing, and hosting the data $Costs_b$. As a result, the equation of the loss can be achieved by injecting Equation (3.1), which amounts to:

$$\begin{aligned} Loss_b &= Revenue_i + Costs_b \\ &= \alpha * q_i(\theta) + Costs_b \end{aligned} \tag{3.6}$$

Equation (3.7) describes the utility of the broker during this first step of the negotiation with the data owner and uses Equations (3.5) and (3.6):

$$\begin{aligned} U_b &= Revenue_b - Loss_b \\ &= \mu * q_b - \alpha * q_i(\theta) - Costs_b \end{aligned} \tag{3.7}$$

Anonymizing the data changes its utility from $q_i(\theta)$ to q_b depending on the parameter of the anonymization k. This transformation can be represented as a function g in which:

$$q_b = g(q_i(\theta),k) \tag{3.8}$$

The function g is decreasing seeing as the higher the value of k the lower the utility q_b. Let us replace q_b in (3.7) with its expression in (3.8). Hence, the new formulation of the data broker's utility is:

$$U_b = \mu * g(q_i(\theta),k) - \alpha * q_i(\theta) - Costs_b \tag{3.9}$$

Each value of α defines a strategy that the "player" data broker can follow.

3.3.3. Utility function of the data user

The data user makes a profit by repurposing the anonymized data provided by the broker. $Revenue_u$ is proportional to the utility of anonymized data q_b.

$$Revenue_u = \gamma * q_b \tag{3.10}$$

Where γ is the monetary value of a single piece of data after being put to use by the data user. As for the loss $Loss_u$, it corresponds to the monetary compensation paid to the data broker $Revenue_b$ to acquire the anonymized data. Equation (3.11) highlights the expression of the data user's loss:

$$Loss_u = \mu * q_b \tag{3.11}$$

Finally, the utility of the data user U_u is:

$$\begin{aligned} U_u &= Revenue_u - Loss_u \\ &= (\gamma - \mu) * g(q_i(\theta),k) \end{aligned} \tag{3.12}$$

Each couple (k, μ) defines a strategy for the "player" data user.

3.4. Equilibrium strategy

The first negotiation involves the data owner and the broker. The solution to which is proposed based on the backward induction method [**242**] to attain a Nash equilibrium.

3.4.1. Data owner

The data broker proposes an offer $off_{1,\mathrm{b}} = \alpha$ to the data owner. The goal is to find the optimal strategy for the data owner considering the aforementioned offer. Equation 3.13

shows the optimization problem of the data owner, which is the maximization of this player's utility U_i:

$$\arg\max_{\theta} U_i = max_{\theta}\left[\left(\alpha - \frac{1-\theta}{k}\right) * q_i(\theta)\right], \theta \in [0,1] \tag{3.13}$$

Guo *et al.* [**245**] establish the relationship between the utility of personal data and the preferences in terms of confidentiality as seen in Figure 3.5. This was based on experimental observations.

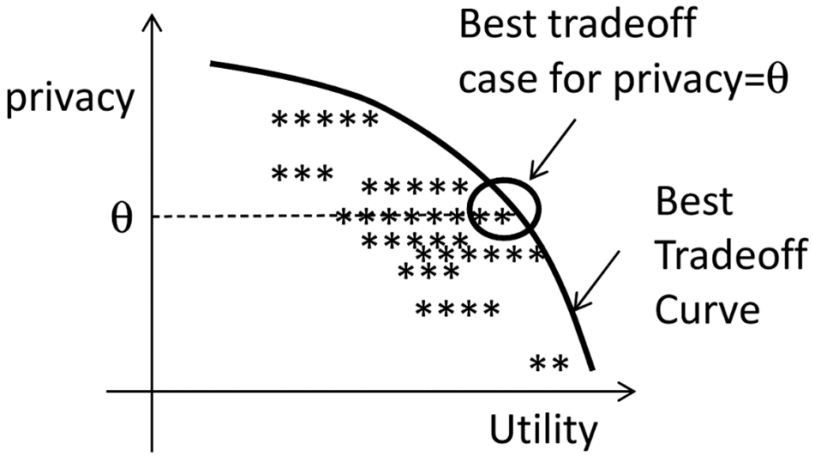

Fig. 3.5. Trade off between privacy and utility [**245**]

Inspired by this approach, I posit that $q_i = f(\theta)$ where f is a continuous concave decreasing function defined on the interval [0,1] such as f(0)=1 and f(1)=0. It can be written as $f(\theta)$:

$$f(\theta) = 1 - \frac{\theta^2}{1+\sqrt{1-\theta}},\ \theta \in [0,1] \tag{3.14}$$

Figure 3.6 shows the representation of the function $f(\theta)$. It is the variation of the utility of the data owner as a function of the preferences in terms of confidentiality.

If the optimal value θ^* exists, then, the data owner accepts the offer $off_{1,\mathrm{b}}$, if the utility is strictly positive, meaning that $U_{\mathrm{i}}(\theta^*)>0$. In the alternative case, the data owner rejects it. The best response Res_i^b of the data owner to all the possible offers of the broker's offer is:

$$Res_i^b = \begin{cases} Accept\ and\ choose\ \theta^*\ if\ U_i^* > 0 \\ Reject\ if\ U_i^* \leq 0 \end{cases}. \tag{3.15}$$

3.4.2. Data broker

After determining the best response Res_i^b of the data owner to the offer of the broker, the next objective is to find the optimal offer $off_{1,\mathrm{b}}^*$ presented by the latter to the former. The

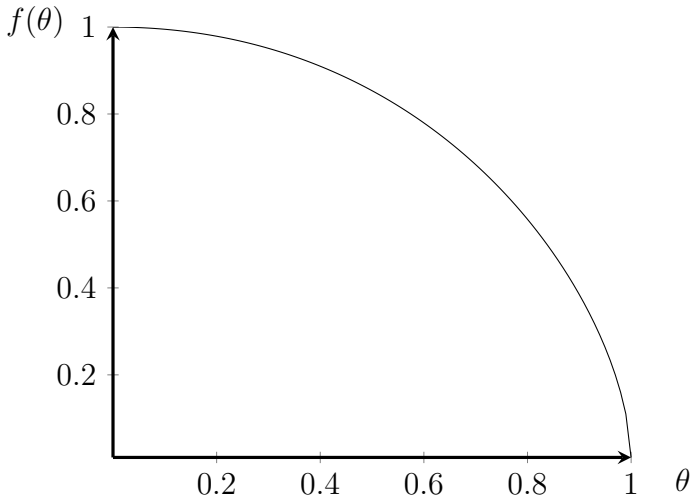

Fig. 3.6. Representation of the proposed $f(\theta)$ function

combination of these optimal strategies $(off^*_{1,\mathrm{b}}, Res^b_i)$ is the perfect Nash equilibrium in the game involving the data owner and the broker. Supposing that the data owner accepts the offer $off^*_{1,\mathrm{b}}$, the solution to the problem would be choosing α^*, the solution to the following optimization problem:

$$\arg\max_{\alpha} U_b = max_\alpha \left[\mu * g(\theta^*,k) - \alpha * q_i - Costs_b\right], q_i(\theta^*) \in [0,1] \tag{3.16}$$

Where:
θ^* depends on α (as shown in Figure 3.3) or else U_b would always be maximal when $\alpha = 0$. $\mu > 0$ and $g(\theta^*,k)$ is the utility of $q_i(\theta)$ after applying k-anonymization. Xu *et al.* [**246**] formulate the general relationship between the utility of raw data q_i and anonymized data q_b (or $g(\theta,k)$) based on which I derive the following special case:

$$g(q_i(\theta),k) = \frac{1}{2} * (1 + \frac{1}{k^{0,25}}) * q_i(\theta) \tag{3.17}$$

The value of $g(q_i(\theta),k)$ increases as k decreases and q_i increases. The data broker takes part in two negotiations. Let us examine the equilibrium in both of them:

- First negotiation: If θ^* and α^* exist, then, the Nash equilibrium exists between the data owner and the broker. For this couple of players, the strategies corresponding to the equilibrium are: (Offer $off^*_{1,\mathrm{b}} = \alpha^*$, accept with θ^*) and (Offer $off^*_{1,\mathrm{b}} = \alpha^*$, reject).
- Second negotiation: If α^* exists, then the broker follows this strategy: It accepts the offer $off^*_{2,\mathrm{u}}$ as long as U^*_{b}>0. In any other situation, it rejects the offer. This can be summarized as:

$$Res^u_b = \begin{cases} Accept\ and\ choose\ \alpha^*\ if\ U^*_b > 0 \\ Reject\ if\ U^*_b \leq 0 \end{cases} . \tag{3.18}$$

3.4.3. Data user

Considering the optimal response Res_b^u, the data user must determine the optimal offer $off_{2,\mathrm{u}}^*(k^*,\mu^*)$ that can be presented to the data broker in exchange for the anonymized data. Hence, it formulates the offer to maximize its own utility U_u while facing the response Res_b^u. The combination of the optimal strategies $(off_{2,\mathrm{u}}^* = \alpha^*, Res_b^u)$ leads to the Nash equilibrium between the data broker and the data user. If the broker accepts the offer $off_{2,\mathrm{u}}^*$, the optimal strategy of the data user consists in choosing $k^*$and μ^*, which present the solution to the following problem:

$$\arg\max_{k,\mu} U_u = max_{k,\mu}\left[(\gamma-\mu) * g(q_i(\theta),k)\right], k \geq 1 \; and \; \mu > 0 \tag{3.19}$$

$q_i(\theta)$ is fixed and no longer considered a variable at this point because it depends on the negotiation between the broker and the data owner without the direct involvement of the data user.

If α^* and (k^*,μ^*) exist, then the Nash equilibrium for the two successive subgames exists. The strategies adopted by the players (user, broker) that correspond to it are:
(Offer $off_{2,\mathrm{u}} = (k^*,\mu^*)$, accept with α^*) and (Offer $off_{2,\mathrm{u}} = (k^*,\mu^*)$, reject)

3.5. Example

This example serves as a way to understand how these equations work in a realistic scenario assuming that the system has all the inputs it requires to initialize the negotiation. The following is an example of a negotiation scenario between the data owner and the broker. Back when it was still active, *Datacoup* was a data broker that offered monetary compensation in exchange for the personal data of individuals. This includes social media activity, financial transactions, medical applications, and/or online activity such as Google searches. For this specific example, let us suppose that a data collector like Datacoup wishes to draw the maximum profit from the information it collected. The anonymization parameter is **k=10** and the offer proposed by the broker to the individual is: $\boldsymbol{off_{1,\mathrm{b}} = \alpha = 0.3}$. The latter, as a rational agent, tries to maximize its utility. If we plug the aforementioned variables in Equation 3.13, the result is as follows:

$$\arg\max_{\theta} U_i = max_\theta \left[\left(0.3 - \frac{1-\theta}{10}\right) * q_i\right], \theta \in [0,1]$$

One of the easiest ways to solve this is using a graph of the function, which can be seen in Figure 3.7. The solution to the optimization problem of the data owner is $\boldsymbol{\theta^* = 0.34.}$ Whether the data owner accepts or rejects the offer is encompassed in the response Res_b^i as explained in Equation 3.18. According to the graph, $\boldsymbol{U_i^* = 0.22 > 0.}$ Thus, the individual accepts the offer $\boldsymbol{off_{1,\mathrm{b}} = \alpha = 0.3.}$.

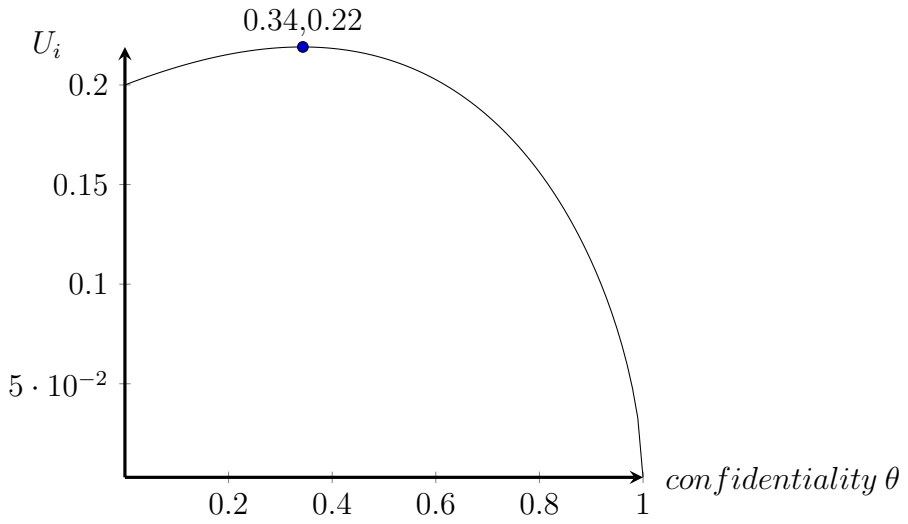

Fig. 3.7. Graph of the utility U_i as a function of the confidentiality θ

The validation of the negotiation-based system requires numerous samples, which I acquired by simulating the interaction between the three agents.

3.6. Validation of the negotiation mechanism

The simulations are performed using *Matlab* in order to capture the parameters of the Nash equilibrium for the two successive games involving all three parties. The purpose is to observe the adaptive behaviour of the agents and the outcome of the negotiation. This serves to validate the proposed equations.

3.6.1. Results of simulating the first negotiation

To simulate the negotiation, there is a need to find the optimal confidentiality value of the data owner θ^*, which serves as a response to the offer of the broker $off_{1,\mathrm{b}}$. To do so, the value of α is defined as $\alpha \in [0,1]$, which is applied to Equation 3.13. The experiments are conducted by testing multiple values of k in order to observe the impact of the degree of k-anonymity on the decisions of the broker and the data owner. In practice, the value of k that is generally applied to raw data prior to redistributing it is $k \geq 5$ [**247**] [1]. Hence, in the simulations, the values considered are $k = 5, 6, .., 20$. Figure 3.8 highlights the results of the simulation.

As the data broker increases the value of the offer to the data owner, the latter tends to lower their degree of confidentiality and as a result, express more willingness to disclose private information. Facing considerable remuneration, the individual is prepared to risk confidentiality in exchange for money. As such, it is more advantageous for the broker to

[1] `https://www.ccohealth.ca/sites/CCOHealth/files/assets/CCODataandDisclosurePolicy_0.pdf`

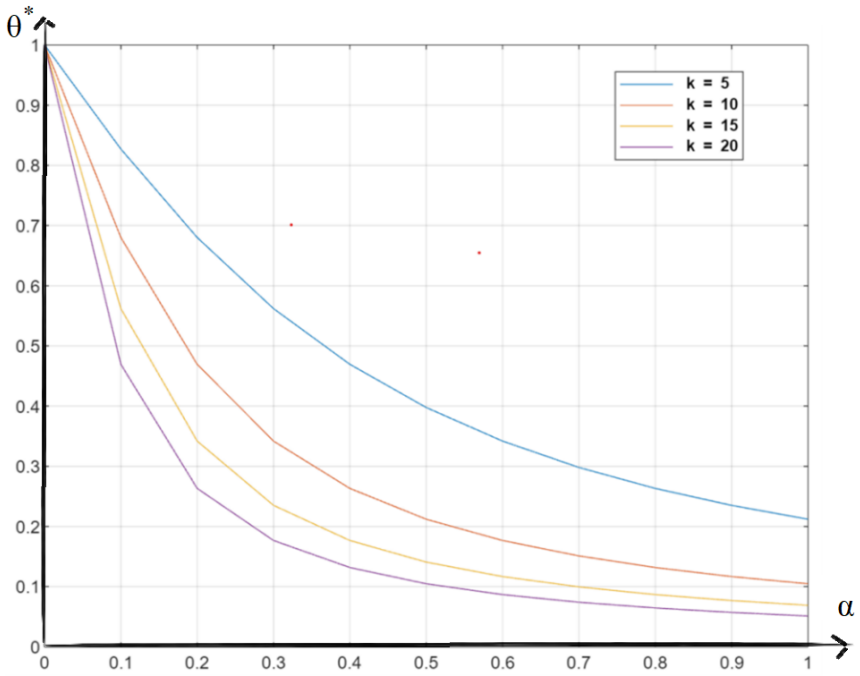

Fig. 3.8. The impact of the monetary compensation α on the level of confidentiality θ^*.

pay the data owner more to acquire better quality data and increase the probability of the offer being accepted. Figure 3.8 also shows that for a specific monetary compensation α, the confidentiality parameter θ^* decreases when k increases. This indicates that the individual is willing to lower their confidentiality preferences and sell their sensitive information if they are reassured that they would be anonymized before being used or resold, with a high degree of confidentiality (k). To summarise, the broker needs to adopt a higher level of anonymity to be able to convince the owner to sell their data. Moreover, this phenomenon is better observed for the smaller values of $k \leq 10$.

The optimal utility of the data owner U_i^* is approximately proportional to the monetary compensation α as shown in Figure 3.9. This result aligns with the assumption made at the beginning of this chapter, which is that the owner's utility depends on the strategy of his "opponent" in the game, the broker. Additionally, from a certain degree of anonymity, in this simulation $k \geq 10$, the optimal utility becomes less sensitive to the increase of k. In other words, the utility of the individual converges to a fixed value when the data owner adopts a good data protection (anonymization) process.

3.6.2. Results of simulating the second negotiation

The next negotiation involves the broker and the data user. Same as in the previous simulation, the values of k and μ are fixed such as $\mu \in [0,1]$ and $k = 5, 6, \ldots, 20$ with the goal of calculating the variation of α^* using Equation 16. Furthermore, there is an

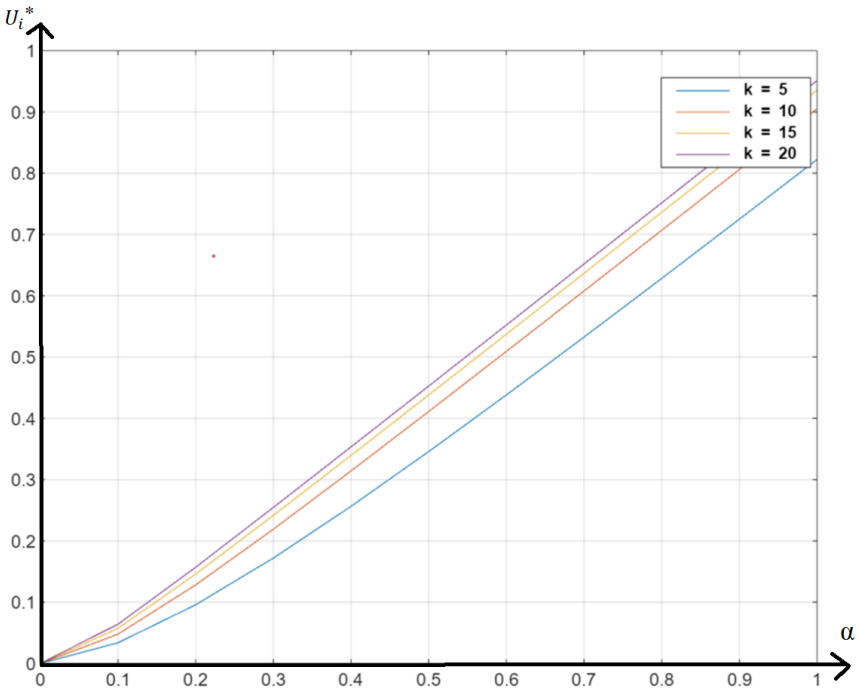

Fig. 3.9. The impact of the monetary compensation α on the utility U_i^*.

observation of the impact of the variation of the monetary compensation μ offered by the data user to the broker and the level of anonymization k on the quality and the quantity of the personal data q_b^*. The latter parameter $q_b^* \in [0,1]$ is specific to the data sold by the collector in response to the offer $off_{2,\mathrm{u}}$.

Figure 3.10 shows that for a given value of k, when the data user increases the monetary value of their offer μ, the data broker also increases the optimal compensation it is willing to pay the owner. Let us consider the example in which a data user wants to acquire a large quantity of personal information, which was anonymized using a specific confidentiality parameter. Hence, it is in the data user's best interest to present a satisfactory offer to guarantee the desired quantity and quality of anonymized data. Given a fixed μ value, increasing the anonymity parameter k results in a decrease in the optimal compensation α^* attributed to the individual (Figure 3.10). This can also be interpreted as the data owner accepting to sell their sensitive data at a low price if they are reassured that they would be anonymized with a high k value. Their reasoning aligns with a risk-averse behaviour in which the person would rather have lower gain as long as they are not subject to potential high loss. The broker can adapt by lowering the compensation they offer an individual who exhibits such traits.

In addition, Figure 3.11 showcases how the optimal utility of anonymized data q_b^* increases when the monetary compensation paid to the broker increases as well, given a specific value of k. For a fixed value of μ, the optimal utility of anonymized data q_b^* decreases when the

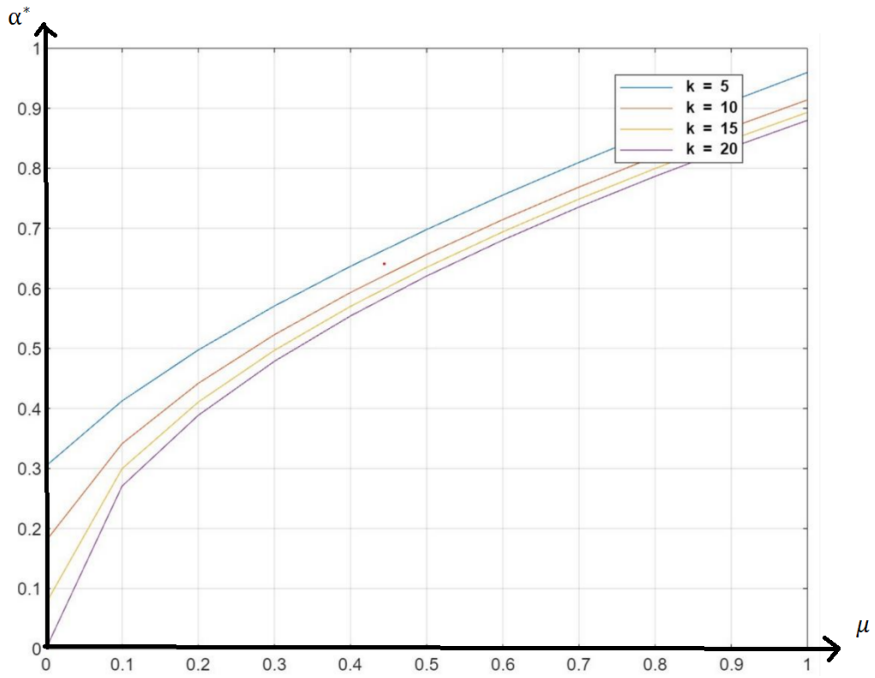

Fig. 3.10. The impact of the variation of the monetary compensation μ (paid to the broker) on the optimal value of the compensation paid to the individual α^*

degree of privacy protection k increases. This is due to the fact that anonymization reduces the quality of the collected personal data.

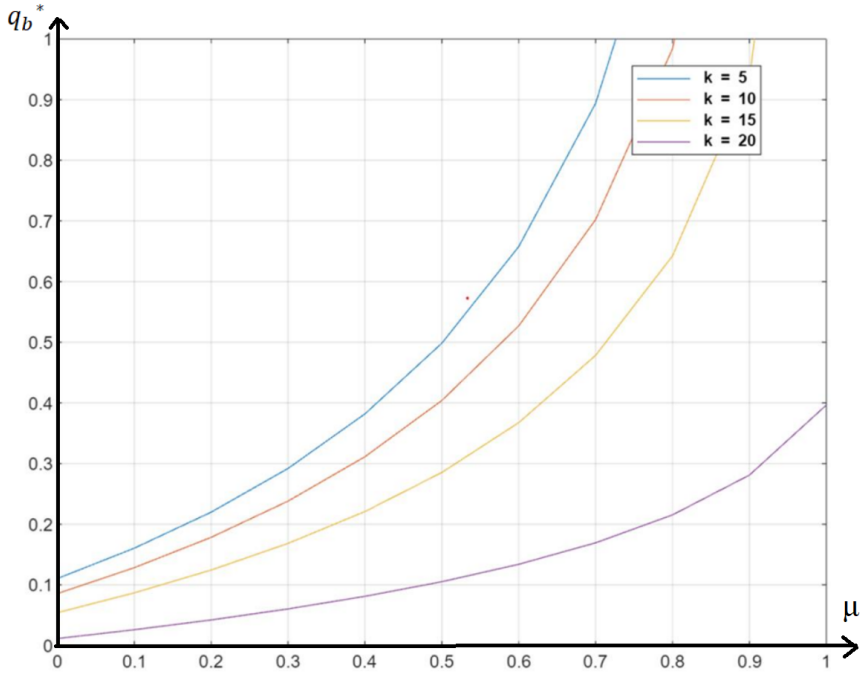

Fig. 3.11. The impact of the variation of the monetary compensation μ (paid to the broker) and the anonymization level k on the optimal utility of the anonymized data q_b^*

3.7. Conclusion

This chapter tackles the subject of the economics of privacy using multi-agent negotiation. Each agent is rational and aims to maximize their utility in terms of monetary gain while facing some constraints. The data owner has privacy and confidentiality concerns, the data broker must respect the owner's preferences and also appease the data user while maintaining a profitable transaction. These interactions are modelled as two sequential dynamic games; the solution to which is based on the extensive body of research on game theory. The focal point is figuring out the optimal parameters leading to an equilibrium between all the parties. This means that, first, the data owner gets to decide the price at which they want to sell their data while considering their privacy preferences. Second, the broker can determine the price at which it wants to purchase data taking into account its budgetary constraints. Third, the data user possesses control over the degree of anonymization that it requires for its envisioned purpose and profit. All of their objectives are intertwined and cannot be considered independently, hence, the need for such multi-agent system negotiation.

To summarize, this chapter tackled self-disclosure in the context of selling one's private information in exchange for monetary gain. Now that we are done with what ought to be, we can move on to what is. The next step is to focus on one area namely SNS and design a framework to explain and detail the factors leading an individual to divulge information in the form of self-disclosure or multiparty disclosure.

Chapter 4

MULTIPRIV: A Framework for MULTIfaceted PRIVacy Decisions

Privacy issues extend to every aspect of the online universe such as shopping platforms, emails, blogs, and most notable of all SNS. Their exponential growth is accompanied by a spike in the amount of time spent by users online, especially teenagers to young adults allocating an average of three hours a day for various types of SNS. From commenting, and sharing opinions to selfies, and videos, they expose numerous personal pieces of data, jeopardizing their privacy. Nowadays, the most vulnerable demographic, kids, can have their entire life on the Internet for everyone to see from the day they are born. New mothers go as far as to disclose that their babies are born prematurely or with a heart condition. The repercussions of this dawned upon parents starting around 2016 as the first case of a teen suing her parents for sharing her photos became a reality [**248**]. As soon as the first generation of children who were born and raised in the era of normalized oversharing grew into young adults, disclosure on social media became one of the most controversial issues of the digital age. This is proven by the various cases in which offspring take legal action against their parents because of this special case of multiparty disclosure. Since the aforementioned case, multiple court hearings ruled in favour of the children. In particular, a mother has to pay a 10,000 euro fine if she posts pictures of her teenage son on Facebook without his consent [**249**].

The two main issues with this behaviour are: first, the fact that users often underestimate the repercussions of self-disclosure on their personal lives and second the impact on others and second, the concern around multiparty disclosure. It consists in revealing information about other members of society who might not be willing or consenting to the exposure. Nudging users to reduce this phenomenon is a popular behavioural reinforcement method. Making these guiding interventions as well-received as possible is of great interest to this research area.

This chapter proposes a user-centric framework called *Multipriv* to contribute to this task. It is a crucial step towards improving the framing of nudges. The existing frameworks are mainly corporate-devised (or corporate-oriented) models for privacy preservation not intended for the general public. To the best of my knowledge, Multipriv is the first framework that studies the factors around both self-disclosure and multiparty disclosure from the perspective of the sharer to improve the nudging process.

4.1. Introducing Multipriv

There is a need for a human-centric solution to guide users through cues and suggestions that benefit them and their social circles. Nudges, as part of a growing body of research on positive behavioural alteration, hold the potential of addressing this. Designing Multipriv allows us first, to understand the factors and parameters around both self-disclosure and multiparty disclosure and second, to offer the basis for formulating user and context-specific nudges, which tend to get more acceptance the more tailored they are. Sections 4.1.1 and 4.1.2 further elaborate on this, but before diving deeper into the subject, Table 4.1 shows examples of scenarios along with their respective nudges. I envision pushing a first nudge that is less pressing (phase 1) and if the user rejects it, it is followed by a second one that urges them by providing more information. The latter is an explainable nudge. *One-size-fits-all* nudges, which are generic and non-user specific, are included in the table for comparison.

To the best of my knowledge, such a framework that illustrates the factors that sway the sharer's privacy decisions is yet to be introduced. It is with this aim of shedding light on the individual, societal, and contextual factors that this work presents itself. The ultimate aim as far as this chapter goes is to offer a recipe or set of guidelines for the implementation of well-framed nudges. Hence, I propose the design for Multipriv and then, I discuss the lessons learned, areas that remain controversial, as well as issues of which to be aware. Figure 4.1 shows an overview of Multipriv, specifically, the factors and catalysts that either incite or deter disclosing data about people other than oneself.

Tableau 4.1. Scenarios showcasing the difference between one-size-fits-all and my enhanced nudges

Scenario	**My proposed enhanced nudge**	**One-size-fits-all nudge**
Scenario 1: Alice is stressed out about her financial situation during a **looming recession**. While having a discussion at work with her **greatest** colleague Sam, she finds out the difference between their salaries. She wants to post on social media details about her employment situation and how she is getting underpaid compared to Sam who has the same position at work. Alice is usually **self-disciplined (conscientious)**, but she has been in a financially difficult situation and finding out the difference in wages made her act on **impulse**.	**Phase 1**: You have previously indicated that information related to your colleagues is very sensitive. This post does not align with your previously shared content, and you might regret sharing this down the line. Would you like to schedule the post for later?	Work-related information in general can be sensitive. We advise you not to share this. The system **does not consider**: • her past behaviour. • that she is conscientious, and regret is something she is susceptible to. • that she is acting on impulse (as such it cannot suggest scheduling for later, which is often a good solution for emotional bias). • that there is an economic struggle.

	Phase 2: Did you know that there have been many incidents where sharing wage-related information on social media costs someone their job? Below are examples of news articles referring to this. The colleague to whom you refer might suffer the consequences, especially during this economic struggle.	
Scenario 2: Bob's old **friend** Alice is visiting. So, he takes her to a party where alcohol is served. They take a **photo together** showing the crowd and activities in the background to commemorate the occasion. Bob wants to share the photo as he usually does in these situations. Alice, however, does not want to be associated with these circumstances especially with **alcoholic drinks** as it is highly frowned upon in her **workplace**. She has never shared such a post before.	**Phase 1**: You are about to share a photo including someone else at a party. According to your history, you have not included them in a similar context before. Why don't you ask for their permission before proceeding?	The system might advise against the disclosure simply because another person is in the photo if it considers "photo" or "selfie" to be inherently sensitive. If this is not the case, the system would not react at all. The system **does not consider**: • The past interactions between Bob and Alice. • The context of the photo.

	Phase 2: Although you have shared similar content with others before, this can be problematic in the current situation. The photo shows partying and alcohol in the background. The other person might have a cultural background that does not permit this type of disclosure, they might not be very extroverted, or they could have a strict workplace that deters them from this behaviour. Why don't you ask for their permission to make sure?	
Scenario 3: Alice enjoys travelling and recording candid **videos** of the locals and their customs. While in **Switzerland**, she captures one where the subject is a girl wearing traditional clothes at a festive event. She wants to share it on Facebook.	Phase 1: In your current location, Switzerland, the regulations forbid individuals from sharing visual content focused on other people even if it is not done for the purpose of redistributing the image for commercial gain. Why not get the people's permission first, if you have not already done so?	Photos in general can be sensitive. We advise you not to share this. The system **does not consider**: • The current location. • The specific customs and regulations associated with the current context. • Potential solutions that guarantee Alice some gain while respecting the individuals involved in the present situation.

	Phase 2: If obtaining permission is no longer possible then think about editing the photo to make the individual unrecognizable. An easy solution to this is to use a sticker to hide their face.	

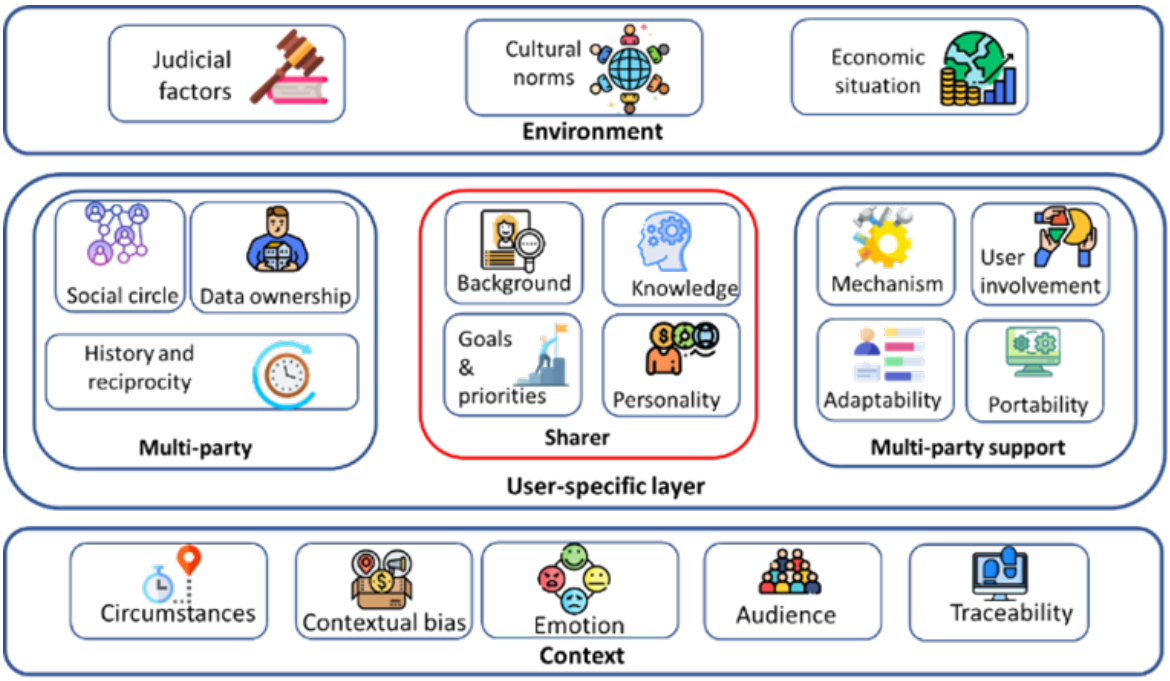

Fig. 4.1. Multipriv: The framework for MULTIfaceted PRIVacy decisions

The framework is represented in a layered model. From top to bottom, the criterion is genericity to the specificity of the factors that contribute to sharing information about someone else. In the following subsections, I detail each layer and sub-component but, at this point, what should be retained is that the environment is less subject to change and less specific to a unique user. Context, on the other hand, is the most capricious variable, which is specific to the user and the situation. Hence, if a user makes a different decision at t=1 from the one they made at t=0, it is unlikely due to a change in the environment and more likely to be a change of context. The model encompasses different modules or layers, namely:

• **Environment**: This layer combines judicial, economic, and societal factors that set the rules for disclosure. The impact of legally binding laws and behavioural constraints imposed by traditions or habits needs to be considered.

• **User-specific layer**: It refers to the middle layer. It is called user-specific because it focuses on the user also referred to as the sharer. Aside from their personal details, how they are influenced by multiparty specifics is important as well as the existing support tools.

- multiparty: As detailed in previous sections, a party, in this context refers to anyone other than the person sharing the information. This submodule details the factors related to the people concerned with the disclosure whomever they might be (family and close friends, colleagues/classmates, or the public), and in all forms (written text, audio recording, video, etc.).
- Sharer: The person who instigates the disclosure. Whether they were casually asked to do so by someone else, pressured into it, or decided to pursue it of their own volition, they are considered to be the party that shares data about themselves or another person.
- Multiparty support: This denotes the existing measures to reduce disclosure. They will be briefly discussed because they have already been detailed in Chapter 2, In general, the existing support mechanisms hinge on the cooperative aspect between the sharer and the other parties to reach a consensus. This makes it a "many to many" framework in which not only the intricacies of the single sharer are accounted for but also those of every other person involved.

• **Context**: It encompasses the surrounding circumstantial factors starting from who instigated the disclosure (who prompted it) to the bias compelling the sharer.

4.1.1. Environmental layer

This is the first layer at the top of the framework. The surrounding of the sharer shapes their perception and also restricts their actions. The environment consists of:

• **Judicial factors**: There are no comprehensive laws that address all the facets of disclosure. However, there are regulations and prohibitions that address the "right to one's own image" such as the one in Switzerland, which is included in scenario 3 (Table 4.1). Photos, for example, should not be published without the explicit consent of the subject even if the photographer does not intend to use them for commercial purposes. This does not apply everywhere around the world and in many countries, such as Canada and the US, there is no expectation of privacy in public spaces.

Moreover, there are laws specific to doxing, which refers to the intentional publishing of private or identifying information about a person or organization without their consent. Private data ranges from the full legal name to residential addresses, and governmental records such as *Social Insurance Number* (SIN) in Canada. The tricky part of the definition

as well as what makes it ambiguous and subject to interpretation is the word "intentional". If the sharer says that they made a mistake should they be given the benefit of the doubt and should the disclosure be written off as "unintentional"? There is a spectrum of differing laws from one jurisdiction to the other but the prevailing definition by lawmakers heavily relies on the intent. An example of this, in recent years, is the doxing regulation in Singapore. This act is officially classified as an offence under the *Protection from Harassment Act* (POHA). One of the examples provided is: Sharing a person's mobile phone number in a social media post with insulting remarks intended to harass them. This specific action is penalized by a fine of up to $5000 and/or jail for up to 6 months. One might wonder what becomes of someone who shares another person's mobile phone number out of animosity but does not include any insults. In this case, the intent is not as crystal clear, and the sharer can argue against this disclosure being maliciously fueled.

- **Cultural norms**: This is another environmental factor that has a major impact on the individual's perception. More conservative groups tend to value privacy more, in general. A problem occurs when the sharer and the other party do not adhere to the same cultural norms, which is the case in the 2nd scenario in Table 4.1. Another way to look at is through the lens of collectivism versus individualism. The former emphasizes the needs and goals of the group as a whole over the needs and desires of each individual. Collectivists are more accepting of the intrusion of groups and organizations into the private life of an individual. On the other hand, individualist cultures are more self-centric, and people tend to be more concerned about online privacy [**250**]. This is a concept that preceded the advent of SNS as discussed in Cutler's work on *interpersonal communication and technology* in cyberspace [**251**], which is the field of research focused on the impact of the advances in computer science on relationships or communication between people. The author explained that a sense of community creates trust and encourages individuals to seek support from the other members. This leads to the revelation of intimate aspects of one's private life.

This factor (cultural norms) is important and ties with multiparty disclosure because someone with a collectivist disposition might be more likely to share information about their neighbours for example if they perceived the behaviour to be harmful to the group. Someone on the other side of the spectrum might not, as long as it does not concern them in particular. This is far from being a general rule but multiple studies on privacy and culture have corroborated the correlation between the two [**252**]. Trepte *et al.* [**253**] pointed out that the cultural dimension crucially influences the perception of SNS risks and benefits and privacy as a whole. Within the United States, the study reported that 35% of Asian Americans, African Americans, and Hispanic Americans never managed their SNS privacy settings compared to a lower percentage of 21% when it comes to White Americans. Thinking

of the culture's impact on the perception of privacy in general and multiparty in particular, China's *social credit* is connected to both the judicial and cultural norms. Eleven years after its formal introduction, it has become a societal hallmark whose impact on the perception of privacy is yet to be fully investigated.

• **Economic situation**: The final parameter in the environment is the economy. From local to regional, or global, the economy plays a part in the drive for disclosure. Recently, during the economic stagnation in late 2019 and 2022 due to the COVID-19 pandemic, unprecedented threats to privacy have been reported. This is due to two main things: the desperation of individuals and the struggle of companies. The latter is not entirely the focus of Multipriv but, for context, it is because companies, going through hard times, seek to reduce costs mainly from departments that don't directly generate income such as internal audits. As a result, vulnerabilities are exacerbated and the data belonging to customers becomes an easy target.

Since Multipriv revolves around the individual, not organizations, the main area of interest, here, is the link between the economy and the sharer's decision. The declining economy affected the urgency with which Internet users share information about themselves or others. To highlight this point, we shall go back to the example of the actor Alice who is usually careful about privacy issues, under normal circumstances. However, following the specific situation in scenario 1 (Table 4.1), she felt frustrated and victimized, which is amplified by the global economic struggle. Following this, she attempted to vent out despite her better judgment. The effect of the global economic distress and her situation, in particular, compelled her to compromise her colleague's privacy and potentially even their livelihood. If a careful conscientious person can be susceptible to such circumstances, it is no wonder that the general public is more vulnerable than ever to social engineering when the economy is crashing. The *Consumer Financial Protection Bureau*, a US government agency, reported that during the pandemic there was a 54% increase in fraud compared to 2019 [**254**].

Before moving on to the 2nd layer, I would like to point out that judicial, cultural, and economic factors are intertwined. Regulations rely on the current state of the territory as well as the culture. Although they are detailed in this section one factor at a time, in real life, they impact one another.

4.1.2. User-specific layer

The middle layer focuses on user-specific factors starting with the multiparty details.

4.1.2.1. multiparty. This subsection details the factors relating to the other party that the sharer evaluates and takes into account before deciding to proceed in cases of multiparty disclosure.

• **Social circle**: It refers to the social group to which the person concerned with the disclosure belongs. There is a difference between publishing information about a close friend or a co-worker, etc. Three social circles are defined here, and they range from closest to most distant: close friends and family, co-workers or colleagues, and finally the general public. Alice is very close to her mom, so, she makes sure to consult her before posting a photo of the two of them, but she has no qualms about sharing a selfie taken in public regardless of who is in the background. For this reason, the social circle plays an important and sometimes even decisive role in the sharer's decision.

• **History and reciprocity**: The second component, reciprocity means accommodating the other person as much as they themselves have been considerate towards the sharer. If Bob had never asked Alice's opinion before sharing her information, Alice, in return, would not feel inclined to ask for his approval either and vice versa. In general, people tend to pay back good intentions with equally good behaviour except for blatantly malicious people, which is a point to reiterate later in the sharer parameters (in the same layer). Since the ultimate goal is to provide the best nudge framing, the history component includes the interactions and how previous disclosure situations were solved. Going back to scenario 3 where the system suggests that Alice either blurs or posts a sticker over the other individual's face, the system would propose the solutions that have the highest likelihood of being accepted.

• **Data ownership**: In general, ownership refers to the possession or proprietorship. In particular, in Multipriv, it is basically an indicator of how much of the shared data belongs to the multiparty. It is devised to draw inspiration from existing privacy frameworks that focus on how corporations and organizations handle personal data. This is already far more complex with intangible assets like data in comparison with tangible goods, but it gets even more complicated when it involves multiple people. When a company intends to store user data, in their terms of service, they usually specify the duration and the ownership rights to the information. So, companies can be held liable in case of a breach of these agreed-upon terms between the individual user and the organization.

Human relationships do not come with a manual that showcases the right to a piece of information or how to divide assets in case of disagreements. A lot is left to interpretation. In terms of specific situations, let us consider the following setting: If Alice takes a selfie with her friend Bob and he agrees to be in it, she might assume that it is he is also implicitly agreeing to share it. In this case, she would think she has 100% full ownership of the photo.

Bob, then, expresses his dissatisfaction as he believes that she crossed a boundary and that the photo is his as much as it is hers so, she should not have free reign over it. I shall take it one step further: If Alice believes she has ownership of the photo she might add some popular filters to it that make faces look funny and distort it. She sees nothing wrong in this because she edited a photo of hers and shared it. Bob is not fond of these trends, and he feels like she is out of line. Ultimately, determining who has true ownership over the photo is what decides who is in the wrong.

Even the law does not have an obvious ruling in such situations. Admittedly, according to an assessment in the *Michigan Technology Law Review* [**255**] copyright ownership vests in the author of a work, and in the case of photography, the author is normally the person who literally creates the photograph by pressing the shutter. The article discusses authorship from a legal point of view and argues that even this definition of photo ownership is not indisputable as the author can equally be someone who chose the lighting and the creative vision for the photo but was never directly involved with capturing it. Hence, a question arises: if the law itself cannot provide an indubitable attribution of ownership then who has the authority to do so? In the user-centric framework, Multipriv, which revolves around the sharer, ownership relies on their perception as they are the decision-maker, which brings forth the next subsection.

4.1.2.2. Sharer. The sharer is the person who decides to disclose information about others. Their action is impacted by many factors such as their background, knowledge, etc.

• **Background**: This includes their gender [**256**], age [**257**], the highest level of education [**258**], main occupation [**259**], and origin [**260**]. These factors have been correlated with the perception of privacy in self-disclosure situations, I extend them to multiparty disclosure, as well. Each parameter cannot single-handedly influence the sharer, instead, they are indicative of more determinant factors such as knowledge.

• **Knowledge**: A user who has a higher diploma in computer science and cybersecurity is more informed on the subject and aware of the potential consequences. Knowledge comes from many sources and sometimes experience teaches individuals the most. People who have been victims of fraud or identity theft or have seen it happen to others become acutely aware of hackers and scammers. Let us suppose that Alice's mom is the main actor in this scenario. She is not up to date on recent technological developments and has no idea what phishing is nor how it is possible. However, a close friend of hers fell victim to an email disguised as a communication from the government urging people to fill out the attached form in preparation to get monetary help because of the COVID-19 pandemic. The mom recalled this incident and avoided the scam.

• **Goals and motivations**: They can tip the scale in favour of revealing multiparty data or concealing it depending on the user's perception. If Alice possesses information of value about her colleague Sam who confided in her how he managed to overcome a health issue through a specific diet and practices, Alice can sell this information to a news outlet for monetary compensation. The mental calculus she does compares the money she would get versus having her relationship with Bob deteriorate and potentially other colleagues seeing her as an opportunist. She still decides to do it because, to her, money is higher on her list of goals and priorities compared to maintaining relationships. Next, more examples of aims are included, specifically, three goals are considered, inspired by[**261**]:

Financial gain: The motivation is monetary gain. It can take the form of cash, digital, or virtual currency such as Bitcoin, discount codes, and vouchers. An example of this would be: filling out surveys for gift cards for specific retailers. In fact, this has become a lucrative business where applications and websites host surveys. If the user responds with their own information as well as that of their household (parents, siblings, partner, kids) such as supermarkets they shop at the most or holiday destinations they would like to go to, they receive points.

Personal gain: This encompasses all nonmonetary services such as exclusive access to premium services. To illustrate this type of gain, Alice intends to sign up for a gym membership and is told that she can be granted access to a more spacious area if she signs up another family member. She agrees and in doing so discloses information about her mother including her full name, and personal phone number.

Moral gain/altruism: The user who aims to achieve an altruistic goal does not expect any form of money or service in return. Motivated by a sense of morality and virtue, they think that their disclosure contributes to the better good of society. Case in point, Bob is hospitalized and incapable of paying his medical bill. Alice takes it upon herself to start a go-fund-me campaign for him. To reassure the people who want to donate that this is not a scam, she includes specific information about his family situation, financial struggle, illness, and procedure he has undergone. This is a major case of multiparty disclosure regardless of how well-intentioned it is.

So, why is identifying the disclosure goal relevant to improving the framing of nudges? For example, if Alice is highly motivated by financial gain, a nudge highlighting the monetary value of what she is gratuitously divulging can resonate with her better than a different approach. There is existing research on how to put a price tag on personal information based on the value accorded by brokers, illegal sellers on the dark web, and groups of individuals

who are the owners of the data [**262**]. It is a subject that I have also previously investigated [**263**].

• **Personality**: It plays a big role in the sharer's actions. A glimpse into this was part of the previous layer when explaining the judicial factors. As discussed before, there is an ambiguous use of "intention" when dealing with doxing, which ties in with the sharer's personality. A conscientious person, for example, would carefully think about the ramifications of their actions. Going further into this point, the earlier example shows Alice being frustrated over the lower wage compared to Sam (scenario 1, Table 4.1). After being reminded of the harm she might cause her co-worker and being conscientious, she might be more likely to refrain from posting the content.

Another personality trait that Multipriv considers is empathy, and this is relevant in the context of putting others at a disadvantage. If someone is concerned about their own privacy and goes to extremes to preserve it, with a high level of empathy, they are likely to be equally concerned for the other party. This could again be Alice when she is nudged to focus on the potential harm to Sam were she to involve him in this situation. The third personality trait that is considered is agreeableness. This will, in fact, be connected to the 3rd layer because depending on who is asking for the information, assuming that the disclosure is prompted by a third party, the decision can change. These are but examples from the *big 5 personality traits*.

4.1.2.3. Multiparty support. There are a few characteristics that determine how successful support tools are.

• **Mechanism**: It can mitigate the disclosure, but it mostly relies on the representation of everyone involved [**12**]. In game-theoretic approaches, for example, the preferences of the sharer and everyone involved are known by respective agents who negotiate to reach a consensus. To put this in terms of a sharing scenario: the support system is familiar with both Alice and Bob. It knows that Bob dislikes sharing his geotagged location. Alice is about to share a photo she took with Bob with a geotag. The automated agent representing Alice who wants to go fulfil her desire to share negotiates the issue with Bob's representative (another automated agent). Following specific protocols and strategies, they convey their compromise to Alice: remove the geotag and keep the photo or keep everything as it is but remove Bob from the photo (through blurring or cropping).

• **User involvement**: It is also referred to as user engagement. It stands for the ease of use and effort that the user must dedicate to using the support tool. Basically, if a system keeps asking the sharer to make complicated decisions or if it requires their manual input at every stage of the process, this can become burdensome very quickly. This is why the

user's involvement is complex because on one hand, users can feel in control and express their preferences but on the other hand, they might not have the necessary prerequisites to do so.

- **Adaptability**: It refers to the ability to keep up and evolve with the user. If Alice keeps rejecting the system's suggestion to blur her friend's face repeatedly, the support tool needs to adapt to this preference. It might ask her instead to schedule the post for later to offer more room for thinking. In other cases, the system might deduce that the other party is her mom and that the disclosure is agreed upon and requires no guidance in the future.

- **Portability**: It is yet another support criterion and is the system's ability to keep up with Alice across different platforms. If she needs to get reacquainted with different solutions, one per social media platform, that is tedious for her and she might stop using all of them. This can be solved through portability, which is usually associated with software usability and integration across different environments. This aspect is important to improve the nudges.

4.1.3. Context

The final lower layer is the most specific one. In other words: For the same user Alice and the same multiparty member Bob, using the same support tool and within the same environment (the same top and middle layers), two different contexts can lead to widely different outcomes. Context is referred to as the most specific factor because the same person does not change their goals and personality overnight, but the context might. The different factors taken into account are the following:

- **Circumstances**: they represent two main things: the geo-temporal setting and the initiating prompt. The former explains why someone who is at work during the day might have come to a different conclusion than if they were at home relaxing on the weekend. The latter part of circumstances classifies the disclosure as prompted or unprompted. While unprompted disclosure starts with the user's decision to share the information, the former is the opposite. Someone would have to ask the sharer to tell them a piece of information, following which multiple parameters are at play to determine the outcome. In the middle layer, specifically in the personality factor, we mentioned agreeableness as something that is intertwined with the context: Following a prompt by someone close to the user, the degree of agreeableness can push them to disclose.

- **Contextual bias**: It encompasses Cialdini's principles of persuasion and all the biases detailed in Chapter 1. Reciprocity, scarcity, authority, commitment/consistency, liking, and

consensus (or social proof) can be exploited to compel people into making seemingly counterintuitive decisions. For instance, if user Alice has a close friend Bob who has always done favours for her, she might be inclined to share confidential work information with him that she would otherwise not do. In fact, she might feel like the situation required reciprocity thanks to his previous acts of altruism. Another bias is the *Fear Of Missing Out* (FOMO). Social media sites facilitate oversharing and make it a social event of sorts. Users are encouraged to post a status update, a photo, an event, or a "check-in" with the click of a button and to follow emerging trends. This can lead to an echo chamber of oversharing.

• **Emotion**: To put it simply: people can be driven by impulses. Alice is usually careful and avoids uncalled-for multiparty disclosure, but she is very angry and frustrated due to her workplace situation (scenario 1 in Table 4.1). She is about to reveal personal incidents and mishaps that Sam was involved in, which could not have been resolved without her intervention. Alice thought of this in the spur of the moment to quell her negative emotions and it clouded her judgement. This is not specific to frustration or anger, all intense emotions can perturb someone's thoughts or assessment abilities to a certain degree. Bob is feeling impatient as he rushes to the airport and receives an email from Alice's co-worker with whom he is not well-acquainted. Sam asks about Alice's address for an emergency at work. Bob would not have replied positively to the request under normal circumstances but this time, he does. Emotions can override knowledge sometimes.

• **Audience**: It refers to the group of people that can view the content. A malicious or maybe vindictive person might be motivated by publishing a defamatory post to the largest possible group of people. Another person can decide against it. If we go back to the first scenario in Table 4.1, the fact that Alice's colleagues and boss could see the post motivated her to vent her frustration and commit to the disclosure, which is a testament to the power of having the target audience see the intended post. If Alice was sure no one from work would ever see it, she might not have done it. The opposite can be said about some other situations where the user simply wants to relieve their stress by explaining the work situation to close friends and family.

• **Traceability**: It refers to the perceived anonymity and the trust that the sharer has in the medium. For example, Alice has an issue at work related to her boss' behaviour and how he assigns her too much work while underpaying her, but she is hesitant about confronting him. She wants advice and describes specific details about the situation and her boss in a Reddit post to get advice. Her trust in remaining unknown is crucial to her. It is the reason why she chose this platform.

This section has elaborated on all three layers going from the most generic to the most specific. Nonetheless, based on the current technology, one might wonder how feasible implementing such a framework is.

4.2. Challenges

Detecting the disclosure is a core part of the process following which the framework can be used to better frame the nudges. This is an expanding area of research that has gained momentum, especially in the past decade [**264**]. Starting with text-based disclosure, Petrolini *et al.* [**265**] use transformer-based classifiers to automatically detect sensitive data in written form. The categories that the authors base their work on include "politics", "health", "religions", and "sexuality". A focus on social media exclusively has been developed such as detecting privacy leaks in the form of vacation plans in Twitter posts [**266**]. Moreover, while going through the different components of Multipriv, I explained the need to put disclosure within the legal scope (applicable regulations) and that is not far-fetched with the existing research. The findings of Tesfay *et al.* [**267**] allow the detection of sensitive information in accordance with the norms defined in the EU's GDPR.

This progress is not unique to text input, photo, and video-based disclosure detection have come a long way, as well. An example of this is detecting children's faces in real-time with an accuracy of 92.1% [**268**]. However, we should also mention that this process is not flawless and that some forms of disclosure remain challenging. Table 4.2 shows some scenarios that are harder to tackle. It is a challenge to design nudges for wearable devices in particular (scenario 4). Even if we do not consider the small dimension of the screen and the low processing power of the device, there are more hurdles to overcome. A smartwatch is worn while on the move and it is unlikely that a jogger or hiker is going to stop because they were nudged.

Scenario 5 raises another issue: The detection of loneliness, which is the first step in the process, is not easy. Being impulsive can be detected to a certain degree such as the user using all capitalized texts, strong language, expressing anger, etc. But, being lonely is harder to automatically figure out. I can think of a few ways to approach it, but none are conclusive. For example, if a user stops being physically active, which can be detected through their phone or wearable device, this can be a sign of loneliness [**269**]. Another sign can be receiving fewer texts and phone calls [**270**]. Another idea is leveraging GPS, which all smartphones are equipped with to detect the user's location throughout the day. If they are spending most of their time at home, research has backed up the hypothesis that this leads to more loneliness compared to individuals who go out more [**271**].

Tableau 4.2. Scenarios that are more challenging.

Scenario	**My proposed enhanced nudge**	**Challenges**
Scenario 4: Alice has started exercising with her friend Bob. She is getting into the habit of proudly sharing the information her wearable health-monitoring watch records including her hike trail, the length of their exercise, and the time to finish it. Her social media account **easily synchronizes** with her watch without any deterrent (warning, nudge, or any multiparty support.).	The information that you are about to share can be dangerous because you are disclosing your current location, which is quite secluded and less frequented by the general public. You are advised to only share your new record (duration of the hike) without the location.	• The small dimension of the screen. • The low processing power of the device. • Difficulty in designing nudges for users on the move.
Scenario 5: During **COVID 19**, Bob has gotten **lonelier** and decides to make a profile on a dating app to meet potential romantic partners. When he did not get matches, he decided to post **more personal information** and he made his current **location** public with virtual strangers.	You are about to share details about your personal life that you have previously considered sensitive. Would you like to reconsider and remove these details or delay the post for further revision?	• Detecting Bob's loneliness. • Convincing the vulnerable Bob not to disclose personal data.
Scenario 6: Alice lives by the principle of "I am an open book and I have nothing to hide". This is reflected in her social media activity as she leaves almost nothing to the imagination.	You are about to share content that can be uniquely identifiable when combined with your recent posts. Would you like to edit the post or view your recent content to check what you have previously shared?	• Fallacies are harder to tackle than lapses of judgment.

Finally, scenario 6 presents the "I have nothing to hide" fallacy that is discussed in Chapter 2. As Solove pointed out, it is wrong to equate not doing anything bad with revealing one's private information. The two are different and this misinterpretation can lead to dire consequences such as being monitored and stalked. The difficulty of scenario 6 is how to approach Alice and rectify her perspective. While nudges are great as reminders and

snippets of educational content, they are not the best at countering logical fallacies. The latter type requires more extensive explanations, examples, counterexamples, and usually longer discussions. My proposed enhanced nudges can spark the conversation, but there needs to be a more robust long-term approach to address cases like Alice's (Scenario 6).

4.3. Discussion

This chapter details the intricacies of the proposed framework called Multipriv and how it merges human, societal, and a few technical factors. In the following, multiple points and lessons are discussed including a comparison between Multipriv and some of the existing frameworks and models.

Self-disclosure: A catalyst for multiparty disclosure

There is something very human about the perception of disclosure. It is similar to making jokes about someone else, if the person telling it starts with a self-deprecating funny remark, in their mind, they do not feel as bad following that with a joke at others' expense. The mentality of "I made myself look worse than you so you should not be complaining" is something a lot of people practice in their lives. A similar analogy can be drawn with disclosure. The sharer Alice can start with a bad personal experience at work and then, move on to speaking about another colleague. Alice can perceive her action to be inoffensive and not to the detriment of her colleague's privacy. This also ties in with the fact that self-disclosure often stems from a lack of concern for one's privacy, due to recklessness or being unaware of the consequences. Both of these are equally problematic on the multiparty level.

Shifting the interest to human-centricity in privacy

Let us take a step back and look at the 90s when the main threat to privacy was theorized to be the systematic side of technology. The concept of *Privacy by Design* (PbD) was coined to address the effects of *Information and Communication Technologies* (ICT) and prioritize privacy as the foundation rather than a feature or an after-the-fact measure. At the time and before the rise of SNS, the general idea was that as long as companies were kept in check, privacy would be preserved. Three decades after this, it has become clear that human actions are often the most impactful. This does not mean that major data breaches hold no weight but there is a general awareness of the impact of such incidents. Companies prepare audits to counter or mitigate them, etc. On the other hand, there are no clear plans on how to address the lack of awareness among individuals. Even when protected by the strongest laws, judicial power does not equal awareness. The latter comes before the fact whereas penalties come after it. It is becoming more apparent that educational tools and positive reinforcements are needed now more than ever. Someone who lacks the needed knowledge for

privacy preservation could be offered the most secure cutting-edge system and still sabotage themselves.

Comparison with existing frameworks and models

This chapter's contribution is the introduction of Multipriv, which provides a comprehensive view of the factors that impact multiparty disclosure because this is a necessary step for the future of raising awareness of the issue. Knowing what sets in motion, amplifies, or mitigates the action of sharing can offer a deeper understanding of the sharer themselves. This is different from existing frameworks that focus on the organization side such as the *Privageo* [**272**] representation that details seven factors: *Legal and risk management*, *third parties*, *personal information management*, *data protection*, *privacy leadership and control*, *Communication and training*, and *consumer trust and consent*. Similar to this, the *National Institute of Standards and Technology* (NIST) proposes a framework centred around the organization and aims to examine the framework risk[1]. The goal is to improve privacy through risk management. It is evident that neither frameworks apply to multiparty disclosure by individuals, which is the current void that my framework aims to fill. More details are showcased in Table 4.3.

ISO, in the table, refers specifically to ISO 27701: 2019[2], the extension to ISO 27001/2[3] for privacy information management. The considered NIST framework is the latest release called "*NIST Privacy Framework: A Tool for Improving Privacy through Enterprise Risk Management*"[4]. The framework and model designed by the *Online Trust Alliance* (OTA) is specific to the field of *Internet of Things* (IoT). All of these examples are organization-specific and tend to be used especially by larger companies. In this sphere, user-centric privacy solutions are scarcer and tend to adopt traditional one-size-fits-all approaches such as mandatory training followed by evaluations to test the information retention ability of employees. For comparison, the model of perception of privacy [**12**] was included in the table to highlight the interest amongst academics to tackle the subject, albeit, multiparty disclosure remains largely unexplored.

Multiparty disclosure: A force for good, or is it?

Another point to discuss is that multiparty disclosure is not inherently wrong in all situations. If that were the case, then whistleblowers who reveal company misconduct or fraud would not be celebrated as they are. They should not be discouraged from doing so just because they are revealing private information about the CEO for example. This brings forth another

[1]Source: `https://www.iso.org/obp/ui/#iso:std:iso:31000:ed-2:v1:en`
[2]Source: `https://www.iso.org/standard/71670.html`
[3]Source: `http://www.vulnerabilityassessment.co.uk/iso27001.html`
[4]Source: `https://www.nist.gov/system/files/documents/2020/01/16/NIST%20Privacy%20Framework_V1.0.pdf`

Tableau 4.3. Comparison of Multipriv with some other framework or models.

Framework or model	**General aim**	**Focus**	**Multiparty consideration**	**Year**
Privageo privacy framework	Roadmap for secure data management	Organization-centric	No	2019
A Model of Perception of Privacy	Factors in correlation with self-disclosure	User-centric	No	2019
ISO NIST OTA	Risk management Breach prediction and mitigation	Organization-centric	No	2019 2020 2020
Multipriv	Framework for multifaceted privacy decisions	User-centric	Yes	2023

debate "at what point does a person lose their right to privacy?". Is that only for convicted criminals or does it apply to those suspected of unethical yet legal actions? The problem is that as soon as one exception is accepted, others pile on. Revealing misconduct can easily become a form of harassment as revenge.

The dilemma of nudging: ethical considerations

Most ethical concerns stem from the notion of autonomy and whether nudging compromises it. Whenever it comes to persuasion or guidance, the issue of freedom of choice always arises. This is particularly prevalent when the designer of the nudge seeks their own interest over the person receiving the nudge. An example of an unethical situation is the following: A privacy nudge redirects the user to "get more information" on a specific website, which benefits the designer through advertisement.

I think that a framework like this, which contributes to personalized user-specific nudges operates within the perimeter of ethical nudging thanks to informed consent. The offered set of guidelines is designed for the social good of the sharer and others. No one is manipulated

into using the eventual system (part of which is disclosure detection and resolution). The purpose is explained to each individual and it is up to them to accept or refuse before the system has any information on them. If their free choice changes through the means of a reasonable augment that manages to convince them, it does not negate their right to freedom any more than a conversation with an acquaintance does. A changed behaviour or decision is not necessarily one achieved through manipulation and underhanded methods or else any awareness-raising campaign is guilty of this. Nature itself is full of nudges that encourage specific actions. If it is windy and raining outside, a person feels somewhat "compelled" to drink something hot and comforting. If we establish that nudges are not innately unethical, then, the next step is to make them as transparent as possible and to keep the user in the loop.

Takeaway message

Privacy is a very crucial topic in today's world as oversharing impacts not only the individual but also their social circle. Moreover, contrary to prior technological innovations, the Internet has made it possible to get real-time feedback and feel the impact of the disclosure as it is happening. Someone could have their entire life flipped upside down in a matter of minutes. It is no wonder that the career of some individuals ends abruptly after someone digs up something they tweeted 10 years ago.

This is a call for more privacy awareness, educational tools, preventive positive reinforcements such as nudges [**261, 273**], gamified approaches, etc. The point on which I would like to insist is that privacy-protecting tools have indeed come a long way and can potentially solve several issues, but they are not a substitute for user awareness. Support and coping mechanisms [**274**] can never take control and deny the user the right to disclose whatever they want. Since the sharer is at the centre of potential issues, they should also be the centre of the solution. Many researchers focus on making this process fun and attractive through gamification with varying degrees of success [**275**]. Moreover, a change of perspective could prove to be beneficial to solving the problem. Privacy is often seen as something that impacts the individual. Hence, encouraging someone to adopt proper measures, they can be told "Make sure not to respond to suspicious emails because you can get your identity stolen and you will not be able to live a normal life". Although the impact on the person is undeniable, it is not the whole picture. The impact on family, friends, and even colleagues can be colossal. To deter someone from compromising their privacy and that of their social circles we should start reminding them of the consequences and whom they can hurt other than themselves. Another point to get across in the awareness-raising process is that every action despite its seemingly small magnitude leaves traces of data everywhere [**276**] leaving a personal digital print.

4.4. Conclusion

This chapter is the first step in a roadmap towards disclosure mitigation by understanding the catalysts and factors leading to it. The framework Multipriv in itself is not a solution nor a standalone work. It is a guideline to improve future approaches to enhance multiparty privacy-preserving nudges. It highlights considerations overlooked before, such as the environmental and contextual impacts on the sharer. I acknowledge, however, that this is a multifaceted multidisciplinary issue that requires collaborations from all sides from human psychology to technological factors (the tools) and socioeconomic parameters. Some researchers in law call for a flexible framework that goes beyond privacy. Bartlett [**277**] argues that the existing laws are mere "shoehorn policy responses" stating that Silicon Valley should not be the only place for innovation and that regulations can be equally adaptive. On the ethical and social side, Sharon *et al.* [**278**] revitalize Gofman's civil inattention, the social norm of showing a proper amount of indifference to others as a means of privacy preservation. The introduced ethics of inattention can be explained as follows: If no one intrudes on another person, no one's privacy is truly compromised. If such ethics become embedded in the beliefs of society, it can be very positive for the development of youth. Without privacy, children and youth's ideas of self, trust, and authority are affected[5].

Finally, my takeaway is that being active on social media is tantamount to being behind the wheel. Unintentional disclosure, due to the sharer being ignorant or self-absorbed, is similar to reckless driving. The person is not motivated by malice, but they have no regard for other vehicles or pedestrians, which can cause harm to the driver and everyone around them. Everyone has somewhere to go, a goal to achieve, and is subjected to a certain pressure including, amongst other factors, time constraints. However, one cannot simply drive hastily into traffic as if their own motive overrides the environment itself.

A parallel can be equally drawn with intentional disclosure, where the objective is to bring detriment to others. An aggressive driver can decide to tailgate the car in front of them out of annoyance. This person might be prone to such behaviour on a regular basis (personality trait) or it might be fueled by the current context (circumstances, emotions, etc.). Either way, one thing remains true, which is that they wanted the other driver to be inconvenienced or maybe even suffer severe consequences. The 2020 movie "Unhinged" comes to mind as the dramatization of the real problem that is road rage. In a similar way, virtual discord on social media can escalate very fast and implicate many people.

The world's first comprehensive traffic code was introduced in 1903 to protect each driver on their own, but also everyone around them, be it in their respective vehicles or passerby

[5] `https://priv.gc.ca/media/1731/yp_201003_e.pdf`

pedestrians. It is high time we established a privacy code of sorts to achieve collective welfare on social media. The next chapter takes a step further in this direction by tackling the proposed nudge-based system.

Chapter 5

Multi-agent Nudge-based Approach for Disclosure Mitigation Online

In Chapter 2, I explained the notion of nudging as a means of "*helping individuals steer away from irrational behaviour*" [**199**]. I also detailed the types of nudges and the ethical concerns around the topic. In section 2.5 of the same chapter, there is an explanation of the reason why I opted for this term throughout the dissertation over "recommendation". Furthermore, Chapter 4 is dedicated to explaining my framework Multipriv, which lays the foundation for my proposed context-aware personalized nudges that this chapter delves deeper into.

While the factors leading to disclosure have been investigated and explored in the literature, to the best of my knowledge, none of the researchers have incorporated their findings in a disclosure-mitigating solution. In other words, the existing work focuses on one of the following three approaches: identifying said factors [**279, 280**], proposing nudges based on the user's preferences [**281, 282**], or proposing generic one-size-fits-all nudges [**283**].

In this chapter, section 5.1 gives an overview of my multi-agent nudge-based approach for disclosure mitigation. It offers a novel perspective on the subject by adding a voice of reason agent called *Aegis*. Additionally, the proposed multi-agent system handles both self-disclosure and multiparty disclosure thanks to the *personal* and *multiparty* agents. Section 5.2 is centred around one module of the aforementioned system, which is domain knowledge. It offers an understanding of the notions and concepts based on which nudges are eventually pushed. This includes the *disclosure categories*, *motivations*, *behavioural biases*, and *data valuation*. Section 5.3 is dedicated to the disclosure detection module. This is paramount to trigger the rest of the system to respond when mitigation is needed. Both text and image formats are handled. Section 5.4 details the context-aware user modelling process, specifically, the perceived data sensitivity and the disclosure appetite. The latter is a concept I coined inspired by the term *"risk appetite"*, which is traditionally known as the amount

of risk a financial organization is willing to take in pursuit of objectives it deems valuable [**284**]. Establishing the user model leads to the definition of each actor in the aforementioned multi-agent module as described and explained in section 5.5. Each of the agents represents a party involved in the disclosure starting with the user who is represented by the personal agent. Aegis is a privacy-focused objective agent that does not cater to the individual's preferences. The multiparty agent stands in for the other people involved in the user's post. The mediator acts as an arbitrator whose objective is to reach a consensus that appeases all parties. In section 5.6, I discuss the implications of the proposed system and explain points that can serve to quell some of the ethical concerns that might arise with such a proposition. Finally, this chapter ends with a conclusion that acts as a reminder of the objectives achieved throughout it as well as an occasion to put this piece of the puzzle in the wider context of privacy preservation on SNS.

5.1. General architecture

The goal of the proposed system is to mitigate disclosure and alter the behaviour as it is about to happen while considering both the personal and multiparty aspects. Figure 5.1 illustrates the steps that the system undergoes before pushing a nudge.

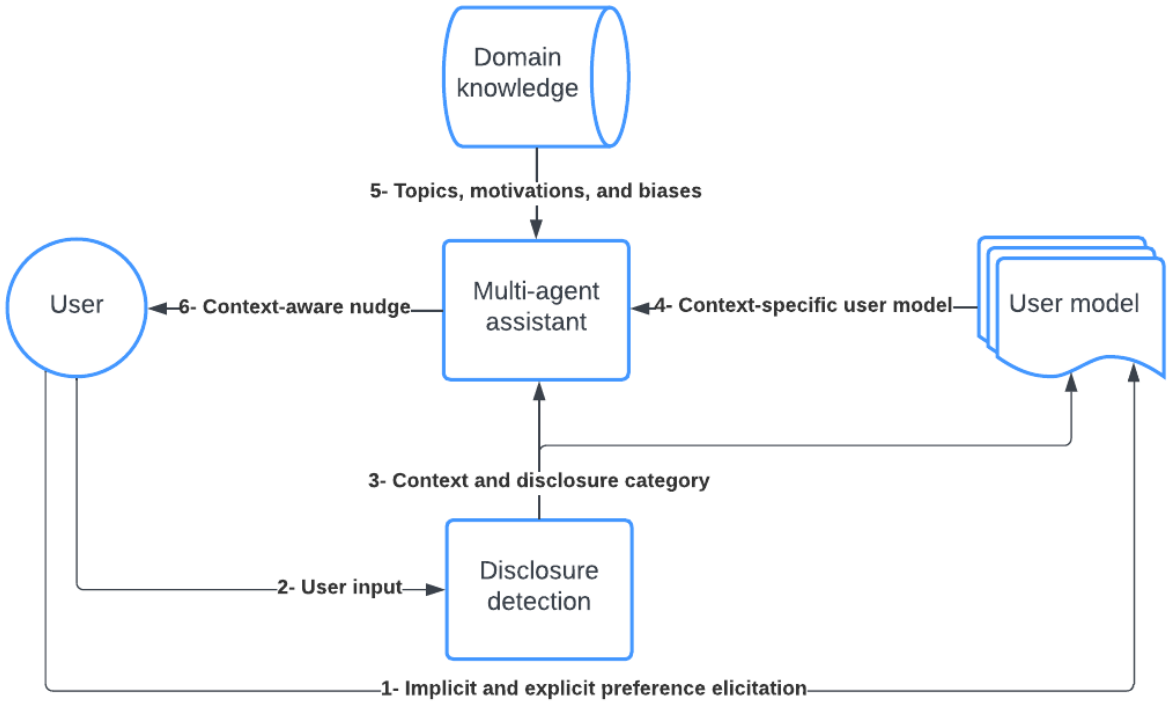

Fig. 5.1. General architecture of the nudge-based system

The *first* step is to construct the user model based on the individual's past posts (implicit) and the answers they provide to a questionnaire posed by the system (explicit). This allows the system to overcome the *cold start problem*. This term refers to situations in which the individual's preferences and behaviour are either completely or partially unknown to the system, which limits its ability to make customizations and to address the user personally [**285**]. *Second*, comes disclosure detection, which happens when a user is about to share any content whether it is text or image. If deemed sensitive, the multi-agent assistant is activated. As a result, it receives the context and the disclosure category as the *third* step. The *fourth* and *fifth* steps feed the user model and the domain knowledge to the same core component, both of which are needed for the negotiation process between the agents. This is detailed in section 5.5. The output of this process is the context-aware personalized nudge that the user receives as the *sixth* step.

5.2. Domain knowledge

In data science, the term domain knowledge is used to refer to the general background knowledge of a specific field or discipline in which a system operates. In this case, it offers the foundation for all the modules to perform their assigned tasks. The disclosure detection, user modelling, and nudge-pushing tasks cannot be performed without knowing which topics users tend to share, which motivation guides them, and which biases they can be susceptible to. This section is broken down into three parts: disclosure topics, disclosure motivations, and disclosure biases.

5.2.1. Disclosure topics

Let us start by considering a scenario to clarify what is meant by the term topic in this section. Alice posts the following Facebook status update about her upcoming vacation: "*The girls and I have booked a stay at the amazing SLS Baha Mar hotel for this weekend. We are going to have an amazing time together Julia Samantha Kendal #girls-trip #SLS-Baha-Mar #Bahamas*". The topic can be defined as "*Experiences*", the pieces of data disclosed are: *vacation hotel*, *vacation time*, and the *friends tagged in the post*. This is an instance of both self-disclosure and multiparty disclosure involving Julia, Samantha, and Kendal.

There are two main purposes for modelling disclosure topics:

- The system cannot detect what it does not know. As such, in order to detect the disclosure and push nudges, it needs to establish the topics first.
- Another goal is to study the interests of people as groups. This allows the system to detect patterns such as *female individuals aged 24-35 being the most likely demographic to disclose travel data*. The importance of this, beyond reporting findings relevant to specific groups, lies in the potential of bypassing the cold start problem

for unknown users, which has been explained in Section 5.1 and will be tackled in Section 5.5.2. If Alice's user model is missing some key parameters needed for future intervention, they can be predicted thanks to the other users' models.

Topic modeling is a Natural Language Processing task that allows the unsupervised discovery of topics in a collection of documents. The chosen method is *Latent Dirichlet Allocation* (LDA) [**286**] in which the term "latent" refers to the fact that the topics are yet to be determined. I selected it because it is known for producing sensible meaningful topics [**287, 288, 289, 290**]. Moreover, it performs well on short texts [**290**], which is a characteristic of social media posts. Furthermore, LDA has been tested on tweets previously, although it was solely for sports subtopics, nevertheless, it outperformed the other methods [**291**]. Despite these encouraging results in the literature, the output should be examined by humans before finalizing the process.

As for the number of topics, which is a manually set parameter, a large number is generally discouraged [**292**]. I test whether this number of topics is optimal or not thanks to two metrics *perplexity* and *coherence*. The first indicates how well the model describes the data: a lower perplexity suggests a better fit. It captures how a model is expected to perform on new data that has not been seen before. The second evaluates the quality of the topics based on the assumption that words with similar meanings tend to co-occur within a similar context [**293**]. A high coherence value is desired but the goal is not to continuously increase it or else we might end up with 100 topics, in extreme cases. The objective is to figure out the point after which the increase in coherence score is no longer worth the increase in the number of topics. Both metrics are calculated as shown in equations 5.1 and 5.2 based on the work of Mimno *et al.* [**294**] and Newman *et al.* [**295**] :

$$Perplexity = \frac{-\sum_{d=1}^{M} \log p(W_d)}{\sum_{d=1}^{M} N_d} \tag{5.1}$$

where:
$p(W_d)$ is the probability of observed words of document d.
N_d is the total number of words in document d.
M is the number of documents.

$$Coherence = \sum_{i=1}^{j-1} Score_{\text{UMass}}(W_i,W_j) \tag{5.2}$$

$$Score_{\text{UMass}}(W_i,W_j) = log\frac{D(W_i,W_j)+1}{D(W_i)}$$

where:
$D(W_i)$ is the frequency of documents that contains the word W_i.
$D(W_i,W_j)$ is the frequency of documents that contain both words W_i and W_j, and D is the

total number of documents in the corpus.

To model the topics as part of the domain knowledge, I used a dataset [**296**], which contains a pre-processed sample of 12811 Facebook posts, specifically status updates, collected for research purposes. I used *Gensim*[1] to plot the coherence values as highlighted in Figure 5.2.

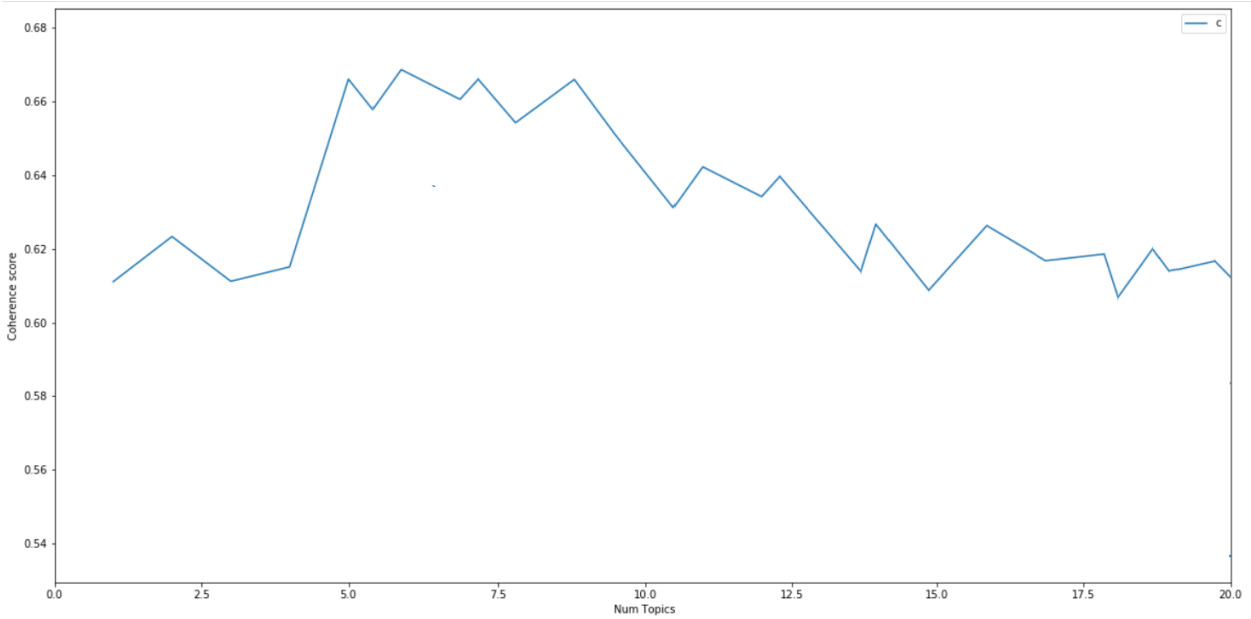

Fig. 5.2. Coherence score as a function of the number of topics

Beyond number 5, there is no significant increase in the coherence to warrant increasing the number of topics. As such, I thought about choosing five topics, but first I had to further check the optimality of this decision using the perplexity. Figure 5.3 shows the variation of this metric as a function of the number of topics and this process uses the same aforementioned library, Gensim.

For a number of topics between 5 and 7, perplexity is the lowest. Considering this with the coherence values, the number of topics is fixed as five and they are as follows after I manually labelled them:

(1) Identity (race, gender, sexual orientation).
(2) Location (most frequented restaurant, workplace, etc.).
(3) Experiences (interpersonal relationships such as dating, childhood trauma, etc.).
(4) Alcohol, drug consumption, and health information (health records for example).
(5) Religious and political views.

[1] `https://pypi.org/project/gensim/`

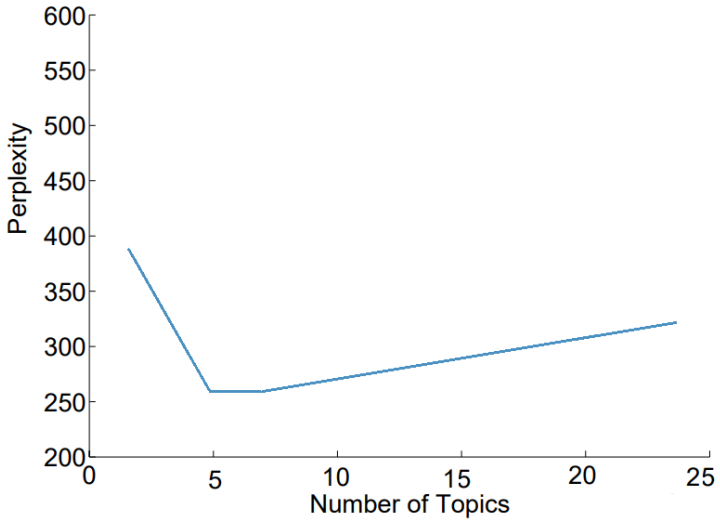

Fig. 5.3. Perplexity as a function of the number of topics

It is not enough to push nudges solely based on the disclosure topic because users share personal data for various motivations. Identifying which one is behind the current post offers an understanding of the user's decision-making process, which helps to push better nudges.

5.2.2. Disclosure motivations

In this section, I elaborate further on what has been partially explained in Subsection 4.1.2.2 of Chapter 4. Motivation plays a major part in the decision-making process leading to disclosure. Alice might not gratuitously reveal her home address to anyone who asks for it, but she might if she responds to a message on SNS promising free samples of a novelty product. In a similar case, Bob who is feeling lonely and wants to make new friends can act against his better judgement and open up about his fears and anxieties to strangers with whom he hopes to make a connection. I summarize this in six motivations:

(1) Financial gain: The motivation is monetary gain. It can take the form of cash, digital or virtual currency such as Bitcoin, discount codes and vouchers.

(2) Personal gain: This encompasses all nonmonetary services such as exclusive access to premium services.

(3) Moral gain/altruism: The user who aims to achieve an altruistic goal does not expect any form of money or service in return. Motivated by a sense of morality and virtue, they think that their disclosure contributes to the better good of society.

(4) Social compliance: This goal encompasses Cialdini's principles of persuasion. At its core, his work details how peoples' decisions are highly influenced by their surroundings and relationships within their social circles. Reciprocity, scarcity, authority, commitment/consistency, liking, and consensus (or social proof) can be exploited to

make people end up making seemingly counterintuitive decisions. For instance, if user Alice has a close friend who has always done favours for her, she might be inclined to share confidential information with him that she would otherwise not do (banking information for example). In fact, she might feel like the situation required a quid pro quo due to his previous acts of altruism. The same scenario could happen if "close generous friend" is replaced with "successful credible coworker", the difference is that this changes from depicting the principle of reciprocity in the former situation to authority in the second. Having defined the categories of personal data and the goals of disclosure, the next subsection details the process of "preference elicitation" and establishing the user profile.

(5) Self-expression: This category includes actions made with the purpose of sharing one's happiness, sadness, grievance, complaint, preferences, past experiences, etc. Self-expression has long since been associated with the use of SNS, however, it became even more prevalent with the growth of short format-based platforms. In their research paper, Claresta *et al.* [**297**] conclude that self-disclosure on TikTok is mostly triggered by a feeling of wanting to express oneself and get comfort thanks to the presence of other individuals who do the same thing on this social media. Other scholars focus on "authentic self-expression" [**298**], which refers to realistic posts unbound by self-idealization, and its correlation with well-being. Distressed individuals can benefit from honestly sharing their frustrations, anger, and fear online. The findings of Zhang *et al.*, revealed that higher levels of disclosure on SNS were observed amongst very stressed-out participants [**299**]. While the benefits and drawbacks of this phenomenon are subject to debate, there is no doubt that self-expression is one of the greatest motivations for disclosure online.

(6) Development and maintenance of interpersonal relationships: The term interpersonal relationship refers to a social association or connection with varying degrees of intimacy and duration. This ties in with the social penetration theory explained in Chapter 1. Disclosure is one of the most commonly used strategies to develop, deepen, and maintain these affiliations. It is through sharing one's deepest secrets, memories, feelings, and aspirations, that relationships become stronger and acquaintances step into the close circle of the individual. The level of disclosure depends on the type of connection. While bearing one's heart to their romantic partner or family member might come naturally to most people, doing the same with a boss or co-worker is deemed unprofessional.

The motivations are closely tied to behavioural biases as social compliance can be considered as both a perceived goal and a cognitive bias (social validation, which is Cialdini's fifth principle). Both are based on the desire to conform to societal norms. Moreover, it is

often a perceived reward that blinds users and leads them to deviate from rational decision-making. A job seeker in desperate need of money is likely to be financially motivated and also susceptible to the scarcity tactic that scammers use in their phishing attacks. Hence, after exploring the perceived gains, next come the biases.

5.2.3. Behavioural biases

Considering behavioural biases is very important to decide what approach to use with the user, for two main reasons:

(1) First: It is relevant to know which biases sway users' decisions in order to push nudges when the context promotes disclosure.
(2) Second: Behavioural biases can be a force for good when yielded to deter from disclosure.

To explain this let us consider this example: Alice is susceptible to the scarcity bias and the social validation bias, both of which have been discussed in Section 5.2.3 of Chapter 1. When a situation in which the reward seems limited both in time and in the number of people who can benefit from it, she is about to share her personal data. The first purpose of utilizing her biases is knowing that in this situation, there is a strong drive for disclosure that the ayatem needs to respond to.

The second purpose manifests itself as follows: The process of mitigating the disclosure can use the fact that Alice is susceptible to the other one, the social validation bias, to push a nudge. This bias describes the phenomenon where people tend to conform to the actions of others within a group. Thus, the following nudge might be the most effective to deter her from data disclosure: "*You are about to disclose personal information that is deemed very sensitive by over 90% of the other users. They stated that under no circumstances would they ever share this unless it is with a highly trusted party*".

The behavioural biases that this work considers have all been detailed in Chapter 1 and they are as follows:

(1) Emotional bias as described in Section 1.7.1
(2) Default bias
(3) Reciprocation/reciprocity
(4) Authority
(5) Social validation

The first bias encompasses whether the user is currently feeling a heightened emotion and how that reflects on their decision-making process. Overall, these are all contextual and circumstantial rather than fallacies. Tackling the latter generally requires changing the person's faulty reasoning and challenging their beliefs over time and through multiple iterations, which this dissertation does not focus on.

5.3. Disclosure detection

The proposed system is agnostic to the mechanism and technology leading to the disclosure detection task. This is, however, a relevant task that precedes the nudge-based intervention. The aim is to figure out the following based on the user's input:

- Topic, motivation, bias, and audience.
- Disclosure category: self-disclosure or multiparty disclosure.

A simple method like searching by keywords and structure can be used by parsing the text, searching for the following combination: "I am/I'm/my name is" + name (from a database of names), and deducing that the user is self-disclosing their name, for example. More sophisticated solutions such as *transformer-based models* can achieve state-of-the-art performances on various NLP tasks. I, **first**, explored this route in a collaborative research effort with Ramyasree Vedantham with the objective of training a domain-specific language model based on *Electra*[2] to detect Self-disclosure on social media [**300**]. The process relied on *Named Entity Recognition* (NER), which is a subtask of *information extraction* that aims to locate and classify named entities into pre-defined categories such as *proper nouns*, *locations*, *organizations*, etc. The difference in terms of accuracy between the domain-specific Electra that we worked on and the pre-trained version *Google Electra-small* was disappointingly negligible and the approach ended up increasing the processing time, which led me in a different direction.

The **second** approach aims to classify the disclosure as the first proposition did, but it uses *semantic similarity* to compare the current post with data that has been previously seen by the model. The output is the most likely topic, motivation, and bias. Semantic similarity measures the closeness in terms of the meaning, not the characters in the strings. This is achieved by comparing the *embeddings* of the two words (or sentences). An embedding is a vector comprised of an array of numbers. There are quite a few models capable of achieving this such as *Word2Vec*[3] and *GloVe*[4]. The issue here is that in human communication, written or otherwise, the meaning of a word is often dependent on the sentence. This is something that both GloVe and Word2Vec fail to capture as they associate each word with a fixed representation. If Bob shares "*Metallica is my favourite heavy metal band*" and Alice writes "*Exposure to a heavy metal such as Mercury for a long period of time can be very dangerous to humans*", it is obvious that the same group of words "*heavy metal*" is used. To those who understand the language, there is a clear distinction between the two meanings, but to a system that represents words statically, a high similarity is likely to be inaccurately reported.

[2]ELECTRA is a method for self-supervised language representation learning. It can be used to pre-train transformer networks using relatively little computation.

[3]`https://www.tensorflow.org/tutorials/text/word2vec#:~:text=word2vec%20is%20not%20a%20singular,downstream%20natural%20language%20processing%20tasks.`

[4]`https://nlp.stanford.edu/projects/glove/`

As such, the **third** and final approach led me to work with *BERT* and benefit from its ability to produce word presentations dynamically by considering the other components of the sentence. More specifically, the version used is *DistilBERT*[5], a transformer-based model that is smaller, lighter, and faster than BERT. It runs 60% faster while preserving over 95% of BERT's performances. By using it, the system is indeed able to capture the difference in the "*heavy metal*" example because it examines the surrounding words and captures their meaning, as well. A variation of Bert was chosen despite the existence of other language models capable of achieving the same task because it has proven to perform well in similar contexts related to social media posts [**301, 302, 303**].

Table 5.1 shows examples of the output of the disclosure detection process using DistilBERT. The system ran three separate classification tasks to predict the topic, motivation, and bias. I, then, grouped them together in the output shown in the table. The output "null" in the third example reflects the fact that DistilBERT did not report a significant probability of the sample belonging to any of the classes.

Tableau 5.1. Applying the disclosure detection process to samples from the dataset

Samples from the dataset	**(Topic, Motivation, Bias))**
"Spent another sleepless night tossing and turning, only to get up at 9 am and deal with a kid who, I swear, has the loudest, most ear-piercing scream I have ever heard (and uses it). To top it all off, I got a call from the clinic saying that I'm anemic, and they want to put me on yet ANOTHER iron supplement that I can't afford. FTW."	(Alcohol, drugs consumption, and health information, Self-expression, Emotional bias)
"Negative A blood urgently for a little girl at the Rainbow ward suffering from a stomach tumour. Could you please make an announcement at your respective lists to see if anyone could donate. This is a rare blood group. If you have anyone, they need to urg"	(Alcohol, drugs consumption, and health information, Altruism, Social validation)
"Watched the President's speech yesterday. Wow, I'm SOOO indoctrinated! [//sarcasm] Stupid radical jerks..."	(Religious and political views, Self-expression, None)

In addition, text-based disclosure detection can be enriched by considering the photos as well. For example, Alice might post "*I'm having the time of my life*", which, alone, does not allow the model to classify it as one disclosure topic over another. It might not even identify it as an instance of disclosure, to begin with. However, if she posted that along with a selfie while drunk at a party, it can be given the topic label: "alcohol, drug consumption,

[5]`https://huggingface.co/docs/transformers/model_doc/distilbert`

and health information". The challenging part is that such a task requires a *multimodal approach* capable of handling both texts and photos. Fortunately, there are multiple existing tools that can automatically generate a caption (text) from an image such as *Astica vision*[6]. Figure 5.4 shows an example of applying this to an image readily available online[7]. Both the text written by the user and the caption generated from the image are used to detect the disclosure.

Fig. 5.4. Results of image to text using Astica vision applied to an example[8]

So far, the process of identifying the topic, motivation, and bias has been explained. Detecting the audience is much more straightforward and does not require the use of a language model because, on most SNS, it is a preexisting feature that the user manually sets. As for deducing whether the post showcases self-disclosure or multiparty disclosure, I designed a few rules for the system:

- Use of proper nouns other than the user's name: Multiparty disclosure.
- Use of subject, object, possessive, or reflexive pronouns of the first-person singular forms "I", "me", "mine", and "myself": Self-disclosure.
- Use of subject, object, possessive, or reflexive pronouns other than the first-person singular forms such as "he", "she", "they", and "theirs": Multiparty disclosure.

[6] `https://www.astica.org/vision/describe/`
[7] `https://beachsideteen.com/wp-content/uploads/2020/04/Partying-Alcohol-Abuse-Beachside.jpg`

- Use of subject, object, possessive, or reflexive pronouns of the first-person plural forms "we", "us", "ours", and "ourselves": Both self-disclosure and multiparty disclosure.

If we revisit the examples in Table 5.1, the first sample shows an abundant use of the subject "I", which makes it an instance of self-disclosure. The second one is marked by the presence of multiple pronouns ("you", "they", and "anyone") none of which are first-person singular or plural forms. Hence, this would be classified as multiparty disclosure.

To conclude, the system is capable of detecting the topic motivation, bias, audience, and disclosure category. This process allows it to push a nudge considering the context-aware user model, which the next section delves into.

5.4. Context-aware user model

User modelling is the process of constructing a representation of a specific user whether it be their preferences, knowledge, or behaviour for the purpose of offering a customized adapted response from the system. To adopt a personalized approach to privacy assistance, there is a need to construct a user model. The two main sub-components I chose to represent the user are the *disclosure appetite* β and the *topic sensitivity* γ. Both of which can be seen in Figure 5.5.

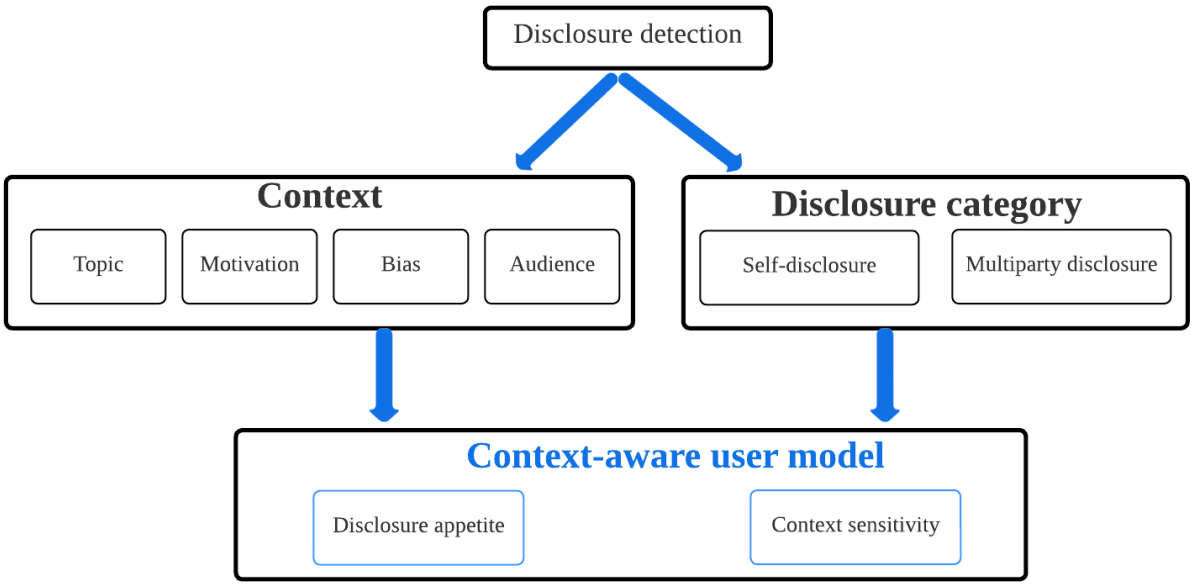

Fig. 5.5. The inputs and sub-components of the context-aware user model.

I use Dey's definition of context [**304**], which is commonly referred to in the literature: "*Context is any information that can be used to characterize the situation of an entity. An*

[8]Source: `https://beachsideteen.com/wp-content/uploads/2020/04/Partying-Alcohol-Abuse-Beachside.jpg`

entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and application themselves". Furthermore, *context awareness* refers to a system's ability to gather information about the environment and react accordingly [**305**]. *Context-aware user modelling* captures not only the user's past recorded actions to infer their preferences and tendencies but also elements of the context. It is often used to improve the quality of recommendations. For example, Bob might get a digital advertisement for discounted clothes as he is walking by a specific store. His phone reported the current time along with his location, which triggered the process. Had he not walked by the store or had he been there after closing time, he most likely would not have received this advertisement because it would have been useless.

In this dissertation, the context is defined by the elements that affect privacy decision-making and can be expressed as: $Context_{\mathrm{i,j,k,l}} = (Topic_i, Motivation_j, Bias_k, audience_l)$. Aside from the topic, motivation, and bias, all of which have been detailed in Section 5.2, the fourth component of the context is the audience. There are three options for the latter: *close friends and family*, *colleagues/classmates*, and *the general public*. The people with whom Bob shares the post can encourage or discourage him from pursuing this action. He might aim for the largest audience because he thinks that the funny political meme that he made is engaging and deserves to be seen by more people. This is particularly observed amongst individuals who deem their post to align with the opinion of the majority of potential viewers (a form of social validation) [**306**]. The same element, the audience, might have the opposite effect on Alice who is averse to disclosing such opinions to strangers in fear of negative social sanctions such as feeling criticised and ostracised. In this case, she might end up granting access to the content solely to a niche of family members and friends who understand her sense of humour and would be understanding even if she makes a mistake. There is no general rule connecting the size of the audience to the desire or willingness to disclose private data.

Figure 5.6 illustrates an example in which changing the motivation alone leads to two different contexts and ultimately two different disclosure decisions for the same person. It corresponds to the following self-disclosure scenario: Alice does not like to express vulnerability online so, she would not normally disclose her intimate stories for **self-expression** (motivation 1). Her disclosure appetite and the context sensitivity in this instance are respectively $\beta_{\mathrm{n,s}}$ and $\delta_{\mathrm{s,1,1,1,1}}$, in which the "s" stands for self-disclosure and "n" is the user id. Later on, when she sees many of her friends sharing their sad experiences and emotional struggles during COVID lockdown, she felt **compelled to follow suit** and disclosed her own. Thus, she chose to reveal private information as she seeks the **social compliance** motivation (motivation j), which changes the context sensitivity to $\delta_{\mathrm{s,1,j,1,1}}$.

The user model is comprised of two main sub-components:

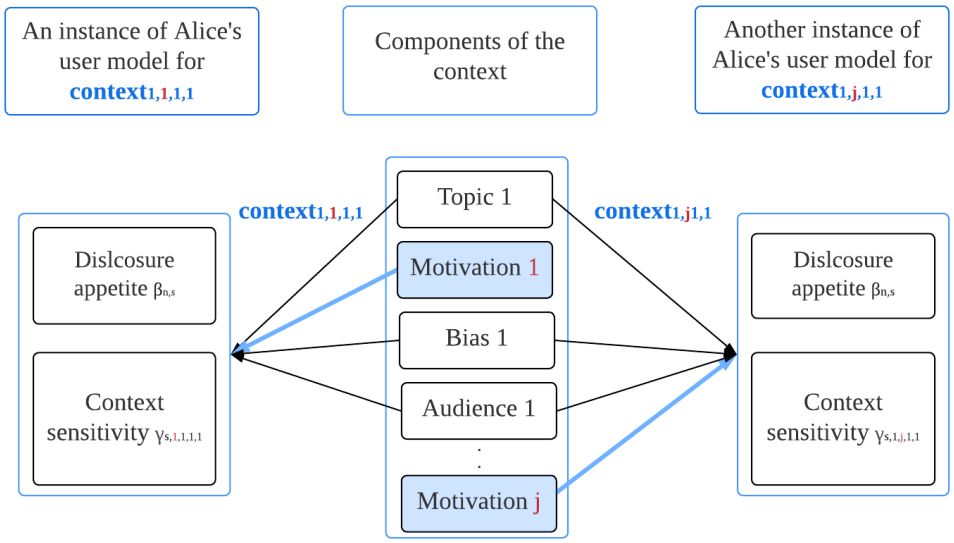

Fig. 5.6. A visual representation of the context-aware user model for self-disclosure.

(1) Disclosure appetite: In the context of *enterprise risk management*, the term *"risk appetite"* has a number of definitions, most with a link to risk acceptability, but also the values and goals that an entity seeks to attain. It is one of the core parameters in decision-making and is often used interchangeably with *risk acceptability* and *risk tolerability* [**307**]. I borrow this term and adapt it into *disclosure appetite* to denote the drive for disclosure at the expense of one's privacy. A high value signifies that the user is very inclined to share private information and a low value indicates that the individual is privacy-aware and not prone to impulsive sharing. Each user n has four disclosure appetite values: one for self-disclosure $\beta_{\mathrm{n,s}}$ and three for multiparty disclosure $\beta_{\mathrm{n,m1}}$ (social circle 1 = family and close friends), $\beta_{\mathrm{n,m2}}$ (social circle 2 = colleagues and classmates), and $\beta_{\mathrm{1,m3}}$ (social circle 3 = the public).

(2) Context sensitivity: Each $Context_{\mathrm{i,j,k,l}}$ is characterised by four sensitivity values: one for self-disclosure $\delta_{\mathrm{s,i,i,k,l}}$ and three for multiparty disclosure $\delta_{\mathrm{m1,i,j,k,l}}$, $\delta_{\mathrm{m2,i,j,k,l}}$, and $\delta_{\mathrm{m3,i,j,k,l}}$ where the letter "m" stands for multiparty disclosure followed by a number from 1 to 3 to indicate the social circle involved in the content: "1" for family and close friends, "2" for colleagues and classmates, and "3" for the public.

This duality of user and context-specific values bears a striking resemblance to the *Rasch model*, which seeks to represent and measure the connection between a test taker's ability (person parameter) and the difficulty of questions or items (item parameter) based on the responses of the former to the latter. Furthermore, Knijnenburg's work [**308**], which tackled self-disclosure using the Rasch model, reported promising results. As such, I was motivated to develop a more comprehensive version including both context awareness and consideration for

multiparty disclosure, both of which have not been considered up until now. My proposition is illustrated in Figures 5.7 and 5.8.

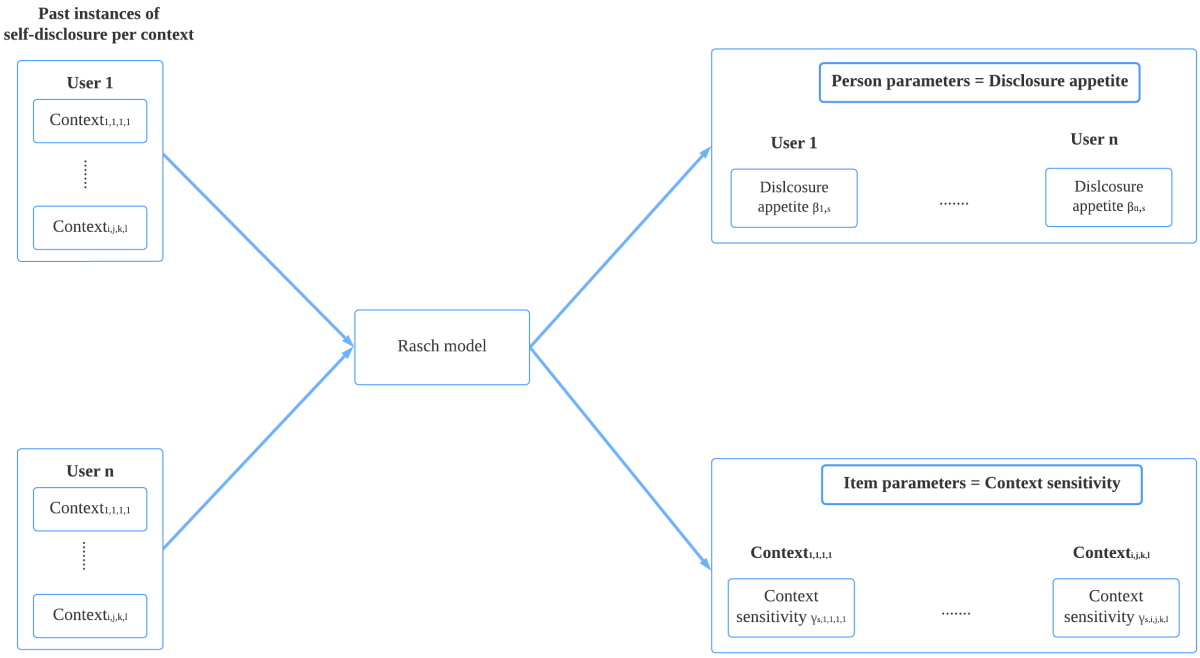

Fig. 5.7. Applying the Rasch model to self-disclosure.

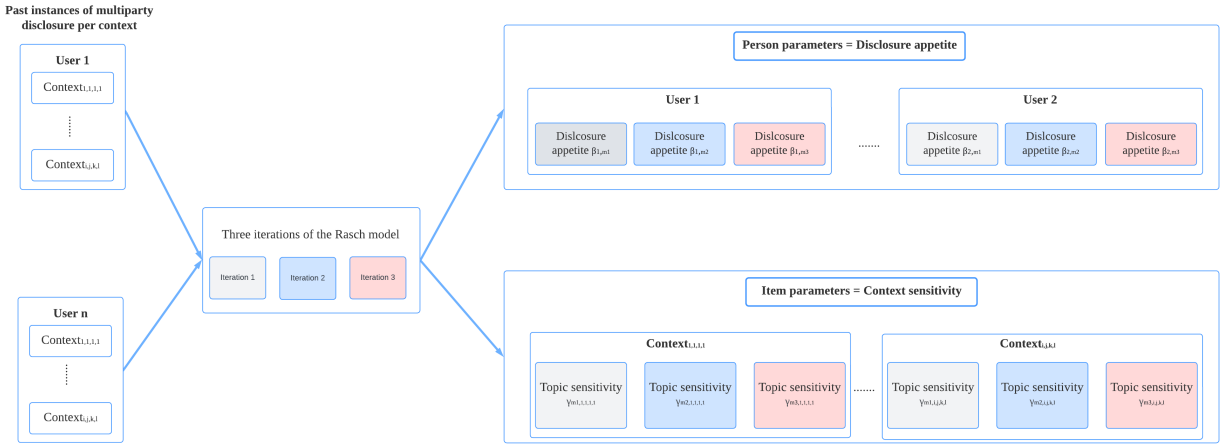

Fig. 5.8. Applying the Rasch model to multiparty disclosure.

One of the advantages of the Rasch model lies in its formulation, which allows the system to predict future responses based on the aforementioned parameters as shown in Equation 5.3. I adapt this in Equation 5.4 to align with my objective.

The original Rasch model:

The objective is to calculate the probability of a person n answering an item i correctly,

based on that person's ability level β_{n} and the difficulty of the item δ_{i}. $x_{\mathrm{n,i}} \in \{0,1\}$ is a dichotomous random variable where, $x_{\mathrm{n,i}} = 1$ denotes a correct answer given by user n to the given assessment item i. $x_{\mathrm{n,i}} = 0$ indicates its incorrect counterpart. The probability of the outcome $x_{\mathrm{n,i}} = 1$ is given by:

$$P\{x_{\mathrm{n,i}} = 1\} = \frac{e^{\beta_{\mathrm{n}} - \delta_{\mathrm{i}}}}{1 + e^{\beta_{\mathrm{n}} - \delta_{\mathrm{i}}}} \tag{5.3}$$

where:
β_{n} is the ability of person n
δ_{i} is the difficulty of item i.

My adapted Rasch model for self-disclosure:
I substitute the ability of the person with the self-disclosure appetite and the difficulty of the item with the context sensitivity. $P\{x_{\mathrm{n,s,i,j,k,l}} = 1\}$ is the probability of user n self-disclosing information about $Topic_{\mathrm{i}}$ in $Context_{\mathrm{i,j,k,l}}$.

$$P\{x_{\mathrm{n,s,i,j,k,l}} = 1\} = \frac{e^{\beta_{\mathrm{n,s}} - \delta_{\mathrm{s,i,j,k,l}}}}{1 + e^{\beta_{\mathrm{n,s}} - \delta_{\mathrm{s,i,j,k,l}}}} \tag{5.4}$$

where:
$\beta_{\mathrm{n,s}}$ is the disclosure appetite of user n
$\delta_{\mathrm{s,i,j,k,l}}$ is the sensitivity of $Context_{\mathrm{i,j,k,l}}$ in the case of self-disclosure.

This also applies to multiparty disclosure cases by replacing $x_{\mathrm{n,s,i,j,k,l}}$ with $x_{\mathrm{n,m1,i,j,k,l}}$ (or m2 or m3). The parameters $\beta_{\mathrm{n,s}}$ and $\delta_{\mathrm{s,i,j,k,l}}$ are also changed to $\beta_{\mathrm{n,m1}}$ and $\delta_{\mathrm{m1,i,j,k,l}}$. The higher the disclosure appetite, the more likely an individual is to share data. Decreasing the context sensitivity has the same effect of increasing the probability of the user pursuing the disclosure. Let us go consider this example:

When Alice uses the system for the first time and expresses her explicit consent, she grants it access to her past posts, comments, and tagged content[9]. These are instances of past disclosure and are used as part of the dataset. It is further enriched by her answers to a questionnaire[10] presenting her with context-specific disclosure scenarios. The combination of the two allows the system to overcome the cold start problem and construct her user model. All of the sections in the questionnaire are in the affirmative form such as "In this situation, I would share" or "In this context, I feel comfortable sharing...". Alice has the option to answer by checking the option "agree" or "disagree".

[9]When another user tags her in a photo or comment for examples. This also includes instances of multiparty disclosure initiated by her in which she tagged others.
[10]Examples of the questions are included in Chapter 6.

Later on, Alice finds herself in the following disclosure situation: she takes a photo in which her **friend** Bob is included while they are both **partying and drinking**. As she is about to share it publicly on Facebook, she gets a popup asking her if she wants to tag Bob. She is one click away from doing so. This marks the presence of the **default bias**, which can lead Alice to tag Bob even though she might not have otherwise done it. For years, Facebook was using an **auto-tag feature** that allows its algorithm to detect the presence of a face in the photo and also identify who it belongs to, which is Bob, in this case [**309**]. If he has previously been tagged in other photos, the Facebook algorithm uses that data to recognize him in the future and prompt his friends to tag him. Table 5.2 illustrates the scenario.

Tableau 5.2. Context parameters for Alice's multiparty disclosure example

($Topic_i$, $Motivation_j$,$Bias_k$,$Audience_l$)	$\beta_{\mathrm{n,m1}}$	$\delta_{\mathrm{m1,i,j,k,l}}$
(Alice, alcohol, drug consumption, and health information, Self-expression, Default bias, Public)	3.63	2.48

This allows the system to calculate the probability of Alice disclosing this post as follows:

$$P\{x_{\mathrm{n,m1,i,j,k,l}} = 1\} = \frac{e^{\beta_{\mathrm{n,m1}} - \delta_{\mathrm{m1,i,j,k,l}}}}{1 + e^{\beta_{\mathrm{n,m1}} - \delta_{\mathrm{m1,i,j,k,l}}}} = \frac{e^{3.63-2.48}}{1 + e^{3.63-2.48}} = 0.76$$

There is a high probability of Alice sharing the post, which she was already attempting to do. The relevance of this calculation becomes apparent when the system attempts to change the context to decrease this probability, which will be discussed in Section 5.5.4 as part of the mediation process.

5.5. Multi-agent assistant

The multi-agent assistant is the core of the disclosure-mitigating system. It is the nudge-generating main component that takes in the output of the disclosure detection module, the user model, and the domain knowledge to produce personalized context-aware nudges. Figure 5.9 shows the different agents interacting with one another.

If a user is disclosing information solely about themselves (self-disclosure), only the personal agent and Aegis are needed to push a nudge. In the other case, when someone else's information is disclosed, then, there is a need for multiparty agents (one agent per party) to represent their best interest.

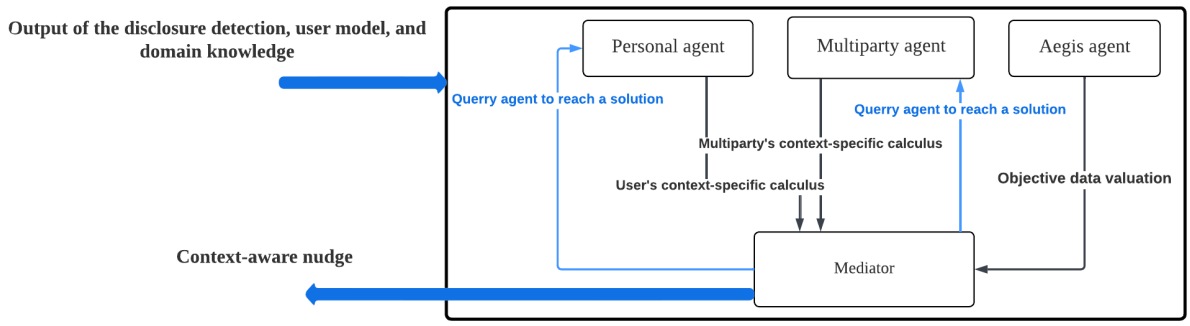

Fig. 5.9. The components of the multi-agent assistant

5.5.1. Personal agent

The personal agent represents the user's perception of privacy and the current context they are in. It uses the disclosure appetite and context sensitivity. The foundation of this agent is inspired by the privacy calculus and the privacy paradox, both of which are detailed successively in Sections 1.5 and 1.6. The former states that users subtract the perceived risk from the benefit to make a privacy decision. The latter proposes that this is not enough as people are subjected to biases leading to discrepancies between their judgment and their actions. It does not, however, point to a concrete way to represent and calculate this. I propose Equations 5.5 and 5.6 as the personal agent's *context-aware privacy calculus* formulas respectively for self-disclosure and multiparty disclosure. The agent is representing user n.

$$Calculus_{\mathrm{n,s,i,j,k,l}} = \beta_{\mathrm{n,s}} - \gamma_{\mathrm{s,i,j,k,l}} \tag{5.5}$$

$$Calculus_{\mathrm{n,m1,i,j,k,l}} = \beta_{\mathrm{n,m1}} - \gamma_{\mathrm{m1,i,j,k,l}} \tag{5.6}$$

where:

$\boldsymbol{Calculus_{\mathrm{n,s,i,j,k,l}}}$: is the context-aware measure of the privacy calculus of user n while subjected to $Context_{\mathrm{i,j,k,l}}$ in the case of self-disclosure.

$\boldsymbol{Calculus_{\mathrm{n,m1,i,j,k,l}}}$: is the context-aware measure of the privacy calculus of user n while subjected to $Context_{\mathrm{i,j,k,l}}$ in the case of multiparty disclosure involving the first social circle (family and close friends). The variable "m1" can be replaced by "m2" or "m3".

The equation is designed to find a middle ground between the privacy calculus and the privacy paradox. Some scholars have debated that one negates the other (Sections 1.5 and 1.6) based on two different perspectives:

- The privacy calculus can be rather monolithic by not considering the factors that can impact the perception of the same individual under different circumstances. As such,

the privacy paradox, which acknowledges the biases and other contextual factors, cannot align with the former theory.

- On the other hand, proponents of the privacy calculus believe that the notion of the privacy paradox can amplify edge cases[11]. Furthermore, the privacy calculus is one of the longest-standing theories to investigate decision-making and human behaviour. Hence, deconstructing it without offering an alternative can be problematic.

To ease the contention around the second point, Equations 5.5 and 5.6 keep the structure of the privacy calculus. Addressing the first point comes in the form of including context awareness, which considers various factors amongst which are the biases. The disclosure appetite and context sensitivity are different from the traditional definition of "perceived gain" and "perceived risk". Furthermore, the proposed context-aware version retains the same meaning as the original privacy calculus: if the calculation is positive, the agent considers that the user it is representing would want to proceed with the disclosure action.

Aside from the personal agent, there are more considerations to take into account because there are two other agents that interact with it starting with the multiparty agent.

5.5.2. Multiparty agent

The multiparty agent assumes the perspective of the other parties involved in the disclosure. I published a journal paper with a focus on this component in 2023 [**310**]. In the scenario where Alice tags Bob, such an agent is assigned to represent him in the system. If she tags Sam as well, then, there are two multiparty agents. This agent uses the same calculus function defined in Equation 5.5. However, the calculation is not as straightforward. Multiparty members are not always known to the system. In the **best-case** scenario, Bob has previously used it and as a result, the dataset contains his past posts and his answers to the initial questionnaire. Hence, an agent can be assigned to him without any issues. In the **most likely case** scenario, the only information known about him is his past interactions with Alice. This is still a trove of data as it is likely that he commented on her posts, tagged her, and was tagged by her, etc. In the **worst-case** scenario, this is the first time that Alice and Bob have ever interacted on their social media accounts, which does not provide the system any foundation to directly build his user model.

The best case and most likely case scenarios are both handled in the same way by constructing the user model based on the available data. The worst case, however, calls for a different approach. The system must rely solely on the publicly available information on Bob's profile such as age, gender, and current occupation to infer his disclosure appetite $\beta_{\text{bob,s}}$. To achieve this, I use a solution inspired by *user-based collaborative filtering* for

[11]An edge case is a problem or situation that occurs only at an extreme (maximum or minimum) operating parameter. In this context, it refers to rare cases that do not apply to the larger population.

recommender systems. If the system is unfamiliar with user Bob's preferences, it calculates the similarity between him and the existing users. The assumption here is that the items of interest to users who are in Bob's proximity (similarity-wise) are likely to interest Bob himself. In the same spirit as this, I use the available information on the newly introduced multiparty member to match them with existing users whose disclosure appetites are known.

One issue arises due to the presence of some categorical data such as gender and current occupation. If Alice is an engineer, Sam is an author, and Bob is a financial analyst, a human being might interpret Alice and Bob to have more similar jobs since they fall under the umbrella of "*Science, technology, engineering, and mathematics* (STEM)". However, if the system is simply given the occupations as labels ("engineer", "author", and "financial analyst"), the connection is not instantly made. It might assign each of them a numerical value: 0 for "engineer", 1 for "author", and 2 for "financial analyst". But in this case, "engineer" and "author" are closer than "engineer" and "financial analyst".
As a result, this approach is not ideal as there is a need to calculate the closeness in terms of meaning between the different labels, which is called *similarity encoding* [**311**]. In a similar process to the one explained in Section 5.3, an embedding is created to represent each string (label) and as a result, similar meanings get similar encodings. Let us consider two users Alice (A) [A1,A2,[A3,A4,A5]] and Bob (B) [B1,B2,[B3,B4,B5]]. The variables A1, A2, B1, and B2 are numerical such as age. On the other hand, [A3,A4,A5] and [B3,B4,B5] represent the embedding of a single categorical variable. So, if the category is marital status, [A3,A4,A5] can represent "married" and [B3,B4,B5] can denote "complicated". With this, the similarity between the two can be calculated as the following *Euclidean distance*:

$$d(A,B) = \sqrt{(A1 - B1)^2 + (A2 - B2)^2 + (A3 - B3)^2 + (A4 - B4)^2 + (A5 - B5)^2} \quad (5.7)$$

The number of similar users is set to three. In other words, the system averages the user model values corresponding to the three closest users to Bob. There is no preexisting method to calculate the optimal number of similar users. It is initialized as three, however, in the case of poor results, this can be changed.

The personal and multiparty agents both represent human perspectives, which can be flawed even without the presence of contextual biases. The users can lack the knowledge to be privacy-aware. As such, there is a need for another actor, which has a firmer grasp on privacy.

5.5.3. Aegis agent

In Greek mythology, Aegis[12] was a shield carried by Zeus for additional protection. I borrow the term and use it for the agent whose objective is to provide a privacy-preserving perspective to the user. Aegis was first introduced in my article, which was published in the proceedings of the conference on *AI, Ethics, and Society Conference* (AIES) [**312**]. It is the answer to the question "*what if the user lacks privacy awareness in general and not as a result of contextual factors overriding their judgment?*".

The role of Aegis is to judge whether an intervention is needed or not using the result of the disclosure detection module. This process is not based on the user's preferences and is instead designed to produce an objective voice of reason. In the literature, this is referred to as *data valuation*, which is a discipline in the fields of accounting and information economics and is focalized on the value of data once it no longer solely belongs to the data owner (as defined in Chapter 3). Each of the five topics detailed in Section 5.2.1 has a value that has been determined based on two sources: crowd data valuation and market valuation.

5.5.3.1. Crowd data valuation. This approach is based on the findings of multiple studies, amongst them is my own which I did as part of the evaluation in the article entitled "The Privacy versus Disclosure Appetite Dilemma: Mitigation by Recommendation" [**313**]. Numerous members of the crowd are recruited and their opinions are averaged to get a final output. They are laymen non-expert members of the general public, which can make one wonder how this can be objective if the people themselves are susceptible to biases and have motivations that can sway their judgment. While that is certainly true on an individual level, the power of the crowd has been considered as good as or even better than an expert's opinion in multiple fields and based on numerous experiments [**314, 315, 316**]. In the absence of a unanimously agreed upon expert opinion on data valuation, crowd wisdom might be the closest thing we have to an "objective" estimation on the subject.

The research by Prince *et al.* [**317**] aims to understand how much privacy is worth around the world with a focus on the United States, Mexico, Brazil, Colombia, Argentina, and Germany. The authors classify data into six data categories, namely: *finances*, *biometrics*, *location*, *networks*, *communications*, and *web browsing*. The paper reports on the average payment consumers would demand for permission to share Data, which is 8.44$ a month for a bank balance, 7.56$ a month for fingerprint information, 6.05$ to read an individual's texts, etc. When it comes to establishing differences based on geographic origin, the

[12]The Aegis is a shield or breastplate used by Zeus and his daughter Athena in Greek Mythology. Doing something "under someone's aegis" means acting under the protection of a powerful and knowledgeable source.

German participants displayed a much greater concern for their private financial information. According to their self-reported assessment, they need to be paid double the average across the entire study to disclose such data. In a similar vein to studying geographic origin-based differences, Schomakers *et al.* [**318**] propose their ranking of 40 pieces of data and compare the results with German, Brazilian, and US American individuals. Using cluster analysis, the authors group the data into three categories high, medium, and low sensitivity. Although there were minor differences between participants based on geographic distinction, it is concluded that there is a consensus on what constitutes sensitivity across nations. Table 5.3 shows a summary of data valuation based on existing literature [**317, 318, 319, 320**], and my own research in which individuals are surveyed and self-report their estimation of the value of personal data [**313**]. Due to the fact that the papers use different scales, I normalize all of them to [0,1]. Based on a value x, which was reported on a scale of $[x_{\min}, x_{\max}]$, its normalized value x' on a scale of [0,1] can be obtained as follows:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{5.8}$$

So, if a sensitivity value is $x = 9$ on a scale of [1,10], $x' = 0.89$ when normalized to [0,1].

Tableau 5.3. Normalized crowd-based data valuation on a scale of [0,1]

Disclosure topics	**Normalized value**
Identity	0.22
Location	0.77
Experiences	0.25
Alcohol, drugs consumption, and health information	0.85
Religious and political views	0.88

As discussed in Chapter 3, since personal data exists within the larger context of the economics of privacy, data valuation cannot be solely dependent on the data owners (users whose data is exchanged and sold) and must include the market valuation which encompasses the interactions between the other two parties: the data brokers and the data users.

5.5.3.2. Market valuation. The market valuation revolves around the economic value and the benefit that brokers and data users generate from collecting and mining data from users. Data brokers collect, analyze, mine, and sell data to companies that might use it for targeted advertising, for example. This is a highly lucrative $200 billion industry. According to a report by Avast[13], a bundle of user data can be worth more than $240 per year while an email address alone can retail for about $89. The company *Invisibly*[14] lets users collect points that

[13] `https://www.avast.com/c-data-brokers`
[14] `https://www.invisibly.com/`

can be exchanged for cash value by giving them access to their web usage. The company can monitor, in real-time, the user's likes, dislikes, and other activities such as saving articles and news etc. The average payment is around 5$ to 10$ according to a spokesperson from the company[15]. On the higher echelon, anonymized health records are valued at 1000$, which is calculated based on the report of *Flatiron Health* [**227**]. It is a Google-backed health-tech company that collects patient data, analyzes medicine performance in the real world, and sells its findings to pharmaceutical companies. The company's website indicates that it has 2.2 million active patient records available for research. So, when it sold for $1.9 billion, some health-tech experts noted that the sale amounted to roughly $1,000 per record.

I averaged the crowd and market valuation to obtain the final data valuation as shown in 5.4. As these values are between 0 and 1, a simple way to classify data into "least sensitivity", "less sensitivity", "high sensitivity", and "highest sensitivity" is by dividing this interval into equal subdivisions: [0, 0.25],]0.25, 0.5],]0.5, 0.75],]0.75, 1] from least to highest sensitivity.

Tableau 5.4. Aegis's normalized data valuation of the disclosure topics on a scale of [0,1]

Disclosure topics	**Normalized value**
Identity	0.26
Location	0.72
Experiences	0.19
Alcohol, drugs consumption, and health information	0.78
Religious and political views	0.77

The final nudge presented to the user who is about to share the content is the outcome of the negotiation between Aegis, the personal agent, and the multiparty agent. The mediator stands in between to ensure the best outcome by considering all of their perspectives.

5.5.4. Mediator agent

The mediator aims to produce a nudge that is accepted by the sharer while also considering the privacy-preserving and the multiparty perspectives. There is no universal algorithm for mediation so, my final approach to this task was achieved after three iterations:

5.5.4.1. First iteration: Risk tolerance-based approach. This approach to mediation is published in the *workshop on Online Misinformation- and Harm-Aware Recommender Systems* (OHARS 2020) [**263**]. The concept of disclosure appetite was yet to be established and instead, the solution was based on the term "*privacy threshold*" or "*privacy tolerance*"

[15] `https://www.fox26houston.com/news/consumers-can-get-paid-for-theirinternet-data.`

both of which denote how much private information an individual is willing to share. The mediation between the drive for disclosure and the potential privacy risk is achieved through comparing the risk with the individual's tolerance and then, pushing nudges if the former surpasses the latter. The mediator considers two valuations of personal data x_{i}: subjective user-specific $W_{\mathrm{i,subj}}$ and objective $W_{\mathrm{i,obj}}$. If $Risk < Threshold$, the disclosure is within the user's tolerance, and the process is terminated. In the other case, the system parses the disclosed pieces of information and selects the candidates x_{i} whose elimination reduces the objective risk below the personal threshold. X is the list of x_{i} that verify the Inequation 5.9.

$$X = \{x_{\mathrm{i}} \mid Risk - W_{\mathrm{i,obj}} < Threshold\} \tag{5.9}$$

The system uses a *greedy approach*[16] and looks for the highest $W_{\mathrm{i,obj}}(x_{\mathrm{i}})$ first. If it is enough to reduce the risk below the threshold, then, the process ends and the user is nudged to not disclose the data in X. If not, the goal is to find two or more pieces of data which, when combined together, satisfy Inequation 5.9. Even if everyone involved does not possess sufficient privacy knowledge, the system can push nudges using objective values. In theory, is a great approach to privacy preservation. However, not considering the user's preferences results in a lower acceptance rate. This is corroborated by the literature [**321, 322**] as well as the evaluation results (Chapter 6).

5.5.4.2. Second iteration: Logrolling approach. To overcome the lack of representation of the user's preferences, *logrolling*[17] was the second iteration of the mediator. This is the act of trading across issues in a negotiation, and it requires that every party involved is knowledgeable about their preferences as well as their opponents'. In a bilateral multi-issue negotiation, if party one gains ownership of issue 1, then, said issue is lost to party 2. Logrolling is designed to overcome potential impasses leading to a breakdown by ensuring that each party involved gains issues that they are interested in. For example, let us consider the following scenario: Someone knows that Alice and Bob have a sweet tooth and offers them a brownie and a cheesecake as long as they can come to an amicable decision on who gets what or else neither of them gets anything. Each of them knows their own preferred dessert and that of the other person involved:
Alice's preferences: (brownie: 0.7, cheesecake: 0.6)
Bob's preferences: (brownie: 0.6, cheesecake: 0.8).
If Alice gets the brownie, she gains a value of 0.7 (her preference for a brownie) and Bob loses 0.6 (his preference for a brownie). Following this, Bob gets the cheesecake and gains 0.8 (his preference for cheesecake) and Alice loses 0.6 (her preference for cheesecake). So, to conclude:

[16] A greedy approach follows the problem-solving heuristic of making the locally optimal choice at each stage.
[17] Logrolling is the trading of favors, or quid pro quo, such as vote trading by legislative members to obtain passage of actions of interest to each legislative member

Alice gains: value of brownie – value of cheesecake =0.7-0.6 = 0.1 > 0
Bob gains: value of cheesecake – value of brownie = 0.8-0.6 = 0.2 > 0
Hence, the negotiation ends successfully with each party getting more than they initially had, which was nil. This is an analogy for the negotiation between the personal agent and Aegis. I consider each disclosed piece of data to be a negotiation issue. If the personal agent representing Alice gets $issue_{\mathrm{i}}$, it means that the nudge will not advise her to eliminate it and that she is free to share it. Once this agent gains possession of the issue, it is lost to Aegis, which denotes the loss of privacy caused by the disclosure. If the latter agent wins $issue_{\mathrm{j}}$ that means that the user is nudged to not share this issue and subsequently, reduce the disclosure. $V_{\mathrm{p,i}}$ is the value of $issue_{\mathrm{i}}$ to the personal agent. In this approach, to mediation, this is the disclosure appetite of the user. $V_{\mathrm{a,i}}$ is the value of the same issue to Aegis, which can be seen as the objective data sensitivity. The system looks for issues i and j that satisfy the following Inequalities and attributes $issue_{\mathrm{i}}$ to the personal agent and $issue_{\mathrm{j}}$ to Aegis:

$$V_{\mathrm{p,i}} - V_{\mathrm{p,j}} \geq 0 \tag{5.10}$$

$$V_{\mathrm{a,j}} - V_{\mathrm{a,i}} \geq 0 \tag{5.11}$$

While this logrolling approach does indeed seek to balance the disclosure appetite (personal agent) with privacy preservation (Aegis), it does not fully embody the complexities of privacy decision-making. It reduces the user's preferences to a static value of their disclosure appetite, which is not the full picture. Hence, there was a need to steer the process in a different direction, which considers the disclosure appetite, the perceived data sensitivity, and the objective data sensitivity.

5.5.4.3. Final iteration: Elimination or mitigation. The final approach is based on this reasoning: If the user cannot be convinced to eliminate the post, whether it contains self-disclosure or multiparty disclosure, then, mitigation is the only solution. In order to simplify this process, Figure 5.10 shows the context-aware privacy calculus values.

In self-disclosure scenarios, the interactions between the personal agent, Aegis, and the mediator is as follows:

(1) Aegis communicates to the mediator that the current disclosure requires an intervention.
(2) The mediator asks the personal agent whether eliminating the content entirely is possible or not given the current context: $Context_{\mathrm{i,j,k,l}}$.
(3) The personal agent considers all the contexts in which the topic and audience are the same but the motivation and bias are different (different j and k values). It calculates the $Calculus_{\mathrm{n,s,i,j,k,l}}$ for each one of them looking for the context associated with a minimum calculus value.

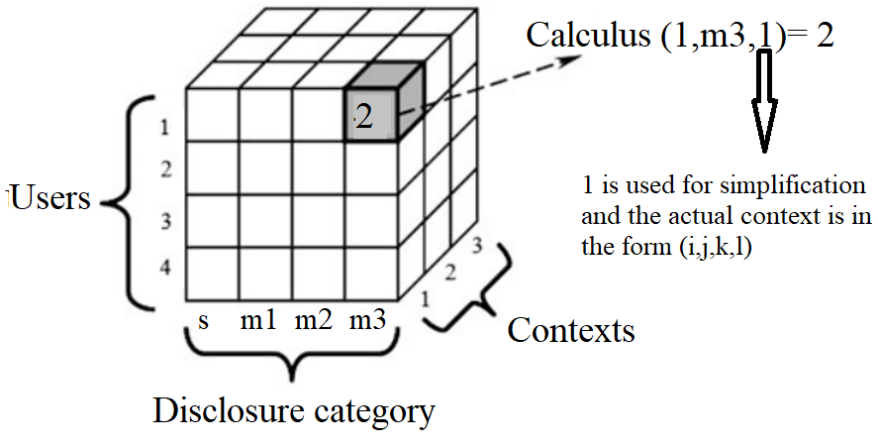

Fig. 5.10. Visualization of the context-aware privacy calculus values depending on the user, disclosure category, and context.

(4) If minimum $Calculus_{n,s,i,j,k,l}$ is negative, this means that if the system manages to override the current motivation and bias, the user would eliminate the content. This takes us back to the definition of calculus and the implication of a negative calculation on the desire to disclose. An example will follow to explain how the new context can override the original one.

(5) If the minimum $Calculus_{n,s,i,j,k,l}$ is positive, total elimination is unlikely since under all circumstances, the user perceives the disclosure to be beneficial. This is communicated to the mediator, which then, switched gears towards mitigation.

(6) The idea behind mitigation is to limit the audience. If the content cannot be eliminated, then, ensuring that it is seen by the closest people is the next best thing to do. Of course, if the user is already sharing with the closest circle, then, the process stops here. If not, the mediator prompts the personal agent to fix the variable "l" to 1 (audience: family and close friends) and look for the variables j and k that minimize $Calculus_{n,s,i,j,k,l}$.

To explain how a context can override another, we circle back to the example of Alice who is about to share a photo showing her drunk at a party. The current context can be expressed as: $Context_{i,j,k,l}$= (alcohol, drug consumption, and health information**,** Self-expression**,** Social validation**,** Public). Aegis considers this sensitive and triggers the mediation process. If the personal agent finds a new context $Context_{i,j',k',l}$= (alcohol, drug consumption, and health information**,** Financial gain**,** Authority**,** Public) that renders the calculus value negative, then, the mediator receives this information along with the indication that the complete elimination of the post is possible. So, how does the user go from the first context to the second? This is achieved through framed nudges: Presenting the user the same options (disclose or not) can lead to widely different outcomes depending

on how the intervention is structured and worded. In this case, it depends on the motivations and biases that can sway the user. Thus, considering $Context_{i,j',k',l}$= (alcohol, drug consumption, and health information**,** Financial gain**,** Authority**,** Public), a framed nudge can be as follows:
According to many privacy and cybersecurity scholars (authority), this post contains private data and by sharing it, you may **incur financial damage to the point of risking your source of income** (financial gain).

The framing effect occurs when people react differently to something depending on whether it is presented as positive, negative, or even neutral. For example, people tend to buy discounted products that clearly display the original price because they think they are getting a great deal. If the store simply shows the new price, they are less likely to make the purchase despite the current price being the same either way. There is an emerging interest in using the framing effect for the purpose of privacy preservation [**28**], which motivated me to explore it in this dissertation.

As for multiparty disclosure, it is not a dissimilar process to this with the addition of other agents (multiparty agents), at least one of them. Algorithm 5.5.1 presents the corresponding mediation process in pseudo-code.
Algorithme 5.5.1. Pseudo code of the mediation process.

For the same user and topic :
Parse all the contexts and return (min, context-min)
 # This corresponds to the new minimum privacy calculus min and its context
If (min $<$ 0), then :
Nudge using context-min
Else :
Add to elimination-set the users whose calculus is negative
 # These are the users who should be eliminated from the content (untag, blur, etc.)
min-audience := intersection between all the closest circles
Nudge user to eliminate from the post users in elimination-set
Nudge user to reduce audience to min-audience

Let us circle back to the example of Alice who wanted to share a photo in which Bob is tagged.
$Context_{i,j,k,l}$= (alcohol, drug consumption, and health information**,** Self-expression**,** Default bias**,** Public)

If there is a context such as (motivation= Development and maintenance of interpersonal relationships, bias= social validation, audience= public) which reduces her calculus to negative, then, the nudge that is likely to eliminate the disclosure is as follows:
"*You are advised to not share this post because according to Bob does not usually share photos showing alcohol consumption.* ***He would be upset*** *if you share this without his consent. Did you know that* ***90% of users who share content like this end up regretting it later*** *especially when it is shared with the* ***public****?*"
The nudge highlights how Bob would not appreciate the action, which plays on the motivation. Also, since Alice is susceptible to social validation bias, the system inserts statistics about the rest of the group, which is likely to dissuade her since she tends to follow the group's judgment.

But what happens when elimination is not possible (no negative values are found)? In this case, the other two agents introduced in Subsections 5.5.2 and 5.5.3 intervene. If there are k parties, other than the user, involved in the disclosure, the system compares their perceived data sensitivity with Aegis' objective value. If the former is above the latter, then the system groups these individuals in what is called "partial elimination group". These people are hyper-aware of their privacy and will not be satisfied with any solution other than elimination. As such the nudge could suggest cropping them out of the photo or untagging them from the text-based disclosure without deleting the whole content. For the other users, the system proceeds with the mitigation by suggesting to reduce the audience to one that they all prefer. If there are three multiparty members (Sam, Alex, and Riley) remaining after the partial elimination of Bob and in case each of them would rather only share with their closest circle (close friends and family), then, the best solution for the multiparty members is to reduce the audience to the intersection of the three. The system considers this along with the context that reduces user Alice's privacy calculus the most and pushes the nudge accordingly. The nudge would be:

"*You are advised to not include Bob in this post by* ***cropping him out*** *of the photo because he does not usually share photos showing alcohol consumption. You can still share this with the friends that you have in common with Sam, Alex, and Riley but not with a wider audience because* ***90% of users who share content like this end up regretting it later*** *especially when it is shared with the* ***public***".

In summary, this process prioritizes elimination if possible because it is the best approach to privacy preservation. However, knowing that people are driven by motivations, biases, and the audience that they want to share with, this cannot be the only approach for two reasons:

- In the short term: If the user rejects it because they are seeking a particular motivation for example, then, they would end up sharing the post as it is. As a result, the system would be rendered useless.
- In the long term: If the user keeps getting stiff strict nudges that solely push for elimination, then, they might simply deactivate the system.

5.6. Discusssion

The discussion points are broken down into two categories: first, the approach itself, and second, the ethics of the proposition. *First*, there is a newly emerging approach to privacy protection that has shifted away from safeguarding personal information to focusing on the purpose for which information is collected and processed, to begin with. Solove argues that data is what data does and that "*to be effective, privacy law must focus on use, harm, and risk rather than on the nature of personal data*" [**323**]. This means that lawmakers should not solely rely on the metric of data sensitivity to protect privacy. However, this does not fully translate into privacy protection on a personal level. For example, if Alice shares unprofessional photos of herself and ends up getting harshly criticized for it by her boss, this is not covered by the aforementioned approach. Her boss did not go out of his way to collect data about her and his intention was not to target her, but as often happens on SNS, her post was deemed of interest to him by the recommender algorithm and he ended up seeing it. While focusing on what data does on a systematic level has the potential of regulating the economics of privacy markets, in my opinion, it needs to be combined with a solution like the one proposed in this dissertation to cover the realm of SNS.

Second, the ethics of fighting fire with fire have always been dubious. As such, one might call into question my use of biases to combat biases. This can appear to have some of the hallmarks of classic paternalism, which relies on an all-knowing authority forcing the individual to adopt a specific behaviour. However, my approach did not stray away from libertarian paternalism because users can still reject the nudge without any repercussions. Furthermore, the system is not robbing them of their free will as it draws inspiration from the concept of *informed consent* by explaining to them how the system operates, what kind of data it collects, etc. This is done before signing up for the platform. Still, it should be acknowledged that this research, similar to all nudge-based approaches, does indeed take advantage of judgemental heuristics and cognitive biases [**8**]. However, the question that I posed and discussed in Section 2.7 of Chapter 1 remains valid: "*Are choices ever completely free of external factors, incentives, and limitations*?". In the absence of "Absolute freedom" and the omnipresence of behavioural biases, my proposed nudges offer a novel non-restrictive solution to disclosure online, granted that the user gives their consent.

5.7. Conclusion

Today's social media users are facing bigger threats than ever. The most concerning fact is that although the harms are becoming more prevalent, at the same time, online users, who are potential victims, are disclosing more private information than ever before. Hence, it is important to highlight how harmful the oversharing culture is and how correlated it is with a lack of digital literacy [**324**] accompanied by behavioural biases. Moreover, the gap between what is perceived to be the correct conduct and the actual actions taken by users is extremely disconcerting and calls for interventions and forms of guidance to help people navigate through these issues.

This chapter addresses the issue of disclosure and proposes a system to help people behave more responsibly and achieve two conflicting goals: personal satisfaction from the disclosure of information while protecting their privacy. Specifically, this work *first* proposes the notion of disclosure appetite to determine and model users' context-specific privacy attitudes. *Second*, it designs a personalized nudge-based system, which provides context-specific reinforcements that balance the users' need to disclose with the protection of their privacy. The model developed for the system not only considers the privacy of the user but also the privacy of others who might be affected by the disclosure. In addition to the perspective of the personal agent (which takes into account the viewpoint of the user given the context), another agent, Aegis, weighs on the disclosure mitigation with the initially declared concerns that the user expressed before being subjected to bias. Effectively, Aegis represents the unbiased perspective. The remaining objective is to evaluate the proposed system through a proof of concept.

Chapter 6

Implementation and evaluation

This chapter articulates the details of the implementation and the evaluation of *Cognicy*. It is a portmanteau term that blends "cognizance" referring to knowledge and awareness with "privacy". The platform's design, development, and deployment was a collaborative effort [**325**]. Cognicy serves as a proof of concept to evaluate the findings of this dissertation. *First*, this chapter explains the choice of implementation. *Second*, the use of the platform and its different functions are summarized through an illustrative scenario. *Third*, the frontend, backend, and deployment specifics are detailed. *Fourth*, this chapter delves into the strengths and weaknesses, which are deduced from the offline and online evaluation of Cognicy. The latter category was achieved thanks to the participation of 150 real users. The *fifth* section is dedicated to the discussion preceding the conclusion.

6.1. Platform versus navigator extension

The original idea was to implement Cognicy as an extension to be used and tested on real SNS accounts. However, there were two main reasons that led me in the direction of implementing it as a separate environment. *First*, it was challenging to recruit individuals to test the system even with the promise of deleting their data after this experiment. In a follow-up conducted with some of the participants, I tried to get to the bottom of the issue and investigate the reason behind their aversion. Many cited their dislike for all extensions that operate in the background because they seem insidious and resemble keyloggers[1]. A few were more expressive such as someone who replied that being one out of millions of people who are being monitored by companies did not feel threatening. But, being part of privacy-focused research using hundreds of participants, at most, made it feel more personal to them. One of them jokingly said: "I am not English royalty nor am I a celebrity whose data is worth something to the paparazzi, so, I do not fear someone hacking me or desperately spending

[1]Keyloggers, or keystroke loggers, are tools that record what a person types on a device. This malicious software records every keystroke on the victim's device and sends it to the attacker.

time and resources to profile me. It is different when I am one of the few people being observed because every entry I make is scrutinized. Also, 99% of my posts are interactions with my friends and I do not feel comfortable revealing that to strangers, researchers or not.". To address this issue, participants were informed that they could replace the real names of their friends with pseudonyms, blur the photos, etc.

Second, to fairly judge the proposed system, I had to isolate the other factors that can impact the decision-making process. On real social media platforms, a user can delay disclosure not because they had an epiphany about the consequences of their action but because they received a message or got served an advertisement as they were drafting the post. In order to know for sure if the action is a direct response to nudging, I opted for a simulated environment that offers the same functions as a real SNS, but with more control over what the user is subjected to during the experiment.

6.2. Scenario showing the use of Cognicy

This scenario shows the different steps and functions that participants experience on the platform as illustrated in Figure 6.1. The user starts by signing up, answering a questionnaire, and setting their profile attributes. At this point, they can start using the platform to share their own content or interact with existing posts. Due to the limited number of users, the news feed[2] is enriched by synthetic content from artificial profiles. They are designed to look like real people. Furthermore, Figure 6.2 shows the profile settings that Cognicy offers to the user along with a tip, which is a form of a one-size-fits-all nudge to ease their experience on the platform. Another example of this is in Figure 6.3, in which a non-user-specific disclosure-mitigating nudge is pushed to the user in response to sharing their personal email publicly in a comment. *Colour coding*[3] serves to highlight the urgent instances of disclosure. Hence, tips are displayed on a blue background, while warnings triggered by disclosure use a red one. As the user is navigating through the content, the platform captures their interests.

This is referred to in the scenario as the assessment of user attention. The platform captures the interactions (comments, posts, responses to real or synthetic content) to update the user model and the privacy report. The latter is accessible to users at all times. When they click on the logout button, two processes are triggered. First, The user is shown the analytics describing the entire session such as topics that they disclosed the most, motivations that

[2]This is the content shown on the home page without the need to click on a person's profile or a specific page

[3]Colour coding is a marking system, which can be used in various contexts such as schools and professional environments to indicate hazards, convey information, or guide employees. The more severe the consequences (fire hazard, for example) the more attention-grabbing the colour like bright red.

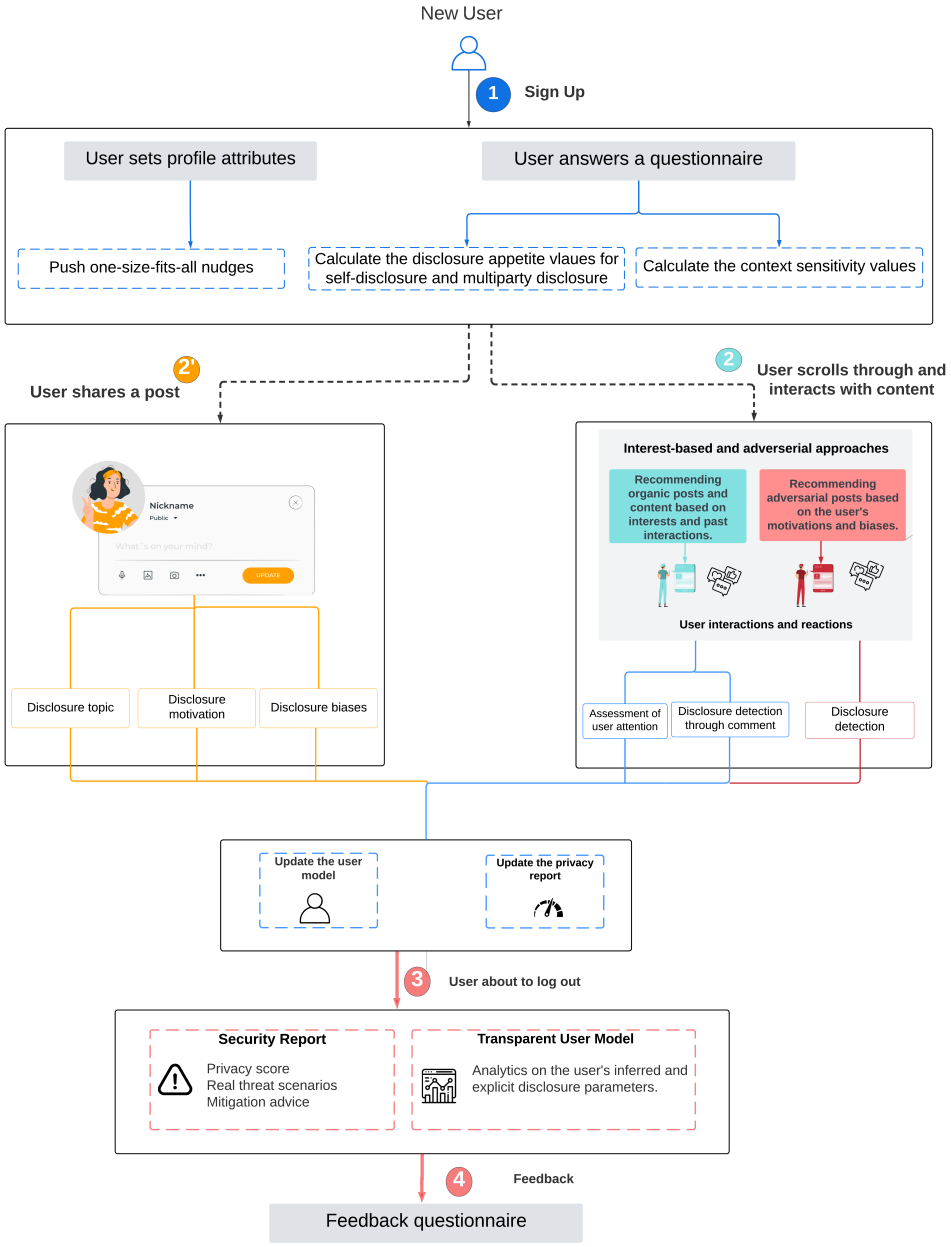

Fig. 6.1. Cognicy scenario

they are driven by, and biases they are susceptible to. Second, they are prompted to leave feedback in which they answer questions and get to express their opinions in their own words.

6.3. Implementation of cognicy

The implementation is broken down into backend, frontend, development tools, deployment, and finally the way through which artificial content[4] was added to the platform.

[4]artificial or synthetic content includes posts and comments that were not added by one of the real users recruited for the evaluation.

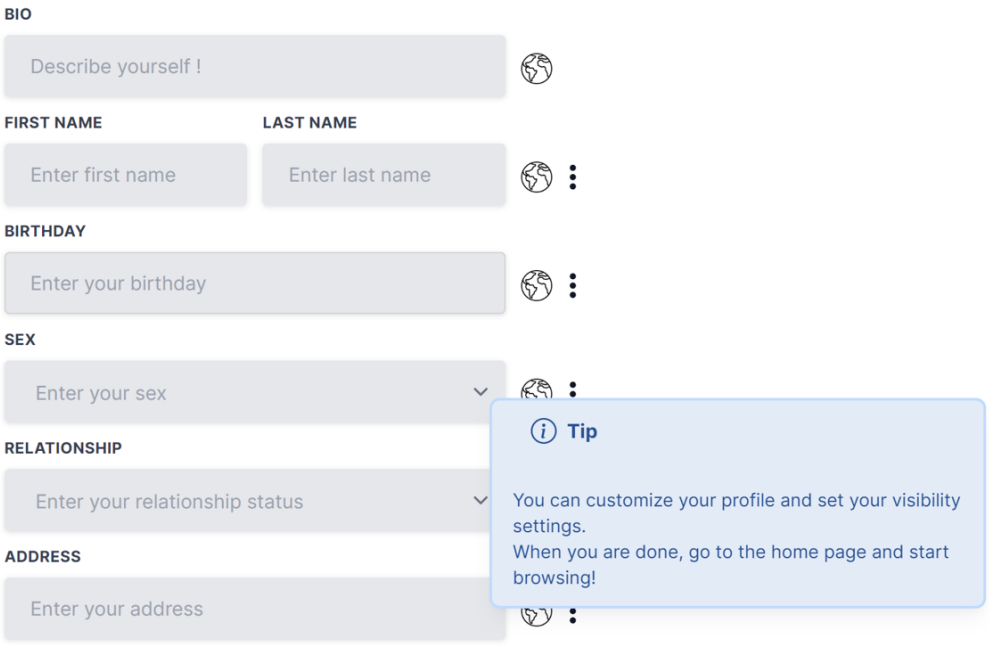

Fig. 6.2. Profile settings

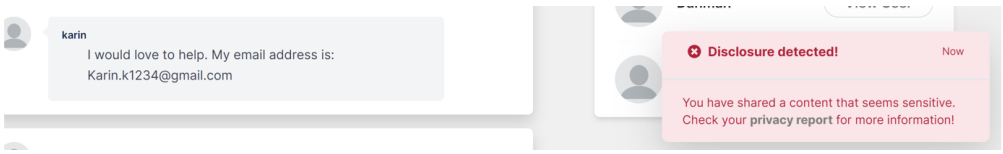

Fig. 6.3. Example of a one-size-fits-all nudge

6.3.1. Backend

This subsection enumerates the different languages, technologies, and frameworks used to implement Cognicy.

- *Python*: It is a high-level, general-purpose programming language. It is the most suitable choice for Cognicy thanks to its ease of use in web development projects and its extensive readily available libraries.

- *Django*: This high-level Python web framework encourages rapid development and clean, pragmatic design. The main advantages attributed to Django are:

- It follows the software design pattern commonly used for developing user interfaces called *Model–view–controller* (MVC). This facilitates the maintenance of the code.
- Embedded within it are advanced security functions such as the protection against brute force and *Cross-Site Request Forgery* (CSRF).
- The ability to efficiently manage a large number of users and their simultaneous activity.

- The documentation of Django is extensive and the community around it is active and interactive.

- *FastAPI*: This is another Python-based framework that has quickly gained popularity among developers due to its ease of use, speed, and robustness. It is used to build, deploy the NLP component of the system, and integrate it in Django.
- *PostgreSQL*: This free and open-source *relational database management system* emphasizes extensibility and *Structured Query Language* (SQL) compliance. PostgreSQL is used to store the users' data, history of past disclosure instances, past responses to nudges, and the different calculations (such as user interest and user model parameters). This management system ensures:

- Reliability and robustness ensuring the integrity of the data.
- Elastic storage, extensibility, and the potential to scale the system.
- Intelligent caching enabling a dynamic use of the RAM for faster access to data.

6.3.2. Frontend

The frontend part of the implementation focuses on the technology needed to bring forth the visual elements of Cognicy.
- *HTML*: The HyperText Markup Language or HTML is the standard markup language for documents designed to be displayed in a web browser.
- *CSS*: Cascading Style Sheets is a style sheet language used for describing the presentation of a document written in a markup language such as HTML or XML.
- *Tailwind CSS*: It works by scanning all the HTML files, JavaScript components, and any other templates for class names, generating the corresponding styles, and then writing them to a static CSS file. It is fast, flexible, and reliable.
- *JavaScript* (JS): It is a programming language, which is one of the core technologies of the World Wide Web, alongside HTML and CSS.
- *Chart.js*: This is a free, open-source JavaScript library for data visualization, which supports eight chart types: bar, line, area, pie, bubble, radar, polar, and scatter. This is particularly relevant for the privacy report section in Cognicy.
- *jQuery*: It is a JavaScript framework designed to simplify event handling, *CSS animation*, and *Ajax*. It is free, open-source software using the permissive *MIT License*.

6.3.3. Development tools

- *Visual Studio Code*: Also commonly referred to as VS Code, is a source-code editor made by Microsoft with the Electron Framework, for Windows, Linux, and macOS. Features include support for debugging, syntax highlighting, intelligent code completion, snippets, code refactoring, and embedded Git.

- *Google colab*: It is a cloud-based Jupyter notebook environment. It runs in a web browser allowing my collaborators and me to access and edit the code.

6.3.4. Deployment

To reach a wider number of participants, Cognicy was deployed for a month using a public IP address that is accessible via a dedicated URL. This called for the use of a *Virtual Private Server* (VPS), specifically, *Hostinger* was selected. The properties of the virtual machine on this server are illustrated in Table 6.1.

Tableau 6.1. Virtual machine details

Properties	*characteristics*
CPU	6 vCores
RAM	6 GB
Disk	120 GB
Hostname	cognicy.machine
OS	Ubuntu 22.04 64bit
IP	83.136.219.199

6.3.5. Content on the platform

The users' published posts, photos, and comments all become part of the content that can be visualized by future users. Moreover, I used Python Reddit API Wrapper to scrape the web to enrich the news feed and to create synthetic or bot profiles. Specifically, I chose 50 posts from each of the following subreddits: r/offmychest, r/antiwork, r/relationships, r/giveaways, r/motivation, r/personalfinance, r/alcoholism, r/AskAcademia, r/diary, r/CryptoTechnology, r/askHistorians, r/worldnews, and r/travel. They were subjected to manual filtering to remove the bot-generated posts such as the one in Figure 6.4 because they are clearly automated messages and defy the purpose of creating human-like synthetic data. They can be identified by detecting keywords like "Rule", "Rules", "Subreddit", and "Guidelines". Subsequently, 19 posts were removed leaving 631 of them to be shown on the home page.

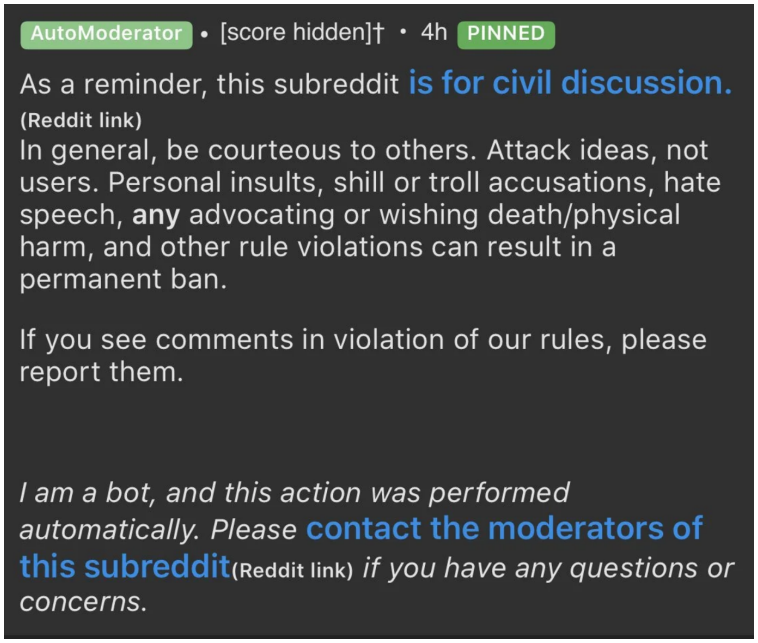

Fig. 6.4. An example of a removed bot-generated post

The next section details the visual elements that the users interact with. The annex includes more screenshots of the different functions that Cognicy offers.

6.4. Evaluation

The points that the evaluation focuses on are:

- The spread of private data on social media. This validates the motivation for my work.
- The effectiveness of disclosure detection as the first step based on which nudges are pushed.
- The usability of the system according to a widely used metric.
- The evaluation of the context-aware user model.
- The acceptance rate of nudges compared to a baseline of non-context-aware nudges.

They are examined through a series of offline and online tests.

6.4.1. Offline evaluation

The offline evaluation refers to the tests conducted on an existing dataset instead of a stream of incoming data. I use this to investigate the need for such a system, the performance of the disclosure detection module, the usability of the system, and the construction of the context-aware user model.

6.4.1.1. The spread of private data. This subsection serves as an argument for the importance of taking action and the potential repercussions of oversharing. The data used to test

the spread of private data comes from the *Stanford large network dataset* collection, specifically the *Facebook social networks* [**326**]. It has 4039 nodes, 88234 edges, and a diameter (longest shortest path) of 8. Although the dataset includes nodes alongside their connecting edges, there is no indicator of the frequency of the communication nor a log of past interactions, at least one of which is needed to compute the spread of private data. To the best of my knowledge, there are currently no repositories that offer this information. This was most likely for privacy reasons because building such a dataset would require gaining access to the records of thousands of users.

The aforementioned Facebook network does, however, include features that are anonymized so, for "*political=Democratic Party*", the publicly available data would simply contain "*political=anonymized feature 1*". For the purpose of this proof of concept, the similarity between user profiles (common features) will be used as the probability of re-sharing a post. In other words, let us suppose that Sam is the sharer and Alice and Bob are both in his immediate *intended audience*. This term refers to a group of individuals who are one connection away from Sam and who belong to the audience that Sam set for his post. If the sharer's user features are closer to Alice's than Bob's, I assume that this means that Alice is more likely to re-share the post with a *probability = similarity (Alice, Sam)*. This is not unfounded as it is, to this day, one of the main pillars of recommender systems. This also applies to social networking sites because they prioritize user engagement by serving content created by similar people [**327, 328**]. Hence, Sam's post will likely be seen by members of the audience who are most similar to him.

Seeing as each user profile in the dataset is presented as a binary vector, I use the *Hamming distance*[5], which returns the number of bits we must change to transform Alice's vector into Sam's. There are 224 features in the dataset. So, a distance of 224 means that the users are polar opposites with no similarity whatsoever. The smaller this distance, the more similar the two users are according to their vectors. I normalize these values (between 0 and 224) to [0,1] to correspond to a measure of probability (probability of re-sharing the post). Table 6.2 shows the results of the *200 simulations* that I ran. In each one of them, a random node was selected as the originator of the post. The immediate neighbours who are one connection away from this node were considered potential re-sharers. As such, if the probability of re-posting the content is over 0.5, they are added to the path. Then, the neighbours of this new node are examined, and so on and so forth until the probability of re-sharing is reduced below 0.5. If the final path is 3, that means that at least one person who is three connections away from Sam (friend of a friend of a friend) managed to view the content.

[5]Hamming distance is a metric for comparing two binary data strings. While comparing two binary strings of equal length, Hamming distance is the number of bit positions in which the two bits are different. The Hamming distance between two strings, a and b is denoted as d(a,b).

Tableau 6.2. Spread of private information based on the Facebook networks from the Stanford large network dataset

Simulation details	**Similarity-based propagation**
Number of simulations	200
Average spread of the private information	5.2
Average count of the unintended audience	47.4

The experiment revealed that on average, around 47 people see a post that the sharer did not intend for them to have access to. The spread of data stands for the average length (number of consecutive nodes or path) that information propagates through. This test was conducted to argue that social media platforms, by design, are fertile ground for disclosure due to the ease with which connections are formed. Due to this, a privacy-aware system can offer users much-needed guidance, part of which is the dislcosure detection module.

6.4.1.2. Disclosure detection module. Although the detection module is not the focal point of my contribution, it is a process that the approach hinges upon. If the output is poor in quality, the system risks overlooking sensitive data or mistakenly flagging non-sensitive pieces of information. Both are just as bad. In the first case, the user would not be nudged even in critical situations. In the second case, they would feel burdened by an uncalibrated system that warns them against sharing the most benign posts.

The dataset for this offline evaluation used contains *9917 posts*, which have been proposed to a pool of 12 volunteers: 7 males and 5 females, aged from 24 to 41 years, mainly postgraduate/Ph.D. students and researchers [**296**]. Their labels per topic, motivation, and bias are hidden from the language model DistilBERT in order to test it using the following metrics:

Correctly classified instances: This is the number of correctly identified samples such as predicting the right topic, bias or motivation.

Root Mean Square Error (RMSE) measures the average difference between a statistical model's predicted values (predicted topic, motivation, or bias) and the actual values. The

calculation of this metric is detailed in Equation 6.1.

$$RMSE(y, \hat{y}) = \sqrt{\frac{\sum_{i=0}^{N-1}(y_i - \hat{y}_i)^2}{N}} \tag{6.1}$$

$\mathbf{y_i}$: Actual label.

$\mathbf{\hat{y}_i}$: Predicted label.

N: Number of data points.

The performance of the offline evaluation is reported in Table 6.3 based on the two metrics: correctly classified instances and RMSE.

Tableau 6.3. Performance of the disclosure detection module in the offline evaluation

Task	**Correctly classified instances**	**RMSE**
Topic	7598	0.27
Motivation	6923	0.43
Bias	7246	0.39

6.4.1.3. Nudge-based assistant. This evaluation relies on scenario-based data that was proposed to 1000 participants who were recruited on Mturk. It was conducted before the deployment of Cognicy, as part of an article entitled "*A Multi-Agent System for Privacy-aware Nudges*", which I submitted to the journal of "*Behaviour & Information Technology*". The demographic details are shown in Table 6.4. The process started with a consent form informing them of the purpose of the research, the affiliation of the researchers, information on the anonymity of the responses, and the irrevocable right of withdrawal. Moreover, information on the ethical board that has approved this research was provided to the users with references to my specific project.

The recruited users answered two consecutive sections: the first one is designed to construct their user model and the second to test their responses given hypothetical scenarios. If the user answers “yes” meaning that they would disclose personal data, then, they receive a nudge. Table 6.5 shows some examples of the scenarios per motivation.

Across all the scenarios, the demographics “18-24” and “45+” showed the highest contrast in terms of being willing to disclose or not (Figure 6.5). Participants aged 45+ showed less inclination to do so but it was still a considerably high percentage of 43%. Each instance of disclosure was followed by a nudge and the following summarizes their acceptance rate per topic:

Tableau 6.4. Demographic characteristics of the study sample

Demographic variable	**Category**	**Frequency %**
Gender	Female Male I prefer not to answer	43% 56% < 1%
Age	18-24 25-34 35-44 45+ I prefer not to answer	30% 32% 20% 17% < 1%
Education	Student Employee Business owner /Self-employed Manager/Official Retired/Unemployed I prefer not to answer	43% 29% 15% 6% 4% 3%

(1) Identity: 55%
(2) Location: 78%
(3) Experiences: 63%
(4) Alcohol, drug consumption, and health information: 84%
(5) Religious and political views: 71%

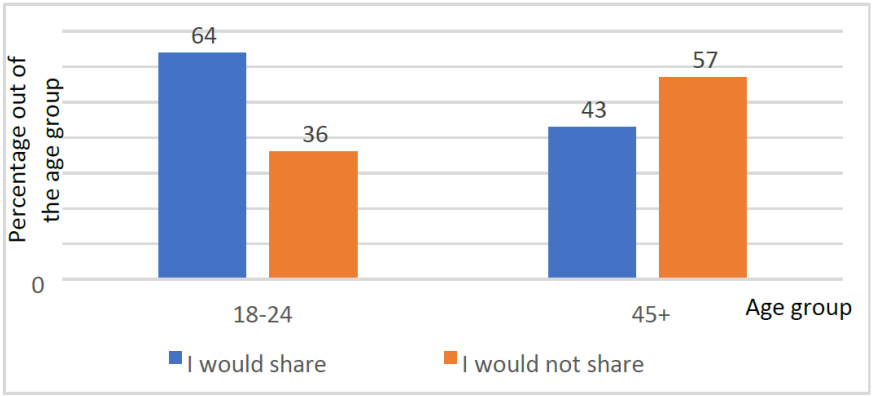

Fig. 6.5. Response to the hypothetical scenarios

The offline tests were promising enough to encourage further investigation using real participants, which brings us to the online evaluation.

6.4.2. Online evaluation

The online evaluation is based on 150 participants who are mainly: fellow researchers at our lab, their family members, and other researchers recruited through diffusing the message by email and on LinkedIn.

6.4.2.1. System usability. *Usability* is defined by the quality of the *User Experience* (UX). System usability does not judge how well the system performs in its privacy preservation task, it is an indicator of whether the individual is satisfied with its ease of use and simplicity. A commonly used way to evaluate usability is the *System Usability Scale* (SUS) [**329**]. A score ranging from 0 to 100 is generated based on ten questions that the user answer on the following five-point *Likert scale*[6]: "partially disagree, disagree, neutral, agree, partially agree". They are:

(1) I think that I would like to use this system frequently and if it is adapted into an extension, I would use it on my real social media accounts.
(2) I found the system unnecessarily complex. Some functions need to be simplified.
(3) I thought the system was easy to use and I got the hang of it quickly.
(4) I think that I would need the support of a technical person to be able to use this system. It is not as intuitive as other social media.
(5) I found the various functions in this system like the privacy report, nudges, and tips to be well integrated.
(6) I thought there was too much inconsistency in this system. These inconsistencies are visual and/or functional.
(7) I would imagine that most people would learn to use this system very quickly.
(8) I found the system very cumbersome to use.
(9) I felt very confident using the system.
(10) I needed to learn a lot of things before I could get going with this system.

These questions are adapted by adding examples and details relating to the proposed system such as mentioning the extension in the first one. But, they are not changed in meaning. My goal was to compare the usability of Cognicy with similar systems and propositions. I wanted to ask the same users whom I recruited to evaluate the other platforms for fairness. However, the researchers did not offer access to their respective systems. The only alternative was to consider the SUS values that they reported. Table 6.6 is not concrete proof showing which approach is better, it simply serves as a way to demonstrate that Cognicy's usability resembles that of other privacy-focused solutions.

[6]When responding to a Likert item, respondents specify their level of agreement or disagreement on a symmetric agree-disagree scale for a series of statements. Thus, the range captures the intensity of their feelings for a given item

According to *Userpeek*[7], a service that provides user evaluation services, a score higher than *68* is above average. It is worth noting that the SUS score of Cognicy was 71.59 during the first evaluation. This led to the addition of a tutorial at the beginning highlighting all the functions with screenshots and images. An "about page" was also added for users to consult at a later time. This increased the score to 76.83 as reported in the table.

6.4.2.2. Disclosure detection module. The offline evaluation of the disclosure detection module based on an existing dataset was detailed in Section 6.4.1.2. The online evaluation was conducted based on the input of the participants. I was going to assign labels to the posts myself and compare them with the output of the disclosure detection module, but I opted for another approach to avoid inserting my biases in the process. At the end of each user's session and after they give the feedback on the platform (Figure 6.6), they are asked to contribute to one last task: labelling other users' posts. Alice would be given other users' posts, which do not contain any unique identifiers and she would label them per topic, motivation, and bias. Each post considered in the online evaluation has been labelled by 5 people, which is an odd number to avoid ties. The majority rule is used: If 3 users label the post as topic1 and the other 2 as topic2, then topic1 is the chosen label. There have not been any cases in which fewer than 3 people agreed on one label. The output of this process is considered the "actual label" and is compared to the "predicted label" to evaluate the online performance of the disclosure detection module (Table 6.7). The formulas of the metrics used have the same definition as in Section 6.4.1.2.

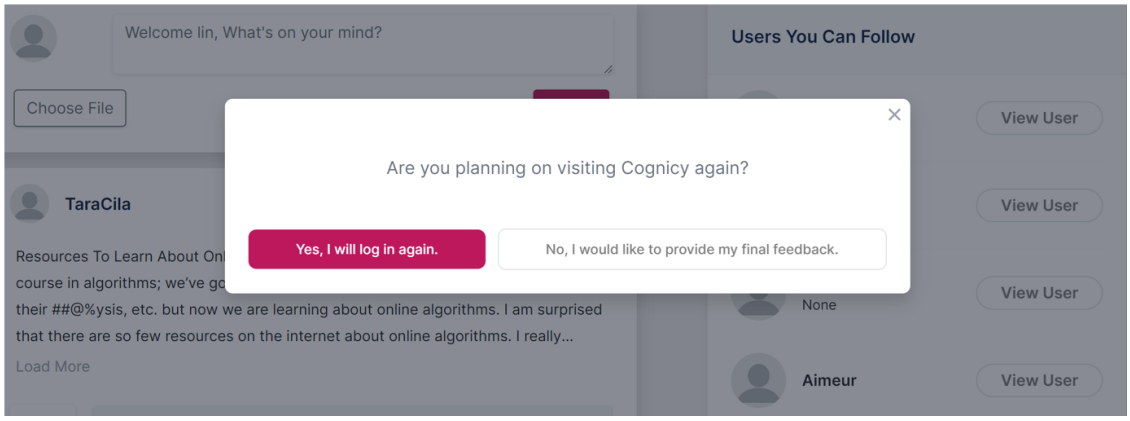

Fig. 6.6. Prompting the user for feedback

Overall, the model performed better in the offline testing compared to the online evaluation (Tables 6.7 and 6.3), which was expected considering that the dataset used for the former had lengthier posts that offered more context. Some of the online posts are three words long such as: "*I feel bad*". They were not removed prior to the testing because the goal is to prepare for actual social media posts outside of the perimeter of Cognicy and in reality,

[7]`https://userpeek.com/blog/system-usability-scale-in-ux-research/`

people do share such short texts from time to time. One way to potentially overcome this in the future is to consider the user history to infer more information about the current post. For example, if Alice has a habit of sharing short negative posts similar to "I feel bad", the system can consider the fact that in the past, her friend Bob replied "*What's wrong, are you still feeling sick?*" to be a helpful piece of context. Of course, this is not a guaranteed result because sometimes, there are not enough additional details to compensate for the lacking information in the original post. Furthermore, the user might post the same short text to mean different things such as "*I feel bad...because I am homesick*" or "*I feel bad...because I am stressed out*".

6.4.2.3. Evaluation of the context-aware user model. The context-aware user model is based on the Rasch adaptation, which in turn is based on two main parameters: person value and item value. The metrics that are conventionally used to evaluate this are: *reliability* and RMSE. The parameters were estimated using the *Joint Maximum Likelihood Estimation* (JMLE), which was chosen since it is capable of being scaled to accommodate large datasets and remains robust against missing data [**332**]. Moreover, its main disadvantage, which is the potential for bias when using a few items, is overcome with an embedded correction in *Winsteps*[8]. Going back to the significance of the metrics, reliability is a measure of how reasonably robust the findings are to the (relevant) data or methods employed [**333**]. It is one way to check that the model is not overfitting and reporting good results that are unlikely to be replicated with new unseen data. I used two different approaches to reliability: *Apha Cronbach score* and *Rasch reliability*. In general, Cronbach Alpha overestimates this measure while Rasch reliability underestimates it [**334**].

So, for good measure, I used both. The former is known for its performance in classical test theory and during this evaluation, it is reported as 0.91, which according to Bond and Fox [**333**] is considered to be the highest (0.9-1.0) and a "*very good and effective level of consistency*". The results of the latter measure, the Rasch reliability, are reported below. They are averaged since the Rasch model was used a total of four times.
For person values:

- Reliability: 0.82
- RMSE: 0.024

For item values:

- Reliability: 0.86
- RMSE: 0.038

Rasch reliability is more conservative and less misleading [**334**], and in this case, corresponds to 0.82 (person value) and 0.86 (item value). When 0.81 is considered to be more than acceptable [**333**], the results of the reliability metrics: Apha Cronbach score (0.91)

[8] `https://www.winsteps.com/index.htm`

and Rasch reliability (0.82 and 0.86) are promising for future reproducibility. As for RMSE values, they support the good fit of the model since a measure of 0.05 or smaller is a "good fit", between 0.0 and 0.10 is considered "mediocre" and $> 0.1 =$ is deemed a "poor fit" [**335**]. Hence, taking the reliability and the RMSE into account, JMLE yielded good results.

6.4.2.4. Evaluation of the nudging mechanism. This subsection starts by comparing non-context-specific nudges with context-specific nudges. The significance of the context has been discussed at length throughout the dissertation, but what does non-context-specificity actually mean? I interpret it to include one-size-fits-all nudges as well as personalized nudges whose approach is user-specific but does not adapt to the circumstances. The former category can be tested thanks to Aegis, which is based on objective non-user-specific domain knowledge. If this agent deems disclosure to be sensitive, a nudge is pushed. Testing the latter required using the Rasch model one more time but instead of providing it with user data that has been labelled to include the context (history + questionnaire responses), the model gets the same bulk of information without any granularity. Hence, Figure 6.7 shows the results of using one single β and δ values for each user regardless of the bias, motivation, and topic to push the nudges. If Alice is about to share a geotagged[9] photo, she might receive the following nudge: "*You are about to share a geotagged photo, which you have previously rated as sensitive. Would you like to revise this decision?*". Its context-aware counterpart would have used the user's motivations and biases to frame the nudge.

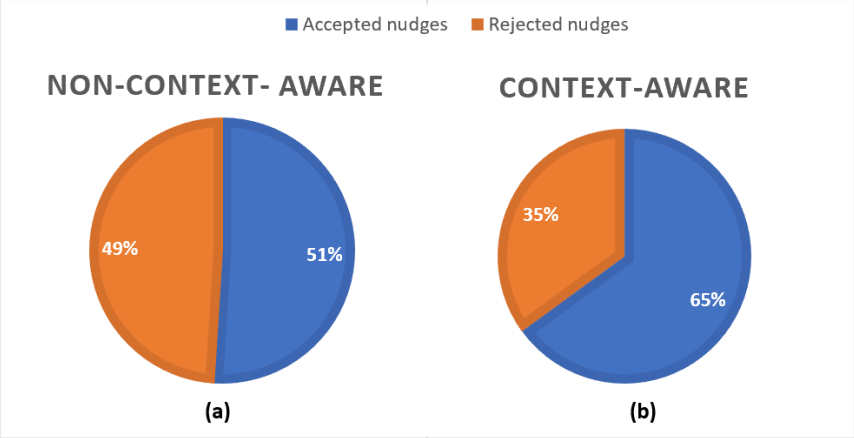

Fig. 6.7. Comparing the acceptance rate of non-context-aware (a) and context-aware personalized nudges (b)

The results reported on the right side of Figure 6.7 (b) corroborate the importance of using context-specific values. While the acceptance rate of non-context-aware nudges barely reached *51%*, my proposed approach increased this value to *65%*. Furthermore, Figure Figure

[9]Geotagging, or GeoTagging, is the process of adding geographical identification metadata to various media such as a geotagged photograph or video, websites, SMS messages

6.8 makes a comparison with a standalone version of Aegis (without the collaboration of the other agents). Comparing (a) with (b), an increase was noted from 12% to 72% for the age group 25-34, from *31%* to *63%* for people aged 18-24, from *33%* to *66%* for the 3rd age group, and finally, an increase of *21%* from *38%* for the category 45+.

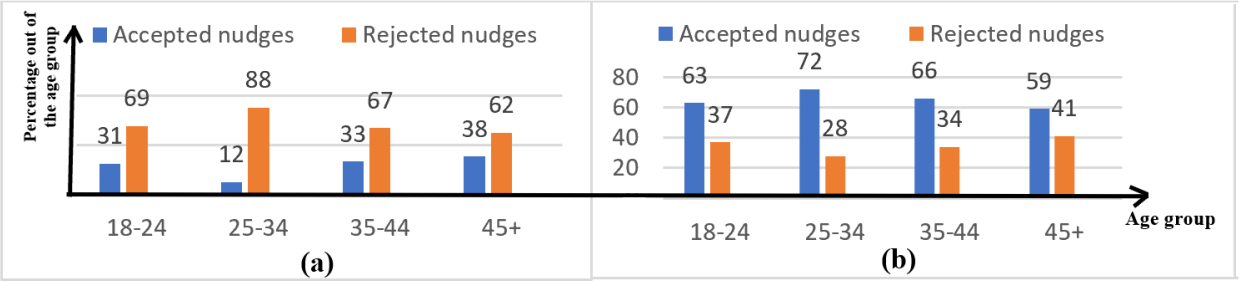

Fig. 6.8. Comparing the acceptance rate of one-size-fits-all (Aegis) (a) and context-aware personalized nudges (b)

The next evaluation was conducted on the context-aware personalized system to investigate the acceptance rate in self-disclosure versus multiparty disclosure scenarios. The participants showed more willingness to accept the latter over the former when the party concerned is either close friends and family (first social circle) or co-workers/classmates (second social circle). As expected, nudging against disclosing information about a stranger from the general public was the least successful (*31%*). The rest of the categories were ordered as follows: self-disclosure (*70%*) followed by the second social circle (*75%*), and finally the first social circle (*86%*).

I scrolled through all the published posts and comments on the platform and I noticed that self-disclosure posts seem genuine and realistic. However, the multiparty disclosure instances are few and far in between and do not perfectly reflect how SNS users usually interact. They seem formal and do not include inside jokes or slang that one would expect to read in friendly communication. I think that this is due to the presence of few users and also, the absence of people with whom users have a strong interpersonal relationship, which promotes self-disclosure and multiparty disclosure. Mimicking years of back-and-forth communication in such a setting is a very difficult task.

Finally, users could give their opinion on the platform, what they liked, disliked, and some general suggestions to improve Cognicy. They particularly appreciated the presence of the privacy report (screenshots of which are added to the appendix). Figure 6.9 highlights their responses when asked if the visualization allowed them to gain more awareness of their behaviour, to which *80%* answered that they agree or partially agree.

In addition, many of them criticized the lack of entertaining elements on the platform such as mini-games, active communities, and fandom pages. There was an interesting suggestion

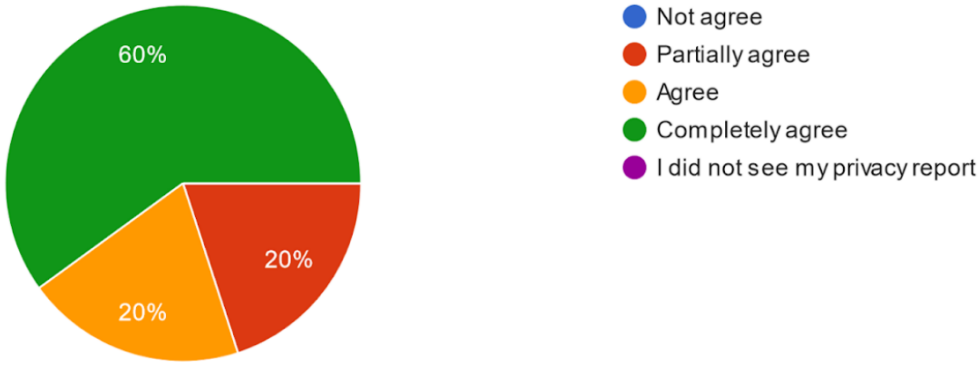

Fig. 6.9. Responses of the users to whether the visualization helped raise their awareness

about creating a future version of Cognicy for children who are eager to be on social media but whose parents are rejecting the idea for their protection. This might be a compromise to give them a closed-off environment that parents can have absolute oversight over. Of course, if such a platform came to be, it would be with a different foundation and approach.

6.5. Discussion

I would like to point out a few limitations with the aim of tackling them in the future. The *first* of which is that due to the limited number of participants in the online evaluation, a follow-up study needs to be conducted with a focus on diversifying and enlarging the pool of users.

Second, the current version of the Rasch model that I have used would not be able to withstand thousands of users, especially since the nudges should be prompt. An extended version can be the solution to managing changing data, missing values, and scalability issues [**336**]. Cognicy itself would need to be changed into an extension and serve as a privacy guide for the user. This comes with its own challenges. If users are not willing to use it now, which is one of the reasons I opted for a platform instead, what would make them eager to do so in the future? For better or for worse, individuals would rather trust companies and large corporations than a small group of researchers, at least based on the follow-up interviews I conducted with the participants. It makes sense because organizations can be held accountable and have a vested interest in preventing major issues. However, that makes deploying Cognicy for a large population difficult. I think that this project should be approached, in the future, in collaboration with a larger body such as a university, company, or government initiative to reach a wider audience and also offer some insurance to users to be willing participants.

Third, while this dissertation contributes to the field of privacy, I acknowledge that disclosure can be beneficial in certain situations and even necessary at times. The Social Penetration Theory [**104**] explains how revealing information about oneself is crucial to the development of interpersonal relationships. Seeking a close relationship with others cannot be achieved without divulging some aspect of one's private life. My work does not propose eradicating every instance of disclosure. The issue is that there is a larger audience online, one that never forgets, and the information can cause unfavorable consequences to the user. It would have been somewhat acceptable if an adult willingly chooses to revoke their right to privacy for whatever objective, but the issue goes beyond that. Children growing up in the digital age are deprived of a normal childhood because numerous parents have fully adopted the practice of oversharing, which extends to their babies from the time they draw their first breath. This is not about rare edge cases since 75% of parents post about their kids on SNS[10]. What is even more alarming than this is that 8 out of 10 of them have followers they have never met. With the rise of deep fakes, who knows what danger the young ones are facing through no fault of their own?

Fourth, comes the issue of "infantilization" as Beck calls it while arguing that persuasive technologies are far from enhancing autonomy [**337**]. Even the authors who popularized the term note that nudges, by design, take advantage of judgemental heuristics and cognitive biases [**8**]. This brings us back to a question that I posed multiple times throughout this work, "if not this, what can we do then?". Users are not going to spontaneously gain knowledge and awareness overnight. If they are not nudged in the direction of privacy preservation, with the proliferating inescapable deep roots that technology has laid into the foundation of our lives, would they not move further in the direction of oversharing? Another point to consider is that the argument against this type of choice architecture can be used in its favour as well. If an adult user gives their informed consent to be nudged after understanding the purpose of this method, it is an equal if not a more severe instance of infantilization to ignore their decision and assume that they were beguiled by the lure of this choice architecture.

6.6. Conclusion

This chapter goes through the details of the implementation from the backend to the front end and deployment. Furthermore, various visual parts are included as part of the user interface. As reported in the evaluation, the context-aware personalized nudges proved better received by users than generic one-size-fits-all and non-context-specific interventions. This was achieved thanks to the collaboration of every module: disclosure detection, domain knowledge, user model, and multi-agent assistant. Cognicy offered users a simulated

[10] `https://campaignsoftheworld.com/digital/deutsche-telekom-a-message-from-ella/`

environment designed to resemble SNS like Facebook and Twitter. The added features like the privacy report and prompt nudges were appreciated by the participants. The evaluation corroborates the potential of the system.

Tableau 6.5. Examples of disclosure scenarios in the questionnaire as part of the offline evaluation

Disclosure motivation	**Disclosure scenarios**	**Disclosure category**
Financial gain	You are healing from a severe illness. You developed specific habits like your diet or your bedtime that could help other patients who suffer from the same disease. You are prepared to sell your medical record including this information.	Self-disclosure
Personal gain	You activated a new account on a video streaming platform. You can get more personalized recommendations but to do so, the platform needs access to your specific location.	Self-disclosure
Moral gain	You have been through a harrowing experience while travelling and although the details are very sensitive, you feel it could help others make better decisions. It is, therefore, necessary to describe everything.	Self-disclosure
Self-expression	After a frustrating day at work, you log into your SNS account and see posts made by other co-workers who are enjoying their evening. It further irks you so, you post a lengthy comment about how you are doing most of the work and the impact of your current situation on your mental health.	Multiparty disclosure
Development and maintenance of interpersonal relationships	As a member of a Facebook group created for new parents, you often see others sharing their experiences such as photos of their kids walking for the first time, throwing tantrums, celebrating birthdays, etc. You want to form deeper connections with the other members to be able to depend on them for support, as such, you follow suit and share photos of your own child.	Multiparty disclosure

Tableau 6.6. Comparion based on the SUS score

System	**Description**	**SUS** score
Hartwig *et al.* [**330**]	Nudging users towards better security decisions in password creation.	71.57
Yoshikawa *et al.* [**331**]	Opportunistic microlearning about online safety and ethics.	74.19
Cognicy	**Nudge-based simulated social media platform Cognicy.**	**76.83**

Tableau 6.7. Performance of the disclosure detection module in the online evaluation

Task	**Correctly classified instances**	**RMSE**
Topic	7312	0.27
Motivation	6647	0.32
Bias	7029	0.40

Chapter 7

Conclusion

SNS are continuously growing in popularity among all age groups. Gone are the days when oversharing on these platforms was out of the norm or frowned upon. One could log into TikTok, for example, at any given moment and watch a myriad of videos in which people reveal sensitive and intimate parts of their lives. Short format content, in particular, seems to highly promote disclosure to capture the interest of a user that is constantly scrolling through their feed. I cannot help but wonder if the current reality of privacy or the lack thereof is something that George Orwell would have found to be an interesting premise for another one of his page-turners. Who would have thought that in the twenty-first century, the most invasive infringements on privacy are self-inflicted?

While privacy is not on its deathbed, it is not thriving either. It is suffering now more than ever, but the silver lining in all of this is that there are still ways to nurse it by raising individuals' awareness and using positive behaviour reinforcement. But first, we must acknowledge that the elimination of any and all forms of disclosure is unrealistic in today's world. As such, my research is focused on understanding the motivations and biases behind self-disclosure and multiparty disclosure before addressing these behaviours through nudging. For a long time, the potential of nudges has been proven time and time again in a multitude of applications such as healthcare, environmental causes, marketing, etc. Combining these behavioural alterations with cybersecurity and privacy awareness is a more recent development.

My contributions in this thesis can be broken as follows: The first is modelling the different actors in the economics of privacy and establishing an equilibrium between the data owner, broker, and data user. This builds a basis for a better understanding of the different transactions and what each party stands to gain/lose. The second accomplished objective consists in designing a framework for multifaceted privacy decisions. Multipriv offers a deeper understanding of the different factors that promote, dissuade or catalyze disclosure actions.

The three-layered framework is detailed from the most general environmental inputs to the most specific context layer. The third contribution is the proposed nudge-based approach that serves as an immediate intervention when the user is about to share sensitive content. To achieve this, I design the domain knowledge, propose a novel context-aware user model to represent each party's perspective, and establish a method for multi-agent mediation. The latter objective uses an adapted version of the Rasch model that I devised to fit the context of privacy and cybersecurity awareness. The fourth major goal that was fulfilled is the evaluation of the aforementioned contributions. This is done through Mechanical Turk (offline) and the personal recruitment of participants (online). The promising results of this process further strengthen the current state of this research and pave the way for more development in the future.

There was a reported sentiment among users expressing how they want to make as little effort as possible or else they might not be interested in using a privacy-preserving system. This corroborates the findings of the literature on privacy fatigue and the general apathy that many people have while surfing the web. It also serves as a guideline for the future to think about the potential but also the implications of further limiting user engagement. While I could tweak the system to propose the edited post to the user instead of pushing a nudge telling them "delete this part" or "blur this face", is that really a step in the right direction? Would that not further feed into the individual's reliance on another party to do all the thinking, which hinders the decision-making process in the long run? There is a fine line between assisting users and rendering them dependent on the system. Figuring out where that line lies is something of great interest to me. Another future aim is the design and implementation of the nudge-based system as an extension on SNS such as Facebook, Twitter, Instagram, and TikTok. This is expected to be a challenging task as these platforms have different content format from text to photos and videos. Furthermore, execution time is of the essence in a real-time application and as such, figuring out how to optimize it without compromising the quality of the output is another question that lies ahead.

7.1. Contributions

Users' privacy decision-making process is susceptible to biases and other contextual factors. Moreover, their perception is malleable and they can get privacy fatigued, a sentiment that is exacerbated by the increasing number of decisions that they have to make on a daily basis. Nudges offer a great opportunity to encourage behavioural changes that mitigate disclosure without reinforcing restrictive *paternalistic* shackles on free will. The end goal of this dissertation is to answer the questions: "*How can personalized context-aware nudges overcome the shortcomings of the one-size-fits-all paradigm? How efficiently can this be applied*

to self-disclosure and multiparty disclosure?" To address this, five main contributions were achieved:

7.1.1. Designing a multi-agent system to understand the economics of privacy

In Chapter 3, I propose a multi-agent system representing all the actors in the economics of privacy: the data owner, the broker, and the data user. The transactions and data exchange between them mimics real-life scenarios. They negotiate and try to maximize their own utility under certain restrictions such as data anonymization, quality, payment per sample of data, etc. The aim of this system is to conceptualize privacy preservation as a complex issue that depends on the collaboration of all three actors. Ensuring that the data brokers and users operate ethically and legally depends on the regulations and restrictions put in place by the judicial system. As for the data owners, it is high time that they regain a sense of control over their own privacy. Sir Tim Berners-Lee who invented the World Wide Web in 1989, said "*As our data is held in proprietary silos, out of sight to us, we lose out on the benefits we could realize if we had direct control over this data and chose when and with whom to share it.*" [**338**]. Out of the three players in the economics of privacy, this dissertation focuses on the data owners with the purpose of understanding their decisions and then nudging them using personalized context-aware libertarian paternalism.

7.1.2. Proposing MULTIPRIV: A framework for MULTIfaceted PRIVacy decisions

With a focus on the user, who is the data owner, Multipriv leverages the rich literature on disclosure to present the multifaceted issue of privacy decisions in a layered framework. The objective of Multipriv is to understand the user in order to nudge them in the most efficient way. Its structure is designed to detail the factors contributing to self-disclosure and multiparty disclosure in order of specificity. The environment, which includes cultural norms and judicial factors, is on the top as it impacts numerous individuals. It is followed by the middle layer, which details long-term user-specific factors such as motivations and the support tools available to the user. The third layer is the most specific and it is the context. The same person can make very different decisions based on who has access to the content, for example. All of this is detailed in Chapter 4 along with the challenges facing the implementation of this framework.

7.1.3. Proposing a nudge-based system to mitigate disclosure

I proposed a system composed of the following main modules: *Domain knowledge*, *disclosure detection*, *user model*, and *multi-agent assistant*. Each of them is detailed in Chapter 5.

The user model offers a novel perspective of the individual's context-aware characteristics that lead to sharing private information. I coined the term *disclosure appetite*, which is used in this module along with the context sensitivity to portray the duality within privacy decision-making. Finally, the multi-agent assistant is the core of the system. It relies on all the other modules to represent each party involved in the disclosure in a way that reflects the topics, motivations, biases, and audiences of the disclosure post. The mediator's goal is to reach a consensus by reaching out to each agent and finding a compromise that fits everyone's user model. I call the proof of concept of this nudge-based system *Cognicy*.

7.1.4. Implementing the system as a simulated SNS called Cognicy and evaluating it

After implementing Cognicy, I conducted a two-fold evaluation: offline using an existing dataset and online based on 150 real participants whom I personally recruited as detailed in Chapter 6. The offline process started by reinforcing the need for the system proposed by this dissertation. It does so by simulating the spread of data in real networks using the *Facebook social networks* [**326**]. Following this, the disclosure detection module is tested and its performance is deemed good. In addition, the nudge-based assistant is evaluated using answers from participants who were recruited on *Mechanical Turk* and presented a scenario-based survey. The online evaluation included the *System Usability Scale*, the disclosure detection, and the nudging mechanism. I compared the one-size-fits-all nudges and my proposed context-aware personalized nudges. The results highlighted the higher acceptance rate of the latter over the former. The feedback that I received from the participants at the end of the experiment further corroborates their appreciation for the personalized privacy report, the visual representation of their motivations and biases, and the timeliness of nudges. They also offered constructive criticism about the lack of entertainment on the platform, which will be considered in future research along with other directives that I have been reflecting on.

7.2. Future perspectives

The findings of this dissertation offer a novel approach to privacy nudges. The results of the evaluation are promising, not solely for the current work, but also for future avenues. I will discuss a few points that I found intriguing and would like to build on:

7.2.1. Studying the long-term impact of the context-aware personalized nudges

Nudges are often evaluated based on their immediate effect. They are deemed successful if the user accepts them and unsuccessful in the other case. However, privacy preservation is not a one-time cure-all elixir, it needs consistency and up-to-date knowledge of the advances in technology. Thus, I would like to study the effect of the proposed system in this dissertation in the long term. If nudges end up being efficient only as immediate solutions, then, I would like to combine them with *serious games* or *persuasive games*. As explained in Section 2.3 of Chapter 2, there is a difference between the two. Serious games are defined as having an "explicit and carefully thought-out educational purpose, not intended to be played primarily for amusement" [**184**]. Whereas, persuasive games take it one step further as they are serious games whose goal is not just to educate people but also to change their behaviour. Both of them can be used in collaboration with nudges with the aim of providing immediate intervention and long-term behavioural alteration.

7.2.2. Using context-aware nudges to mitigate the harms of generative Artificial Intelligence

Generative Artificial Intelligence (AI) is used here as an umbrella term to describe creative machine learning solutions trained on very large datasets in order to produce responses to user prompts. With the rise of this innovative subcategory of AI, privacy is bound to become more jeopardized than ever. Let us consider the example of *Lensa AI*, which is an app that transforms selfies into digital paintings. It gained popularity at the end of 2022 and leading up to 2023 as it amassed about *22.2 million* worldwide downloads [**339**]. A flood of users paid for this service and had to provide 10-20 selfies [**340**] whose use was "***perpetual***, ***irrevocable***, *nonexclusive, royalty-free, worldwide, fully-paid, transferable, sub-licensable license to use, reproduce, modify, distribute, create derivative works of your User Content, without any additional compensation*". This means that the use of data has no given end and that the user cannot request the deletion of their information if they realize later on that they want to protect their privacy. Only after major controversy were the first two terms changed into "*time-limited and revocable use*" [**341**]. Most of the users were following the trend and were completely unaware of Lensa AI's terms and conditions. Others did not mind even after being informed because they thought that their photos were already out there anyway. *ChatGPT* is yet another form of generative AI that is text-based. When I asked it "*Do you harm my privacy?*", it responded with "*As an AI language model, I don't have access to personal data about individuals unless it has been shared with me during our conversation. I am designed to respect user privacy and confidentiality.*" This is an admission

of the fact that any user input is collected to be used in future training. While this is not highly disturbing since almost all websites do this to some degree, generative AI remains fairly uncharted territory. As such, it is hard to predict the true impact of these models on privacy as we know it. This is especially true in the case of chatGPT as GPT-3, a predecessor of the current version (GPT-4), already showed the potential for being used in scams. It wrote more convincing phishing emails than those generated by malicious humans [**342**]. *OpenAi*, which is the company that launched chatGPT is currently being investigated by *federal and provincial privacy authorities* in Canada following a complaint stating that the company ***unlawfully*** *collected, used, and disclosed personal information* ***without consent*** *through ChatGPT* [**343**].

The duality of increasingly invasive technological developments and the growing sense of *privacy fatigue* amongst users can be a recipe for disaster in the near future. In an instance of "*life imitating art*", we might find ourselves living in a similar scenario to the one shown in season six, episode one of *Black Mirror*. It portrays the dire consequences that the main character, Joan, met after finding out that every detail of her life is made into a series for others' enjoyment, called "Joan Is Awful". Every line of dialogue and interaction she had, even with her therapist is replicated. The cherry on top is that she could not take any legal action against the streaming service because she agreed to its terms and conditions when she signed up as a customer.

To summarize, most of the privacy issues on the data owner's side stem from a lack of awareness amongst users, combined with a sense of privacy fatigue, and susceptibility to the "*I have nothing to hide*" fallacy [**109**], all of which were discussed in Chapter 2. They can potentially be addressed through a nudge-based system in the form of a plug-in that acts as a disclosure-mitigating companion to the user.

7.2.3. Investigating the potential for fake news mitigation

The threats online are not exclusive to privacy as *fake news* has been on the rise because of automated tools and bots that facilitate its spread. Almost every platform online uses a form of a recommender system to tailor the newsfeed of users. Hence, one can find themselves in an *echo chamber* of misinformation. Echo chambers can be defined by the segregation of the public space into divided clusters of people who only consume content that fits their opinion or point of view further reinforcing biases, especially the *confirmation bias*. These spaces are social structures systematically excluding sources of information not necessarily by omission but through deliberate action [**344**].

Both privacy and fake news are more intertwined than they appear to be at surface level. A report by a journal on internet regulation in Europe [**345**] found that privacy regulations, namely *GDPR*, limited the collection of personal data, which in return reduced

micro-targeting[1]. In doing so, fewer echo chambers are created because individuals cannot be easily profiled. As a result, this mitigated the spread of fake news thanks to stricter privacy laws. Moreover, I would like to explore my own hypothesis that aside from regulations, *privacy awareness can limit the spread of fake news*. I think that individuals who have developed the ability to criticize and question the harm of online content have already processed their own behavioural biases. As a result, they are unlikely to be driven by confirmation bias, which is at the core of fake news propagation.

This subject felt particularly familiar to me because nudges and recommendations exist within the same scope of "suggestions" made to users. The difference between the two and the reason for which I use the term nudge throughout this dissertation has been discussed in Chapter 2, Section 2.5. Thus, contributing to this growing body of research is a great interest of mine. I already started working on this in collaboration with Dorsaf Sallami in a paper published in the proceedings of the *ACM Conference on User Modeling, Adaptation, and Personalization*, 2023 [**346**]. The proposed approach, *FAke News Aware Recommender system* (FANAR), is an alteration of the *collaborative filtering strategy* with the purpose of preventing the propagation of fake news by detecting and avoiding untrustworthy neighbours. It would be interesting to implement a system that goes one step further and pushes timely nudges to prevent the spread of misinformation. Contexual biases are at play in all decision-making situations. For example, people are more likely to accept information that aligns with their belief (*belief consistency*) without seeking out counterarguments or trying to disconfirm it (*confirmation bias*) [**347**]. I would like to equally investigate the impact of motivation and audience on the decision to share a dubious article. I assume that people who have a larger following would feel more reluctant to do that because they would fear the backlash. However, one could exist within a large echo chamber of like-minded people that encourage such action through reward mechanisms (likes, retweets, re-shares, etc.).

To conclude, this dissertation can lead to even more significant findings in the future that are not solely focused on privacy. Adjacent subjects such as fake news and generative AI can be explored with the same aim of mitigation and behavioural alteration.

[1]Microtargeting is the use of online data to tailor advertising messages to individuals, based on the identification of recipients' personal vulnerabilities.

Références bibliographiques

[1] Alan Glasper. Dispelling anti-vaxxer misinformation about COVID-19 vaccination. *British Journal of Nursing*, 30(6):374–376, mar 2021.

[2] Tiago Oliveira, Benedita Araujo, and Carlos Tam. Why do people share their travel experiences on social media? *Tourism Management*, 78:104041, jun 2020.

[3] Robert H. Gass and John S. Seiter. *Persuasion*. Routledge, feb 2022.

[4] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. Secrets and likes: The drive for privacy and the difficulty of achieving it in the digital age. *Journal of Consumer Psychology*, 30(4):736–758, October 2020.

[5] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. Privacy and human behavior in the age of information. *Science*, 347(6221):509–514, jan 2015.

[6] Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. "i hear you, i feel you": Encouraging deep self-disclosure through a chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, apr 2020.

[7] Yana Dimova, Gertjan Franken, Victor Le Pochat, Wouter Joosen, and Lieven Desmet. Tracking the evolution of cookie-based tracking on facebook. In *Proceedings of the 21st Workshop on Privacy in the Electronic Society*. ACM, November 2022.

[8] Richard H. Thaler and Cass R. Sunstein. *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press, 2008.

[9] Yang Wang, Pedro Giovanni Leon, Alessandro Acquisti, Lorrie Faith Cranor, Alain Forget, and Norman Sadeh. A field trial of privacy nudges for facebook. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, apr 2014.

[10] Sina Ostendorf and Matthias Brand. Theoretical conceptualization of online privacy-related decision making – introducing the tripartite self-disclosure decision model. *Frontiers in Psychology*, 13, oct 2022.

[11] Amanda L. Forest, Kirby N. Sigler, Kaitlin S. Bain, Emily R. O'Brien, and Joanne V. Wood. Self-esteem's impacts on intimacy-building: Pathways through self-disclosure and responsiveness. *Current Opinion in Psychology*, 52:101596, aug 2023.

[12] Nemec Zlatolas, Welzer, Hölbl, Heričko, and Kamišalić. A model of perception of privacy, trust, and self-disclosure on online social networks. *Entropy*, 21(8):772, aug 2019.

[13] Mahamadou Kante, Joel Christian Adepo, and Michel Babri. Towards a model for self-disclosure on social network sites. In *Advances in Healthcare Information Systems and Administration*, pages 229–254. IGI Global, 2022.

[14] Francesca Mosca and Jose M. Such. Elvira: An explainable agent for value and utility-driven multiuser privacy. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '21, page 916–924, 2021.

[15] P. Jayaprabha, K. Paulose Jacob, and K. Preetha Mathew. Fuzzy-based multiparty privacy management in social media using modified elliptic curve cryptography. *Soft Computing*, 25(8):6083–6100, feb 2021.

[16] Onuralp Ulusoy and Pinar Yolum. PANOLA: A personal assistant for supporting users in preserving privacy. *ACM Transactions on Internet Technology*, 22(1):1–32, sep 2021.

[17] Hongqin Lyu, Yongxiong Zhang, Chao Wang, Shigong Long, and Shengnan Guo. Federated learning privacy incentives: Reverse auctions and negotiations. *CAAI Transactions on Intelligence Technology*, feb 2023.

[18] Zhenni Feng, Sijia Yu, and Yanmin Zhu. Towards personalized privacy preference aware data trading: A contract theory based approach. *Computer Networks*, 224:109637, apr 2023.

[19] Yuanping JIANG, Chengming JIANG, Tianyi HU, and Hongyue SUN. Effects of emotion on intertemporal decision-making: Explanation from the single dimension priority model. *Acta Psychologica Sinica*, 54(2):122–140, feb 2022.

[20] Itiel E. Dror. Cognitive and human factors in expert decision making: Six fallacies and the eight sources of bias. *Analytical Chemistry*, 92(12):7998–8004, jun 2020.

[21] Benjamin Enke, Uri Gneezy, Brian Hall, David Martin, Vadim Nelidov, Theo Offerman, and Jeroen van de Ven. Cognitive biases: Mistakes or missing stakes? *Review of Economics and Statistics*, 105(4):818–832, jul 2023.

[22] Ari Ezra Waldman. Cognitive biases, dark patterns, and the 'privacy paradox'. *Current Opinion in Psychology*, 31:105–109, feb 2020.

[23] Mohammad Selim. Anchoring bias in corporate decision making and its effects on net income. In *2021 International Conference on Decision Aid Sciences and Application (DASA)*. IEEE, dec 2021.

[24] Florian Schaub, Bastian Konings, Michael Weber, and Frank Kargl. Towards context adaptive privacy decisions in ubiquitous computing. In *2012 IEEE International Conference on Pervasive Computing and Communications Workshops*. IEEE, mar 2012.

[25] Florian Schaub, Bastian Könings, and Michael Weber. Context-adaptive privacy: Leveraging context awareness to support privacy decision making. *IEEE Pervasive Computing*, 14(1):34–43, jan 2015.

[26] Eugen Dimant, Gerben A. van Kleef, and Shaul Shalvi. Requiem for a nudge: Framing effects in nudging honesty. *Journal of Economic Behavior &amp Organization*, 172:247–266, apr 2020.

[27] Daphne Chang, Roy Chen, and Erin Krupka. Rhetoric matters: A social norms explanation for the anomaly of framing. *Games and Economic Behavior*, 116:158–178, jul 2019.

[28] Hiroaki Masaki, Kengo Shibata, Shui Hoshino, Takahiro Ishihama, Nagayuki Saito, and Koji Yatani. Exploring nudge designs to help adolescent SNS users avoid privacy and safety threats. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, apr 2020.

[29] Colleges Rescinding Admissions Offers as Racist Social Media Posts Emerge (Published 2020) — nytimes.com. `https://www.nytimes.com/2020/07/02/us/racism-social-media-college-admissions.html`. [Accessed 12-Jun-2023].

[30] I got fired for sharing my salary on TikTok — and cried for days straight — nypost.com. `https://nypost.com/2022/07/21/i-got-fired-for-sharing-my-salary-on-tiktok-and-cried-for-days-straight/`. [Accessed 12-Jun-2023].

[31] Tiktok's turning user-submitted content into ads, without user knowledge | social media today. `https://www.socialmediatoday.com/news/tiktoks-turning-user-submitted-content-into-ads-without-user-knowledge/564518/`. (Accessed on 06/13/2023).

[32] Mohsen Minaei, S Chandra Mouli, Mainack Mondal, Bruno Ribeiro, and Aniket Kate. Deceptive deletions for protecting withdrawn posts on social media platforms. In *Proceedings 2021 Network and Distributed System Security Symposium*. Internet Society, 2021.

[33] Privacy (Stanford Encyclopedia of Philosophy) — plato.stanford.edu. https://plato.stanford.edu/entries/privacy/. [Accessed 12-Jun-2023].

[34] George Orwell. *1984*. Tandem Library, centennial. edition, 1949.

[35] Paul Quinn and Gianclaudio Malgieri. The difficulty of defining sensitive data – the concept of sensitive data in the eu data protection framework. *German Law Journal*, March 2021.

[36] IBM. 2023 ibm cost of a data breach report – canadian businesses are being hit hard. https://canada.newsroom.ibm.com/2023-IBM-Cost-of-a-Data-Breach-Report-Canadian-businesses-are-being-hit-hard#:~:text=Canadian%20companies%20are%20still%20paying,see%20the%20highest%20breach%20costs., jul 2023. (Accessed on 08/20/2023).

[37] Qinqi Lin, Lingjie Duan, and Jianwei Huang. Personalized pricing through user profiling in social networks. In *2021 19th International Symposium on Modeling and Optimization in Mobile, Ad hoc, and Wireless Networks (WiOpt)*. IEEE, October 2021.

[38] TikTok faces growing national security concerns: "It's not just the collection or theft of that data" — cbsnews.com. https://www.cbsnews.com/news/tiktok-security-concerns-china-data/. [Accessed 12-Jun-2023].

[39] Emily Baker-White. TikTok Parent ByteDance Planned To Use TikTok To Monitor The Physical Location Of Specific American Citizens — forbes.com. https://www.forbes.com/sites/emilybaker-white/2022/10/20/tiktok-bytedance-surveillance-american-user-data/?sh=38299b206c2d. [Accessed 12-Jun-2023].

[40] The privacy piece: Report on privacy competencies in digital literacy programs in canada, britain, australia, america, and brazil. https://www.priv.gc.ca/media/1740/hamel_201111_e.pdf. (Accessed on 06/11/2023).

[41] Digital literacy - province of british columbia. https://www2.gov.bc.ca/gov/content/education-training/k-12/teach/resources-for-teachers/digital-literacy#:~:text=The%20Ministry%20of%20Education%20and,create%20and%20communicate%20with%20others%E2%80%9D. (Accessed on 06/11/2023).

[42] The Privacy Act in brief - Office of the Privacy Commissioner of Canada — priv.gc.ca. https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-privacy-act/pa_brief/. [Accessed 12-Jun-2023].

[43] What does the General Data Protection Regulation (GDPR) govern? — commission.europa.eu. https://commission.europa.eu/law/law-topic/data-protection/reform/what-does-general-data-protection-regulation-gdpr-govern_en#:~:text=Regulation%20(EU)%202016%2F679,to%20individuals%20in%20the%20EU. [Accessed 12-Jun-2023].

[44] Yang Wang, Gregory Norcie, Saranga Komanduri, Alessandro Acquisti, Pedro Giovanni Leon, and Lorrie Faith Cranor. "i regretted the minute i pressed share". In *Proceedings of the Seventh Symposium on Usable Privacy and Security*. ACM, July 2011.

[45] Sabine Trepte, Michael Scharkow, and Tobias Dienlin. The privacy calculus contextualized: The influence of affordances. *Computers in Human Behavior*, 104:106115, March 2020.

[46] Fernanda Polli Leite, Nicolas Pontes, and Paulo de Paula Baptista. Oops, i've overshared! when social media influencers' self-disclosure damage perceptions of source credibility. *Computers in Human Behavior*, 133:107274, August 2022.

[47] Dana Aizenkot. Social networking and online self-disclosure as predictors of cyberbullying victimization among children and youth. *Children and Youth Services Review*, 119:105695, December 2020.

[48] Tommy K.H. Chan, Christy M.K. Cheung, and Zach W.Y. Lee. Cyberbullying on social networking sites: A literature review and future research directions. *Information &amp Management*, 58(2):103411, March 2021.

[49] Nandita Vijayakumar and Jennifer H Pfeifer. Self-disclosure during adolescence: exploring the means, targets, and types of personal exchanges. *Current Opinion in Psychology*, 31:135–140, February 2020.

[50] Jenny Kennedy. Oversharing is the norm. In *Palgrave Studies in Communication for Social Change*, pages 265–280. Springer International Publishing, 2018.

[51] Kambiz Ghazinour and John Ponchak. Hidden privacy risks in sharing pictures on social media. *Procedia Computer Science*, 113:267–272, 2017.

[52] Tiktok's takeover of marketing and commerce in 2022. `https://www.forbes.com/sites/forbescommunicationscouncil/2022/02/23/tiktoks-takeover-of-marketing-and-commerce-in-2022/?sh=2f182e435b4f`. (Accessed on 06/11/2023).

[53] Kathryn Lundstorm. Nearly half of tiktokers purchase from brands on the app. `https://www.adweek.com/brand-marketing/nearly-half-of-tiktokers-are-buying-stuff-from-brands-they-see-on-the-platform/`, may 2021. (Accessed on 08/20/2023).

[54] Web photo geotags can reveal more than you wish - the new york times. `https://www.nytimes.com/2010/08/12/technology/personaltech/12basics.html`. (Accessed on 06/13/2023).

[55] Deepfakes: Why your instagram photos, video could be vulnerable. `https://www.usatoday.com/story/tech/2019/05/13/deepfakes-why-your-instagram-photos-video-could-be-vulnerable/3344536002/`. (Accessed on 06/13/2023).

[56] Jasmine E. McNealy. Platforms as phish farms: Deceptive social engineering at scale. *New Media &amp Society*, 24(7):1677–1694, July 2022.

[57] Robert S. Laufer, Harold M. Proshansky, and Malcolm Wolfe. Some analytic dimensions of privacy. 1973.

[58] Mary J. Culnan and Pamela K. Armstrong. Information privacy concerns, procedural fairness, and impersonal trust: An empirical investigation. *Organization Science*, 10(1):104–115, February 1999.

[59] Robert S. Laufer and Maxine Wolfe. Privacy as a concept and a social issue: A multidimensional developmental theory. *Journal of Social Issues*, 33(3):22–42, July 1977.

[60] Mary J. Culnan and Robert J. Bies. Consumer privacy: Balancing economic and justice considerations. *Journal of Social Issues*, 59(2):323–342, April 2003.

[61] Heng Xu, Tamara Dinev, H. Jeff Smith, and Paul J. Hart. Examining the formation of individual's privacy concerns: Toward an integrative view. In *International Conference on Interaction Sciences*, 2008.

[62] Fred D. Davis. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3):319, September 1989.

[63] Ji-Won Moon and Young-Gul Kim. Extending the TAM for a world-wide-web context. *Information &amp Management*, 38(4):217–230, February 2001.

[64] Nazanin Andalibi. Disclosure, privacy, and stigma on social media. *ACM Transactions on Computer-Human Interaction*, 27(3):1–43, May 2020.

[65] Yu-Hao Lee and Chien Wen Yuan. The privacy calculus of "friending" across multiple social media platforms. *Social Media Society*, 6(2):205630512092847, April 2020.

[66] Alessandro Acquisti. Privacy in electronic commerce and the economics of immediate gratification. In *Proceedings of the 5th ACM conference on Electronic commerce*. ACM, May 2004.

[67] Kirk Plangger and Matteo Montecchi. Thinking beyond privacy calculus: Investigating reactions to customer surveillance. *Journal of Interactive Marketing*, 50(1):32–44, May 2020.

[68] Teresa Fernandes and Nuno Pereira. Revisiting the privacy calculus: Why are consumers (really) willing to disclose personal data online? *Telematics and Informatics*, 65:101717, December 2021.

[69] 2023 state of the phish report - phishing stats & trends | proofpoint us. `https://www.proofpoint.com/us/resources/threat-reports/state-of-phish`. (Accessed on 06/12/2023).

[70] John W. Thibaut and Harold H. Kelley. *The Social Psychology of Groups*. Routledge, September 2017.

[71] Sarah Spiekermann, Jens Grossklags, and Bettina Berendt. E-privacy in 2nd generation e-commerce. In *Proceedings of the 3rd ACM conference on Electronic Commerce*. ACM, October 2001.

[72] David W. Wilson and Joseph S. Valacich. Unpacking the privacy paradox: Irrational decision-making within the privacy calculus. In *International Conference on Interaction Sciences*, 2012.

[73] Susan B. Barnes. A privacy paradox: Social networking in the united states. *First Monday*, September 2006.

[74] Mike Z. Yao, Ronald E. Rice, and Kier Wallis. Predicting user concerns about online privacy. *Journal of the American Society for Information Science and Technology*, 58(5):710–722, 2007.

[75] Mariea Grubbs Hoy and George Milne. Gender differences in privacy-related measures for young adult facebook users. *Journal of Interactive Advertising*, 10(2):28–45, March 2010.

[76] Ralph Gross and Alessandro Acquisti. Information revelation and privacy in online social networks. In *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*. ACM, November 2005.

[77] Susanne Barth, Menno D.T. de Jong, Marianne Junger, Pieter H. Hartel, and Janina C. Roppelt. Putting the privacy paradox to the test: Online privacy and security behaviors among users with technical knowledge, privacy awareness, and financial resources. *Telematics and Informatics*, 41:55–69, August 2019.

[78] Frederic Stutzman, Jessica Vitak, Nicole Ellison, Rebecca Gray, and Cliff Lampe. Privacy in interaction: Exploring disclosure and social capital in facebook. *Proceedings of the International AAAI Conference on Web and Social Media*, 6(1):330–337, August 2021.

[79] Norshidah Mohamed and Ili Hawa Ahmad. Information privacy concerns, antecedents and privacy measure use in social networking sites: Evidence from malaysia. *Computers in Human Behavior*, 28(6):2366–2375, November 2012.

[80] Hanna Krasnova, Sarah Spiekermann, Ksenia Koroleva, and Thomas Hildebrand. Online social networks: Why we disclose. *Journal of Information Technology*, 25(2):109–125, June 2010.

[81] Daniel J. Solove. The myth of the privacy paradox. *SSRN Electronic Journal*, 2020.

[82] Tobias Dienlin and Sabine Trepte. Is the privacy paradox a relic of the past? an in-depth analysis of privacy attitudes and privacy behaviors. *European Journal of Social Psychology*, 45(3):285–297, July 2014.

[83] Lemi Baruh, Ekin Secinti, and Zeynep Cemalcilar. Online privacy concerns and privacy management: A meta-analytical review. *Journal of Communication*, 67(1):26–53, January 2017.

[84] Jialin Fu, Xihang Li, Xi Zhao, Keyi Zhang, and Nan Cui. How does the implicit awareness of consumers influence the effectiveness of public service announcements? a functional near-infrared spectroscopy study. *Frontiers in Psychology*, 13, March 2022.

[85] bkg_technology_en.pdf. `https://cba.ca/Assets/CBA/Documents/Files/Article%20Category/PDF/bkg_technology_en.pdf`. (Accessed on 06/12/2023).

[86] 2020-21 survey of canadians on privacy-related issues - office of the privacy commissioner of canada. https://www.priv.gc.ca/en/opc-actions-and-decisions/research/explore-privacy-research/2021/por_2020-21_ca/. (Accessed on 06/12/2023).

[87] Catherine L. Anderson and Ritu Agarwal. The digitization of healthcare: Boundary risks, emotion, and consumer willingness to disclose personal health information. *Information Systems Research*, 22(3):469–490, September 2011.

[88] Luisa Pumplun, Amina Wagner, Christian Olt, Anne Zöll, and Peter Buxmann. Acting egoistically in a crisis: How emotions shape data donations. In *Proceedings of the Annual Hawaii International Conference on System Sciences*. Hawaii International Conference on System Sciences, 2022.

[89] Mihir Mehta, Sourya Joyee De, and Manojit Chattopadhyay. Elucidating the role of emotion in privacy-concerns: A text-convolutional neural network (text-CNN)-based tweets analysis of contact tracing apps. *Australasian Journal of Information Systems*, 26, September 2022.

[90] Did cambridge analytica influence the brexit vote and the us election? | politics | the guardian. https://www.theguardian.com/politics/2017/mar/04/nigel-oakes-cambridge-analytica-what-role-brexit-trump. (Accessed on 06/12/2023).

[91] R. B. Zajonc. Feeling and thinking: Preferences need no inferences. *American Psychologist*, 35(2):151–175, February 1980.

[92] Antoine Bechara, Antonio R. Damasio, Hanna Damasio, and Steven W. Anderson. Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, 50(1-3):7–15, April 1994.

[93] Edmund T. Rolls. Précis of ithe brain and emotion/i. *Behavioral and Brain Sciences*, 23(2):177–191, April 2000.

[94] Daniel Kahneman. A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, 58(9):697–720, 2003.

[95] Julie C. Stout, William C. Rodawalt, and Eric R. Siemers. Risky decision making in huntington's disease. *Journal of the International Neuropsychological Society*, 7(1):92–101, January 2001.

[96] Nathalie Camille, Giorgio Coricelli, Jerome Sallet, Pascale Pradat-Diehl, Jean-Rene Duhamel, and Angela Sirigu. The involvement of the orbitofrontal cortex in the experience of regret. *Science*, 304(5674):1167–1170, May 2004.

[97] Alessandro Acquisti, Leslie K. John, and George Loewenstein. What is privacy worth? *The Journal of Legal Studies*, 42(2):249–274, June 2013.

[98] David J. Freeman, Hanh T. Tong, and Lanny Zrill. Default-Setting and Default Bias: Does the Choice Architect Matter? (dp21-08), August 2021.

[99] Robert Cialdini and Brad Sagarin. Principles of interpersonal influence. *Persuasion: Psychological Insights and Perspectives*, pages 143–169, January 2005.

[100] Shipra Gupta and James W. Gentry. 'should i buy, hoard, or hide?'- consumers' responses to perceived scarcity. *The International Review of Retail, Distribution and Consumer Research*, 29(2):178–197, feb 2019.

[101] Is this job posting real? how to avoid falling for a scam. - the washington post. https://www.washingtonpost.com/technology/2022/12/22/job-posting-scam-tips/. (Accessed on 06/13/2023).

[102] Scammers targeting job seekers as layoffs mount - cnet. https://www.cnet.com/tech/services-and-software/fake-online-recruiters-looking-to-scam-job-seekers/. (Accessed on 06/13/2023).

[103] Zick Rubin. Disclosing oneself to a stranger: Reciprocity and its limits. *Journal of Experimental Social Psychology*, 11(3):233–260, jan 1975.

[104] Dalmas A. Taylor. The development of interpersonal relationships: Social penetration processes. *The Journal of Social Psychology*, 75(1):79–90, jun 1968.

[105] Taking action against ad fraud | meta. https://about.fb.com/news/2019/12/taking-action-against-ad-fraud/. (Accessed on 06/13/2023).

[106] Zareef A. Mohammed and Gurvirender P. Tejay. Examining the privacy paradox through individuals' neural disposition in e-commerce: An exploratory neuroimaging study. *Computers &amp Security*, 104:102201, may 2021.

[107] D. Ariely, G. Loewenstein, and D. Prelec. "coherent arbitrariness": Stable demand curves without stable preferences. *The Quarterly Journal of Economics*, 118(1):73–106, feb 2003.

[108] Daphne Chang, Erin L. Krupka, Eytan Adar, and Alessandro Acquisti. Engineering information disclosure. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, may 2016.

[109] Daniel J. Solove. Why privacy matters even if you have 'nothing to hide'. https://www.chronicle.com/article/why-privacy-matters-even-if-you-have-nothing-to-hide/, may 2011. (Accessed on 06/13/2023).

[110] Daniel J. Solove. 'i've got nothing to hide' and other misunderstandings of privacy. *San Diego Law Review*, 44:745, 2007.

[111] Joseph Farrell. Can privacy be just another good? *J. Telecommun. High Technol. Law*, 10:251–264, 2012.

[112] Fithra Faisal Hastiadi, Askar Muhammad, and Jordan Brahmansyah. Economic consequences of digital transformation. *Accelerating Digital Transformation in Indonesia: Technology, Market and Policy*, pages 77–100, 2023.

[113] Charles I. Jones and Christopher Tonetti. Nonrivalry and the economics of data. *American Economic Review*, 110(9):2819–2858, sep 2020.

[114] Annalee Newitz. Man receives the creepiest letter ever from officemax. https://gizmodo.com/man-receives-the-creepiest-letter-ever-from-officemax-1507020325, january 2014. (Accessed on 06/13/2023).

[115] Department of Justice Office of Public Affairs. Marketing company agrees to pay $150 million for facilitating elder fraud schemes | opa | department of justice. https://www.justice.gov/opa/pr/marketing-company-agrees-pay-150-million-facilitating-elder-fraud-schemes, january 2021. (Accessed on 06/13/2023).

[116] Jon Keegan and Alfred Ng. The popular family safety app life360 is selling precise location data on its tens of millions of users – the markup. https://themarkup.org/privacy/2021/12/06/the-popular-family-safety-app-life360-is-selling-precise-location-data-on-its-tens-of-millions-of-user, december 2021. (Accessed on 06/13/2023).

[117] Luc Rocher, Julien M. Hendrickx, and Yves-Alexandre de Montjoye. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, 10(1), jul 2019.

[118] Arnaud J. Tournier and Yves-Alexandre de Montjoye. Expanding the attack surface: Robust profiling attacks threaten the privacy of sparse behavioral data. *Science Advances*, 8(33), aug 2022.

[119] CBS News. "cbs this morning" investigation: Whistleblower reveals information on companies buying and selling your location data. https://www.cbsnews.com/news/location-tracking-whistleblower-reveals-info-on-companies-buying-and-selling-your-location-data/, april 2019. (Accessed on 06/13/2023).

[120] Johana Bhuiyan. Where does your info go? us lawsuit gives peek into shadowy world of data brokers | technology | the guardian. https://www.theguardian.com/technology/2022/mar/23/data-brokers-lawsuit-security-transparency, mar 2022. (Accessed on 06/13/2023).

[121] Forbes staff. 120 million american households exposed in 'massive' consumerview database leak. https://www.forbes.com/sites/thomasbrewster/2017/12/19/120m-american-households-exposed-in-massive-consumerview-database-leak/?sh=b0627f47961b, dec 2017. (Accessed on 06/13/2023).

[122] Kevin Collier. A researcher tried to buy mental health data. it was surprisingly easy. https://www.nbcnews.com/tech/security/researcher-tried-buy-mental-health-data-was-surprisingly-easy-rcna70071, feb 2023. (Accessed on 06/13/2023).

[123] Richard A. Posner. The economics of privacy. *Papers and Proceedings of the Ninety-Third Annual Meeting of the American Economic Association*, 71(2):405–409, February 1981.

[124] Stephan Grynwajc. Privacy at the crossroads: A comparative analysis of regulation in the u.s., the eu and canada. https://www.transatlantic-lawyer.com/privacy-laws-focus-on-a-transatlantic-perspective/, dec 2020. (Accessed on 06/13/2023).

[125] Thorin Klosowski. The state of consumer data privacy laws in the us (and why it matters) | wirecutter. https://www.nytimes.com/wirecutter/blog/state-of-privacy-laws-in-us/, sep 2021. (Accessed on 06/13/2023).

[126] Federal Trade Commission. Data brokers: A call for transparency and accountability: A report of the federal trade commission (may 2014). https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf, may 2014. (Accessed on 06/13/2023).

[127] Brian Fung. Republicans voted to roll back landmark fcc privacy rules. here's what you need to know. - the washington post. https://www.washingtonpost.com/news/the-switch/wp/2017/03/28/republicans-are-poised-to-roll-back-landmark-fcc-privacy-rules-heres-what-you-need-to-know/, mar 2017. (Accessed on 06/13/2023).

[128] Vivek Krishnamurthy. A tale of two privacy laws: The GDPR and the international right to privacy. *AJIL Unbound*, 114:26–30, 2020.

[129] Office of the Privacy Commissioner of Canada. Summary of privacy laws in canada - office of the privacy commissioner of canada. https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/02_05_d_15/#heading-0-0-2-2-2, jan 2018. (Accessed on 06/13/2023).

[130] European Commission. Who does the data protection law apply to? https://commission.europa.eu/law/law-topic/data-protection/reform/rules-business-and-organisations/application-regulation/who-does-data-protection-law-apply_en. (Accessed on 06/13/2023).

[131] Adeola Adegunwa. Analysis of top 10 countries mostly targeted by data breaches. https://informationsecuritybuzz.com/analysis-top-ten-countries-mostly-targeted-data-breaches/, dec 2022. (Accessed on 06/13/2023).

[132] Michael Arrington. Aol proudly releases massive amounts of private data | techcrunch. https://techcrunch.com/2006/08/06/aol-proudly-releases-massive-amounts-of-user-search-data/, aug 2006. (Accessed on 06/13/2023).

[133] Michael Barbaro and Tom Zeller Jr. A face is exposed for aol searcher no. 4417749 - the new york times. https://www.nytimes.com/2006/08/09/technology/09aol.html, aug 2006. (Accessed on 06/13/2023).

[134] Celestin Matte, Nataliia Bielova, and Cristiana Santos. Do cookie banners respect my choice? : Measuring legal compliance of banners from IAB europe's transparency and consent framework. In *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, may 2020.

[135] Iskander Sanchez-Rola, Matteo Dell'Amico, Platon Kotzias, Davide Balzarotti, Leyla Bilge, Pierre-Antoine Vervier, and Igor Santos. Can i opt out yet? In *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security*. ACM, jul 2019.

[136] Alfred Ng and Maddy Varner. The little-known data broker industry is spending big bucks lobbying congress – the markup. `https://themarkup.org/privacy/2021/04/01/the-little-known-data-broker-industry-is-spending-big-bucks-lobbying-congress`, apr 2021. (Accessed on 06/14/2023).

[137] Justin Sherman. Data brokers know where you are—and want to sell that intel | wired. `https://www.wired.com/story/opinion-data-brokers-know-where-you-are-and-want-to-sell-that-intel/#:~:text=They%20openly%20and%20explicitly%20promulgate,US%20government%20and%20military%20personnel.`, aug 2021. (Accessed on 06/14/2023).

[138] Karen A. Scarfone Erika McCallister, Timothy Grance. Guide to protecting the confidentiality of personally identifiable information (pii). *National Institute of Standards and Technology*, 2010.

[139] General Data Protection Regulation GDPR. General data protection regulation (gdpr) – official legal text. `https://gdpr-info.eu/`. (Accessed on 06/14/2023).

[140] Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, Damon McCoy, Sarah Meiklejohn, Thomas Ristenpart, and Gianluca Stringhini. SoK: Hate, harassment, and the changing landscape of online abuse. In *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, may 2021.

[141] Tanja Artiga Gonzalez, Francesco Capozza, and Georg Granic. Political support, cognitive dissonance and political preferences. *SSRN Electronic Journal*, 2022.

[142] Wiebke Mohr, Anika Rädke, Bernhard Michalowsky, and Wolfgang Hoffmann. Elicitation of quantitative, choice-based preferences for person-centered care among people living with dementia in comparison to physicians' judgements in germany: study protocol for the mixed-methods PreDemCare-study. *BMC Geriatrics*, 22(1), jul 2022.

[143] Nathan N. O'Hara. Eliciting health care preferences with discrete choice experiments. *JAMA Network Open*, 5(4):e228794, apr 2022.

[144] Stacia M. Garlach. *Usability of advertising preference tools on smartphones: AdChoices and Facebook Ad Preferences*. PhD thesis, University of Hawaii.

[145] Robert Sugden. Debiasing or regularisation? two interpretations of the concept of 'true preference' in behavioural economics. *Theory and Decision*, 92(3-4):765–784, feb 2022.

[146] Dario Bonaretti, Marcin Bartosiak, Tsz-Wai Lui, Gabriele Piccoli, and Daniele Marchesani. "what can i(s) do for you?": How technology enables service providers to elicit customers' preferences and deliver personalized service. *Information &amp Management*, 57(6):103346, sep 2020.

[147] Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. Ethical aspects of multi-stakeholder recommendation systems. *The Information Society*, 37(1):35–45, oct 2020.

[148] Terri R. Fried, Mary Tinetti, Joe Agostini, Lynne Iannone, and Virginia Towle. Health outcome prioritization to elicit preferences of older persons with multiple health conditions. *Patient Education and Counseling*, 83(2):278–282, may 2011.

[149] Tristan Allard, Tassadit Bouadi, Joris Duguépéroux, and Virginie Sans. From self-data to self-preferences: Towards preference elicitation in personal information management systems. In *Personal*

Analytics and Privacy. An Individual and Collective Perspective, pages 10–16. Springer International Publishing, 2017.

[150] Esma Aïmeur, Gilles Brassard, José M. Fernandez, and Flavien Serge Mani Onana. Alambic: a privacy-preserving recommender system for electronic commerce. *International Journal of Information Security*, 7(5):307–334, February 2008.

[151] Eric P. Kroes and Robert J. Sheldon. Stated preference methods: An introduction. *Stated Preference Methods in Transport Research*, 22(1):11–25, 1988.

[152] Erin Carbone and George F. Loewenstein. Dying to divulge: The determinants of, and relationship between, desired and actual disclosure. *SSRN Electronic Journal*, 2020.

[153] Petr Mariel, David Hoyos, Jürgen Meyerhoff, Mikolaj Czajkowski, Thijs Dekker, Klaus Glenk, Jette Bredahl Jacobsen, Ulf Liebe, Søren Bøye Olsen, Julian Sagebiel, and Mara Thiene. *Environmental Valuation with Discrete Choice Experiments*. Springer International Publishing, 2021.

[154] Alan F. Westin. *Privacy and Freedom*. Bodley Head, 1970.

[155] P. A. Samuelson. A note on the pure theory of consumer's behaviour. *Economica*, 5(17):61, feb 1938.

[156] Humphrey Taylor. Most people are "privacy pragmatists" who, while concerned about privacy, will sometimes trade it off for other benefits. 01 2003.

[157] Tiago Bianchi. Global mobile traffic 2022 | statista. `https://www.statista.com/statistics/277125/share-of-website-traffic-coming-from-mobile-devices/`, apr 2023. (Accessed on 06/14/2023).

[158] Yaacov Trope and Nira Liberman. Temporal construal and time-dependent changes in preference. *Journal of Personality and Social Psychology*, 79(6):876–889, 2000.

[159] Kostas Stefanidis and Evaggelia Pitoura. Fast contextual preference scoring of database tuples. In *Proceedings of the 11th international conference on Extending database technology: Advances in database technology*. ACM, mar 2008.

[160] Kostas Stefanidis, Evaggelia Pitoura, and Panos Vassiliadis. Managing contextual preferences. *Information Systems*, 36(8):1158–1180, dec 2011.

[161] Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.

[162] Ping Xu, Benjamin T. Vincent, Hui Sang, and Xiaodong Li. Examining the role of risk in waiting preference and dynamic preference reversal: An experience intertemporal choice study. *Journal of Behavioral Decision Making*, 35(2), jul 2021.

[163] Kumar Sharad and George Danezis. An automated social graph de-anonymization technique. In *Proceedings of the 13th Workshop on Privacy in the Electronic Society*. ACM, nov 2014.

[164] Ildar Nurgaliev, Qiang Qu, Seyed Mojtaba Hosseini Bamakan, and Muhammad Muzammal. Matching user identities across social networks with limited profile data. *Frontiers of Computer Science*, 14(6), apr 2020.

[165] Kamran Shaukat, Ibrahim A Hameed, Suhuai Luo, Imran Javed, Farhat Iqbal, Amber Faisal, Rabia Masood, Ayesha Usman, Usman Shaukat, Rosheen Hassan, Aliya Younas, Shamshair Ali, and Ghazif Adeem. Domain specific lexicon generation through sentiment analysis. *International Journal of Emerging Technologies in Learning (iJET)*, 15(09):190, may 2020.

[166] Ahmad Hassanpour and Bian Yang. PriMe: A novel privacy measuring framework for online social networks. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, nov 2022.

[167] Awanthika Senarath, Marthie Grobler, and Nalin Arachchilage. A model for system developers to measure the privacy risk of data. In *Proceedings of the Annual Hawaii International Conference on System Sciences*. Hawaii International Conference on System Sciences, 2019.

[168] Xuefeng Li, Yang Xin, Chensu Zhao, Yixian Yang, Shoushan Luo, and Yuling Chen. Using user behavior to measure privacy on online social networks. *IEEE Access*, 8:108387–108401, 2020.

[169] Samia Oukemeni, Helena Rifa-Pous, and Joan Manuel Marques Puig. IPAM: Information privacy assessment metric in microblogging online social networks. *IEEE Access*, 7:114817–114836, 2019.

[170] Shitong Fu and Zhiqiang Yao. Privacy risk estimation of online social networks. In *2022 International Conference on Networking and Network Applications (NaNA)*. IEEE, dec 2022.

[171] Lorrie Faith Cranor. Platform for privacy preferences (p3p). In *Encyclopedia of Cryptography and Security*, pages 940–941. Springer US, 2011.

[172] Tanmoy Chakraborty, Sushil Jajodia, Jonathan Katz, Antonio Picariello, Giancarlo Sperli, and V. S. Subrahmanian. A fake online repository generation engine for cyber deception. *IEEE Transactions on Dependable and Secure Computing*, 18(2):518–533, mar 2021.

[173] Gretel Egan. 2019 state of the phish report: Attack rates rise, account compromise soars. `https://www.proofpoint.com/us/corporate-blog/post/2019-state-phish-report-attack-rates-rise-account-compromise-soars`, jan 2019. (Accessed on 06/18/2023).

[174] Public Safety Canada. National cyber security strategy: Canada's vision for security and prosperity in the digital age. `https://www.publicsafety.gc.ca/cnt/rsrcs/pblctns/ntnl-cbr-scrt-strtg/index-en.aspx`. (Accessed on 06/18/2023).

[175] Homeland Security. Biden-harris administration announces $1 billion in funding for first-ever state and local cybersecurity grant program | homeland security. `https://www.dhs.gov/news/2022/09/16/biden-harris-administration-announces-1-billion-funding-first-ever-state-and-local`, sep 2022. (Accessed on 06/18/2023).

[176] Polona Caserman, Katrin Hoffmann, Philipp Müller, Marcel Schaub, Katharina Straßburg, Josef Wiemeyer, Regina Bruder, and Stefan Göbel. Quality criteria for serious games: Serious part, game part, and balance. *JMIR Serious Games*, 8(3):e19037, jul 2020.

[177] Mohib Ullah, Sareer Ul Amin, Muhammad Munsif, Utkurbek Safaev, Habib Khan, Salman Khan, and Habib Ullah. Serious games in science education. a systematic literature review. *Virtual Reality &amp Intelligent Hardware*, 4(3):189–209, jun 2022.

[178] Ari Min, Haeyoung Min, and Sujeong Kim. Effectiveness of serious games in nurse education: A systematic review. *Nurse Education Today*, 108:105178, jan 2022.

[179] George Papanastasiou, Athanasios Drigas, and Charalabos Skianis. Serious games in pre-k and k-6 education. *Technium Education and Humanities*, 2(3):1–18, sep 2022.

[180] Ifeoma Adaji. Serious games for healthy nutrition. a systematic literature review. *International Journal of Serious Games*, 9(1):3–16, mar 2022.

[181] N. Menelaos Katsantonis, Isavella Kotini, Panayotis Fouliras, and Ioannis Mavridis. Conceptual framework for developing cyber security serious games. In *2019 IEEE Global Engineering Education Conference (EDUCON)*. IEEE, apr 2019.

[182] Sara Rye and Emel Aktas. Serious games as a validation tool for PREDIS: A decision support system for disaster management. *International Journal of Environmental Research and Public Health*, 19(24):16584, dec 2022.

[183] Steven Ashley Forrest, Martina Kubíková, and Jan Macháč. Serious gaming in flood risk management. *WIREs Water*, 9(4), apr 2022.

[184] Clark C. Abt. *Serious Games*. University Press of America, 1987.

[185] Daniel J. Solove. Online privacy training i teachprivacy. `https://teachprivacy.com/`. (Accessed on 06/18/2023).

[186] Charu Singh and Meenu. Phishing website detection based on machine learning: A survey. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*. IEEE, mar 2020.

[187] Gaurav Misra and Jose M. Such. PACMAN: Personal agent for access control in social media. *IEEE Internet Computing*, 21(6):18–26, nov 2017.

[188] Ryan Wishart, Domenico Corapi, Srdjan Marinovic, and Morris Sloman. Collaborative privacy policy authoring in a social networking context. In *2010 IEEE International Symposium on Policies for Distributed Systems and Networks*. IEEE, 2010.

[189] Branka. Facebook statistics 2023 - truelist. `https://truelist.co/blog/facebook-statistics/`, jan 2023. (Accessed on 06/18/2023).

[190] Katia Sycara. Resolving goal conflicts via negotiation. In *Proceedings of the Seventh AAAI National Conference on Artificial Intelligence*, August 1988.

[191] Jeff Kelley. An empirical methodology for writing user-friendly natural language computer applications. In *Proceedings of the CHI' 83 Conference on Human Factors in Computing Systems*, pages 193–196, August 1983.

[192] Jeff Kelley. An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Office Information Systems*, 2(1):26–41, mar 1984.

[193] Dorota Filipczuk, Tim Baarslag, Enrico H. Gerding, and m. c. schraefel. Automated privacy negotiations with preference uncertainty. *Autonomous Agents and Multi-Agent Systems*, 36(2), aug 2022.

[194] Andrew Besmer and Heather Richter Lipford. Moving beyond untagging. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, apr 2010.

[195] Nadin Kökciyan, Nefise Yaglikci, and Pinar Yolum. An argumentation approach for resolving privacy disputes in online social networks. *ACM Transactions on Internet Technology*, 17(3):1–22, jun 2017.

[196] Ron S. Hirschprung and Shani Alkoby. A game theory approach for assisting humans in online information-sharing. *Information*, 13(4):183, apr 2022.

[197] Richard H. Thaler, Cass R. Sunstein, and John P. Balz. Choice architecture. *The behavioral foundations of public policy*, 2013.

[198] Lucia A. Reisch, Cass R. Sunstein, and Wencke Gwozdz. Viewpoint: Beyond carrots and sticks: Europeans support health nudges. *Food Policy*, 69:1–10, may 2017.

[199] William Hagman, David Andersson, Daniel Västfjäll, and Gustav Tinghög. Public views on policies involving nudges. *Review of Philosophy and Psychology*, 6(3):439–453, may 2015.

[200] Dragos C. Petrescu, Gareth J. Hollands, Dominique-Laurent Couturier, Yin-Lam Ng, and Theresa M. Marteau. Public acceptability in the UK and USA of nudging to reduce obesity: The example of reducing sugar-sweetened beverages consumption. *PLOS ONE*, 11(6):e0155995, jun 2016.

[201] Benjamin Wellace-Wells. Cass sunstein wants to nudge us - the new york times. `https://www.nytimes.com/2010/05/16/magazine/16Sunstein-t.html`, may 2010. (Accessed on 07/08/2023).

[202] Shlomo Benartzi, John Beshears, Katherine L Milkman, Cass R Sunstein, Richard H Thaler, Maya Shankar, Will Tucker-Ray, William J Congdon, and Steven Galing. Should governments invest more in nudging? *Psychological Science*, 28(8):1041–1055, jun 2017.

[203] Deloitte. www2.deloitte.com/content/dam/insights/articles/6730_tt-landing-page/di_2021-tech-trends.pdf. `https://www2.deloitte.com/content/dam/insights/articles/6730_TT-Landing-page/DI_2021-Tech-Trends.pdf`. (Accessed on 07/08/2023).

[204] World Health Organization. Saving lives, spending less: the case for investing in noncommunicable diseases. https://www.who.int/publications/i/item/9789240041059, dec 2021. (Accessed on 07/08/2023).

[205] Loni Ledderer, Marianne Kjær, Emilie Kirstine Madsen, Jacob Busch, and Antoinette Fage-Butler. Nudging in public health lifestyle interventions: A systematic literature review and metasynthesis. *Health Education &amp Behavior*, 47(5):749–764, jun 2020.

[206] Christoph Schneider, Markus Weinmann, and Jan vom Brocke. Digital nudging. *Communications of the ACM*, 61(7):67–73, jun 2018.

[207] Luca Congiu and Ivan Moscati. A review of nudges: Definitions, justifications, effectiveness. *Journal of Economic Surveys*, 36(1):188–213, jul 2021.

[208] Valério Souza-Neto, Osiris Marques, Verônica Feder Mayer, and Gui Lohmann. Lowering the harm of tourist activities: a systematic literature review on nudges. *Journal of Sustainable Tourism*, pages 1–22, feb 2022.

[209] Seth H. Werfel. Household behaviour crowds out support for climate change policy when sufficient progress is perceived. *Nature Climate Change*, 7(7):512–515, jun 2017.

[210] Daniela Grieco, Nicola Lacetera, Mario Macis, and Daniela Di Martino. Motivating cord blood donation with information and behavioral nudges. *Scientific Reports*, 8(1), jan 2018.

[211] LUC BOVENS. Behavioural public policies and charitable giving. *Behavioural Public Policy*, 2(2):168–173, aug 2018.

[212] Mohin Banker, Moses Miller, Guy Voichek, Dafna Goor, and Tamar Makov. Prosocial nudges and visual indicators increase social distancing, but authoritative nudges do not. *Proceedings of the National Academy of Sciences*, 119(33), aug 2022.

[213] Luca Congiu and Ivan Moscati. Message and environment: a framework for nudges and choice architecture. *Behavioural Public Policy*, 4(1):71–87, aug 2018.

[214] Francesco Ricci, Lior Rokach, and Bracha Shapira. Recommender systems: Techniques, applications, and challenges. In *Recommender Systems Handbook*, pages 1–35. Springer US, nov 2021.

[215] Alessandro Acquisti. Nudging privacy: The behavioral economics of personal information. *IEEE Security &amp Privacy Magazine*, 7(6):82–85, nov 2009.

[216] Rim Ben Salem, Esma Aïmeur, and Hicham Hage. Towards enhanced privacy-preserving nudges. In *Proceedings of the 2nd Workshop on Adverse Impacts and Collateral Effects of Artificial Intelligence Technologies AIofAI @ the International Joint Conference on Artificial Intelligence IJCAI*, July 2022.

[217] James Turland, Lynne Coventry, Debora Jeske, Pam Briggs, and Aad van Moorsel. Nudging towards security. In *Proceedings of the 2015 British HCI Conference*. ACM, jul 2015.

[218] Hayatullah Amanat. ctvnews.ca/lifestyle/watch-what-you-tweet-poll-finds-most-employers-would-consider-firing-workers-for-inappropriate-social-media-posts-1.6230625. https://www.ctvnews.ca/lifestyle/watch-what-you-tweet-poll-finds-most-employers-would-consider-firing-workers-for-inappropriate-social-media-posts-1.6230625, jan 2023. (Accessed on 07/09/2023).

[219] Bin Liu, Mads Schaarup Andersen, Florian Schaub, Hazim Almuhimedi, Shikun Zhang, Norman Sadeh, Alessandro Acquisti, and Yuvraj Agarwal. Follow my recommendations: A personalized privacy assistant for mobile app permissions. In *Proceedings of the Twelfth USENIX Conference on Usable Privacy and Security*, page 27–41, USA, 2016. USENIX Association.

[220] Rebecca Walker Naylor, Cait Poynor Lamberton, and David A. Norton. Seeing ourselves in others: Reviewer ambiguity, egocentric anchoring, and persuasion. *Journal of Marketing Research*, 48(3):617–631, jun 2011.

[221] Han Shao, Xiang Li, and Guodi Wang. Are you tired? i am: Trying to understand privacy fatigue of social media users. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. ACM, apr 2022.

[222] Sushmita Khan, Mehtab Iqbal, Nushrat Humaira, Nina Hubig, and Bart Knijnenburg. Mitigating digital mindlessness. In *Proceedings of the 1st Workshop on Adverse Impacts and Collateral Effects of Artificial Intelligence Technologies AIofAI @ the International Joint Conference on Artificial Intelligence IJCAI*, August 2021.

[223] Randi Karlsen and Anders Andersen. Recommendations with a nudge. *Technologies*, 7(2):45, jun 2019.

[224] Frederic Raber, Alexander De Luca, and Moritz Graus. Privacy wedges: Area-Based audience selection for social network posts. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, Denver, CO, June 2016. USENIX Association.

[225] Víctor Botti-Cebriá, Elena del Val, and Ana García-Fornes. Automatic detection of sensitive information in educative social networks. In *13th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2020)*, pages 184–194. Springer International Publishing, aug 2020.

[226] Simon Conn. Social media impact on credit rating & mortgage applications - simon conn. `https://www.simonconn.com/blog/social-media-impact-credit-rating-overseas-mortgage-applications/`, nov 2020. (Accessed on 07/09/2023).

[227] Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, Yang Wang, and Shomir Wilson. Nudges for privacy and security. *ACM Computing Surveys*, 50(3):1–41, aug 2017.

[228] Andreas T. Schmidt and Bart Engelen. The ethics of nudging: An overview. *Philosophy Compass*, 15(4), apr 2020.

[229] Till Grüne-Yanoff. Old wine in new casks: libertarian paternalism still violates liberal principles. *Social Choice and Welfare*, 38(4):635–645, jan 2012.

[230] de Haan Thomas and Linde Jona. 'good nudge lullaby': Choice architecture and default bias reinforcement. *The Economic Journal*, 128(610):1180–1206, may 2017.

[231] Shlomo Cohen. Nudging and informed consent. *The American Journal of Bioethics*, 13(6):3–11, jun 2013.

[232] Jennifer Blumenthal-Barby. Between reason and coercion: Ethically permissible influence in health care and health policy contexts. *Kennedy Institute of Ethics journal*, 22:345–66, 12 2012.

[233] Moritz Becker, Christian Matt, and Thomas Hess. It's not just about the product. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems*, 51(1):37–50, jan 2020.

[234] Yannic Meier, Johanna Schäwel, Elias Kyewski, and Nicole C. Krämer. Applying protection motivation theory to predict facebook users' withdrawal and disclosure intentions. In *International Conference on Social Media and Society*. ACM, jul 2020.

[235] Anna Rudnicka, Anna L. Cox, and Sandy J. J. Gould. Why do you need this? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, may 2019.

[236] Karen H. Smith, Francis A. Méndez Mediavilla, and Garry L. White. The impact of online training on facebook privacy. *Journal of Computer Information Systems*, 58(3):244–252, oct 2016.

[237] Georgiana Craciun. Choice defaults and social consensus effects on online information sharing: The moderating role of regulatory focus. *Computers in Human Behavior*, 88:89–102, nov 2018.

[238] Gulsum Akkuzu, Benjamin Aziz, and Mo Adda. Towards consensus-based group decision making for co-owned data sharing in online social networks. *IEEE Access*, 8:91311–91325, 2020.

[239] A. K. M. Nuhil Mehdy, Michael D. Ekstrand, Bart P. Knijnenburg, and Hoda Mehrpouyan. Privacy as a planned behavior: Effects of situational factors on privacy perceptions and plans. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. ACM, jun 2021.

[240] Sarah Spiekermann, Dirk Temme, and Martin Strobel. Drivers and impediments of consumer online information search: Self-controlled versus agent-assisted search. In *Wirtschaftsinformatik 2005*, pages 1661–1680. Physica-Verlag HD, 2005.

[241] Mark S. Ackerman, Lorrie Faith Cranor, and Joseph Reagle. Privacy in e-commerce. In *Proceedings of the 1st ACM conference on Electronic commerce*. ACM, nov 1999.

[242] Charalambos D Aliprantis. On the backward induction method. *Economics Letters*, 64(2):125–131, aug 1999.

[243] David M. Kreps. Nash equilibrium. In *Game Theory*, pages 167–177. Palgrave Macmillan UK, 1989.

[244] R.J. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In *21st International Conference on Data Engineering (ICDE'05)*. IEEE.

[245] Shumin Guo and Keke Chen. Mining privacy settings to find optimal privacy-utility tradeoffs for social network services. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*. IEEE, sep 2012.

[246] Lei Xu, Chunxiao Jiang, Jian Wang, Yong Ren, Jian Yuan, and Mohsen Guizani. Game theoretic data privacy preservation: Equilibrium and pricing. In *2015 IEEE International Conference on Communications (ICC)*. IEEE, jun 2015.

[247] Eugenia Politou, Efthimios Alepis, Maria Virvou, and Constantinos Patsakis. The "right to be forgotten" in the GDPR: Implementation challenges and potential solutions. In *Privacy and Data Protection Challenges in the Distributed Era*, pages 41–68. Springer International Publishing, oct 2021.

[248] HuffPost Post. Woman sues her parents for posting her baby photos on facebook. `https://www.huffpost.com/entry/woman-sues-her-parents-for-posting-her-baby-photos-on-facebook_n_57dc03c3e4b04a1497b3ebdd`, sep 2016. (Accessed on 06/19/2023).

[249] Lydia Smith. Woman faces £9,000 fine if she posts pictures of her son on facebook | the independent | the independent. `https://www.independent.co.uk/news/world/europe/facebook-fines-woman-son-photos-post-social-media-court-italy-rome-a8155361.html`, jan 2018. (Accessed on 06/19/2023).

[250] Nessrine Omrani and Nicolas Soulié. Culture, Privacy Conception and Privacy Concern: Evidence from Europe before PRISM. Technical report, 2017.

[251] Richard H. Cutler. Distributed presence and community in cyberspace. *Interpersonal Computing and Technology*, 3(2):12–32, February 1995.

[252] Sophie Cockcroft and Saphira Rekker. The relationship between culture and information privacy policy. *Electronic Markets*, 26(1):55–72, jul 2015.

[253] Sabine Trepte, Leonard Reinecke, Nicole B. Ellison, Oliver Quiring, Mike Z. Yao, and Marc Ziegele. A cross-cultural perspective on the privacy calculus. *Social Media Society*, 3(1):205630511668803, jan 2017.

[254] Greg Iacurci. Covid fraud costs americans $382 million. `https://www.cnbc.com/2021/03/24/covid-fraud-costs-americans-382-million-dollars.html`, mar 2021. (Accessed on 07/10/2023).

[255] Michigan Technology Law Review. Who owns copyright in a selfie when it's captured with one person's phone but by another person's finger? and what if that other person is actually a monkey? or a bradley cooper? `https://mttlr.org/2014/09/who-owns-copyright-in-a-selfie-when-its-captured-w`

ith-one-persons-phone-but-by-another-persons-finger-and-what-if-that-other-person-is-actually-a-monkey-or-a-bradley-coope/. (Accessed on 07/10/2023).

[256] Sigal Tifferet. Gender differences in privacy tendencies on social network sites: A meta-analysis. *Computers in Human Behavior*, 93:1–12, apr 2019.

[257] Neeraj Pandey and Bhargav Gudipudi. Understanding 'what is privacy' for millennials on facebook in india. *Journal of Data Protection Privacy*, 2(3):224–233, mar 2019.

[258] Neil Zhenqiang Gong and Bin Liu. Attribute inference attacks in online social networks. *ACM Transactions on Privacy and Security*, 21(1):1–30, jan 2018.

[259] Jooyoung Lee, Sarah Michele Rajtmajer, Eesha Srivatsavaya, and Shomir Wilson. Digital inequality through the lens of self-disclosure. *Proceedings on Privacy Enhancing Technologies*, 2021:373–393, 2021.

[260] Hai Liang, Fei Shen, and King wa Fu. Privacy protection and self-disclosure across societies: A study of global twitter users. *New Media &amp Society*, 19(9):1476–1497, may 2016.

[261] Esma Aïmeur, Nicolás Díaz Ferreyra, and Hicham Hage. Manipulation and malicious personalization: Exploring the self-disclosure biases exploited by deceptive attackers on social media. *Frontiers in Artificial Intelligence*, 2, nov 2019.

[262] Jose M. Such and Natalia Criado. Multiparty privacy in social media. *Communications of the ACM*, 61(8):74–81, jul 2018.

[263] Rim Ben Salem, Esma Aïmeur, and Hicham Hage. A nudge-based recommender system towards responsible online socializing. In *Proceedings of the Workshop on Online Misinformation- and Harm-Aware Recommender Systems @ ACM Recommender Systems Conference (RecSys)*. ACM, sep 2020.

[264] Welderufael Berhane Tesfay, Jetzabel M. Serna, and Sebastian Pape. Challenges in detecting privacy revealing information in unstructured text. In *PrivOn@ISWC*, 2016.

[265] Michael Petrolini, Stefano Cagnoni, and Monica Mordonini. Automatic detection of sensitive data using transformer- based classifiers. *Future Internet*, 14(8):228, jul 2022.

[266] Huina Mao, Xin Shuai, and Apu Kapadia. Loose tweets. In *Proceedings of the 10th annual ACM workshop on Privacy in the electronic society*. ACM, oct 2011.

[267] Welderufael B. Tesfay, Jetzabel Serna, and Kai Rannenberg. PrivacyBot: Detecting privacy sensitive information in unstructured texts. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. IEEE, oct 2019.

[268] Alem Fitwi, Meng Yuan, Seyed Nikouei, and Yu Chen. Minor privacy protection by real-time children identification and face scrambling at the edge. *ICST Transactions on Security and Safety*, 7(23):164560, jun 2020.

[269] Dror Ben-Zeev, Emily A. Scherer, Rui Wang, Haiyi Xie, and Andrew T. Campbell. Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health. *Psychiatric Rehabilitation Journal*, 38(3):218–226, sep 2015.

[270] Britta Wetzel, Rüdiger Pryss, Harald Baumeister, Johanna-Sophie Edler, Ana Sofia Oliveira Gonçalves, and Caroline Cohrdes. "how come you don't call me?" smartphone communication app usage as an indicator of loneliness and social well-being across the adult lifespan during the COVID-19 pandemic. *International Journal of Environmental Research and Public Health*, 18(12):6212, jun 2021.

[271] Daniel Fulford, Jasmine Mote, Rachel Gonzalez, Samuel Abplanalp, Yuting Zhang, Jarrod Luckenbaugh, Jukka-Pekka Onnela, Carlos Busso, and David E. Gard. Smartphone sensing of social interactions in people with and without schizophrenia. *Journal of Psychiatric Research*, 137:613–620, may 2021.

[272] CBS News. Killer of tanya van cuylenborg and jay cook eludes police for 31 years – how did cece moore find him in two hours? https://www.cbsnews.com/news/tanya-van-cuylenborg-jay-cook-killer-dna-cece-moore/, jan 2023. (Accessed on 07/10/2023).

[273] Nicolás Ferreyra, Esma Aïmeur, Hicham Hage, Maritta Heisel, and Catherine van Hoogstraten. Persuasion meets AI: Ethical considerations for the design of social engineering countermeasures. In *Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. SCITEPRESS - Science and Technology Publications, 2020.

[274] Jose M. Such, Joel Porter, Sören Preibusch, and Adam Joinson. Photo privacy conflicts in social media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, may 2017.

[275] Leah Zhang-Kennedy and Sonia Chiasson. A systematic review of multimedia tools for cybersecurity awareness and education. *ACM Computing Surveys*, 54(1):1–39, jan 2021.

[276] Heidi A. McKee. Policy matters now and in the future: Net neutrality, corporate data mining, and government surveillance. *Computers and Composition*, 28(4):276–291, dec 2011.

[277] Matt Bartlett. Beyond privacy: Protecting data interests in the age of artificial intelligence. *Law, Technology and Humans*, 3(1):96–108, may 2021.

[278] Tamar Sharon and Bert-Jaap Koops. The ethics of inattention: revitalising civil inattention as a privacy-protecting mechanism in public spaces. *Ethics and Information Technology*, 23(3):331–343, jan 2021.

[279] Mufan Luo and Jeffrey T. Hancock. Self-disclosure and social media: motivations, mechanisms and psychological well-being. *Current Opinion in Psychology*, 31:110–115, feb 2020.

[280] Shiri Melumad and Robert Meyer. Full disclosure: How smartphones enhance consumer self-disclosure. *Journal of Marketing*, 84(3):28–45, 2020.

[281] Peter Story, Daniel Smullen, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. From intent to action: Nudging users towards secure mobile payments. In *Proceedings of the 16th Symposium on Usable Privacy and Security (SOUPS)*, pages 379–416, 2020.

[282] Logan Warberg, Alessandro Acquisti, and Douglas Sicker. Can privacy nudges be tailored to individuals' decision making and personality traits? In *Proceedings of the 18th ACM Workshop on Privacy in the Electronic Society*. ACM, nov 2019.

[283] Tobias Kroll and Stefan Stieglitz. Digital nudging and privacy: improving decisions about self-disclosure in social networks. *Behaviour &amp Information Technology*, 40(1):1–19, feb 2019.

[284] Biplab Chakraborty. Risk appetite. *The Management Accountant Journal*, 57(1), January 2022.

[285] V. R. Revathy and S. Pillai Anitha. Cold start problem in social recommender systems: State-of-the-art review. In *Advances in Intelligent Systems and Computing*, pages 105–115. Springer Singapore, aug 2018.

[286] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, mar 2003.

[287] Kartika Rizqi Nastiti, Ahmad Fathan Hidayatullah, and Ahmad Rafie Pratama. Discovering computer science research topic trends using latent dirichlet allocation. *Jurnal Online Informatika*, 6(1):17, jun 2021.

[288] Arho Suominen and Hannes Toivanen. Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology*, 67(10):2464–2476, sep 2015.

[289] Chyi-Kwei Yau, Alan Porter, Nils Newman, and Arho Suominen. Clustering scientific documents with topic modeling. *Scientometrics*, 100(3):767–786, may 2014.

[290] Rania Albalawi, Tet Hin Yeap, and Morad Benyoucef. Using topic modeling methods for short-text data: A comparative analysis. *Frontiers in Artificial Intelligence*, 3, jul 2020.

[291] Edi Surya Negara, Dendi Triadi, and Ria Andryani. Topic modelling twitter data with latent dirichlet allocation method. In *2019 International Conference on Electrical Engineering and Computer Science (ICECOS)*. IEEE, oct 2019.

[292] Jian Tang, Zhaoshi Meng, Xuanlong Nguyen, Qiaozhu Mei, and Ming Zhang. Understanding the limiting factors of topic modeling via posterior contraction analysis. In *2014 31st International Conference on Machine Learning*, pages 90–198, 2014.

[293] Shaheen Syed and Marco Spruit. Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, oct 2017.

[294] David Mimno, Hanna M. Wallach, Edmund M. Talley, Miriam Leenders, and Andrew Mccallum. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 262–272. DBLP, July 2011.

[295] David Newman, Youn Noh, Edmund Talley, Sarvnaz Karimi, and Timothy Baldwin. Evaluating topic models for digital libraries. In *Proceedings of the 10th annual joint conference on Digital libraries*. ACM, jun 2010.

[296] Livio Bioglio and Ruggero G. Pensa. Analysis and classification of privacy-sensitive content in social media posts. *EPJ Data Science*, 11(1), mar 2022.

[297] Henny Claresta and Daniel Tamburian. Self-disclosure of adolescent girls on TikTok social media. In *Advances in Social Science, Education and Humanities Research*. Atlantis Press, 2021.

[298] Erica R. Bailey, Sandra C. Matz, Wu Youyou, and Sheena S. Iyengar. Authentic self-expression on social media is associated with greater subjective well-being. *Nature Communications*, 11(1), oct 2020.

[299] Shanshan Zhang, Ron Chi-Wai Kwok, Paul Benjamin Lowry, Zhiying Liu, and Ji Wu. The influence of role stress on self-disclosure on social networking sites: A conservation of resources perspective. *Information &amp Management*, 56(7):103147, nov 2019.

[300] Ramyasree Vedantham. Vedantham_ramyasree_2021_memoire.pdf. `https://papyrus.bib.umontreal.ca/xmlui/bitstream/handle/1866/26080/Vedantham_Ramyasree_2021_memoire.pdf?sequence=2&isAllowed=y`, jun 2021. (Accessed on 06/22/2023).

[301] Jie Gao. Chinese sentiment classification model based on pre-trained BERT. In *2021 2nd International Conference on Computers, Information Processing and Advanced Education*. ACM, may 2021.

[302] Valerii D. Oliseenko, Michael Eirich, Alexander L. Tulupyev, and Tatiana V. Tulupyeva. BERT and ELMo in task of classifying social media users posts. In *Proceedings of the Sixth International Scientific Conference "Intelligent Information Technologies for Industry" (IITI'22)*, pages 475–486. Springer International Publishing, oct 2022.

[303] Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. User-oriented fairness in recommendation. In *Proceedings of the Web Conference 2021*. ACM, apr 2021.

[304] Anind K. Dey. Understanding and using context. *Personal and Ubiquitous Computing*, 5(1):4–7, feb 2001.

[305] Angela Carrera-Rivera, Felix Larrinaga, and Ganix Lasa. Context-awareness for the design of smart-product service systems: Literature review. *Computers in Industry*, 142:103730, nov 2022.

[306] Philipp K. Masur, Dominic DiFranzo, and Natalie N. Bazarova. Behavioral contagion on social media: Effects of social norms, design interventions, and critical media literacy on self-disclosure. *PLOS ONE*, 16(7):e0254670, jul 2021.

[307] Fatemeh Yari, Alireza Mehrazeen, Reza Yarifard, and Abolghasem Masihabadi. Risk appetite, risks of business continuity, managerial ability and accountability. *Journal of Accounting and Social Interests*, 11(2), August 2021.

[308] Bart Knijnenburg. A user-tailored approach to privacy decision support. *Dissertation*, 2015.

[309] Cbs San Francisco. Facebook will stop automatic 'tag suggestions' on your friends' faces in photos - cbs san francisco. `https://www.cbsnews.com/sanfrancisco/news/facebook-will-stop-automatic-tag-suggestions-on-your-friends-faces-in-photos/`, sep 2019. (Accessed on 06/20/2023).

[310] Rim Ben Salem, Esma Aïmeur, and Hicham Hage. A multi-party agent for privacy preference elicitation. *Artificial Intelligence and Applications*, 2023.

[311] Patricio Cerda, Gaël Varoquaux, and Balázs Kégl. Similarity encoding for learning with dirty categorical variables. *Machine Learning*, 107(8-10):1477–1494, jun 2018.

[312] Rim Ben Salem, Esma Aïmeur, and Hicham Hage. Aegis: An agent for multi-party privacy preservation. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, jul 2022.

[313] Rim Ben Salem, Esma Aïmeur, and Hicham Hage. The privacy versus disclosure appetite dilemma: Mitigation by recommendation. In *Proceedings of the Workshop on Online Misinformation- and Harm-Aware Recommender Systems @ ACM Recommender Systems Conference (RecSys)*, June 2021.

[314] Ethan Mollick and Ramana Nanda. Wisdom or madness? comparing crowds with expert evaluation in funding the arts. *Management Science*, 62(6):1533–1553, jun 2016.

[315] Zhi Da and Xing Huang. Harnessing the wisdom of crowds. *Management Science*, 66(5):1847–1867, may 2020.

[316] Michael W. Kattan, Colin O'Rourke, Changhong Yu, and Kevin Chagin. The wisdom of crowds of doctors. *Medical Decision Making*, 36(4):536–540, apr 2015.

[317] Jeffrey Prince and Scott Wallsten. How much is privacy worth around the world and across platforms? *SSRN Electronic Journal*, 2020.

[318] Eva-Maria Schomakers, Chantal Lidynia, Dirk Müllmann, and Martina Ziefle. Internet users' perceptions of information sensitivity – insights from germany. *International Journal of Information Management*, 46:142–150, jun 2019.

[319] John M.M. Rumbold and Barbara K. Pierscionek. What are data? a categorization of the data sensitivity spectrum. *Big Data Research*, 12:49–59, jul 2018.

[320] Ereni Markos, George R. Milne, and James W. Peltier. Information sensitivity and willingness to provide continua: A comparative privacy study of the united states and brazil. *Journal of Public Policy &amp Marketing*, 36(1):79–96, apr 2017.

[321] Bart P. Knijnenburg, Reza Ghaiumy Anaraky, Daricia Wilkinson, Moses Namara, Yangyang He, David Cherry, and Erin Ash. User-tailored privacy. In *Modern Socio-Technical Perspectives on Privacy*, pages 367–393. Springer International Publishing, jul 2021.

[322] Nicolás E. Díaz Ferreyra, Tobias Kroll, Esma Aïmeur, Stefan Stieglitz, and Maritta Heisel. Preventative nudges: Introducing risk cues for supporting online self-disclosure decisions. *Information*, 11(8):399, aug 2020.

[323] Daniel J. Solove. Data is what data does: Regulating use, harm, and risk instead of sensitive data. *SSRN Electronic Journal*, 2023.

[324] Karina Polanco-Levicán and Sonia Salvo-Garrido. Understanding social media literacy: A systematic review of the concept and its competences. *International Journal of Environmental Research and Public Health*, 19(14):8807, jul 2022.

[325] Mohamed El Ghazali Kimeche, Kenza Makhloufi, Esma Aïmeur, Rim Ben Salem, and Amina Selma Haichour. Protection de la vie privée par l'étude du comportement des utilisateurs sur une plateforme de réseaux sociaux simulée. jun 2023.

[326] Jure Leskovec and Andrej Krevl. {SNAP Datasets}: {Stanford} large network dataset collection. 2014.

[327] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9), feb 2021.

[328] Daejin Choi, Selin Chun, Hyunchul Oh, Jinyoung Han, and Ted "Taekyoung" Kwon. Rumor propagation is amplified by echo chambers in social media. *Scientific Reports*, 10(1), jan 2020.

[329] Prokopia Vlachogianni and Nikolaos Tselios. Perceived usability evaluation of educational technology using the system usability scale (SUS): A systematic review. *Journal of Research on Technology in Education*, 54(3):392–409, feb 2021.

[330] Katrin Hartwig and Christian Reuter. Nudging users towards better security decisions in password creation using whitebox-based multidimensional visualisations. *Behaviour &amp Information Technology*, 41(7):1357–1380, jan 2021.

[331] Ryo Yoshikawa, Hideya Ochiai, and Koji Yatani. Dualcheck: exploiting human verification tasks for opportunistic online safety microlearning. In *Proceedings of the 2022 Symposium on Usable Privacy and Security (SOUPS 2022)*. ACM, aug 2022.

[332] Yunxiao Chen, Xiaoou Li, and Siliang Zhang. Joint maximum likelihood estimation for high-dimensional exploratory item factor analysis. *Psychometrika*, 84(1):124–146, nov 2018.

[333] Trevor G. Bond, Zi Yan, and Moritz Heene. *Applying the Rasch Model*. Routledge, jul 2020.

[334] Clauser B. and Linacre J.M. Relating cronbach and rasch reliabilities. *Rasch Measurement Transactions*, 13(2):696, 1999.

[335] Karin Schermelleh-Engel, Helfried Moosbrugger, and Hans Müller. Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research*, 8(2):23–74, 2003.

[336] Matthieu J.S. Brinkhuis and Gunter Maris. Dynamic estimation in the extended marginal rasch model with an application to mathematical computer-adaptive practice. *British Journal of Mathematical and Statistical Psychology*, 73(1):72–87, mar 2019.

[337] Birgit Beck and Michael Kühler, editors. *Technology, Anthropology, and Dimensions of Responsibility*. J.B. Metzler, 2020.

[338] Tim Berners-Lee. Tim berners-lee: I invented the web. here are three things we need to change to save it | tim berners-lee | the guardian. `https://www.theguardian.com/technology/2017/mar/11/tim-berners-lee-web-inventor-save-internet`, mar 2017. (Accessed on 06/29/2023).

[339] Amanda Silberling. Lensa ai climbs the app store charts as its 'magic avatars' go viral | techcrunch. `https://techcrunch.com/2022/12/01/lensa-ai-climbs-the-app-store-charts-as-its-magic-avatars-go-viral/?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAE8QQlKEj_aVN12Qlek2NZXpUKfrem69hATusOr9OS_Idq2ONs-VJyTwNeeNCN1TFbFYSOxhSTFo94InTvgRfXX0tW20vC4KtozN-Ynfn9gIx65OFW8qL_n4IVrqImdUZkbvr3GCvrBTM8wHmgNMm1i8IaBO5QATA6R7O8V-o_Bf`, dec 2022. (Accessed on 06/29/2023).

[340] What does the lensa ai app do with my self-portraits and why has it gone viral? | artificial intelligence (ai) | the guardian. `https://www.theguardian.com/culture/2022/dec/09/what-does-the-lensa-ai-app-do-with-my-selfies-self-portrait-photos-magic-avatar-generator-gone-viral`, dec 2022. (Accessed on 06/29/2023).

[341] Terms of use. https://tos.lensa-ai.com/terms#:~:text=You%20acknowledge%20and%20agree%20that,subject%20to%20our%20Privacy%20Policy.&text=TL%3BDR%3A%20You%20can%20upload,no%20ownership%20over%20such%20content., de 2022. (Accessed on 06/29/2023).

[342] Lily Hay Newman. Ai wrote better phishing emails than humans in a recent test | wired. https://www.wired.com/story/ai-phishing-emails/, aug 2021. (Accessed on 06/29/2023).

[343] CBC News. More canadian privacy authorities investigating chatgpt's use of personal information | cbc news. https://www.cbc.ca/news/canada/british-columbia/canada-privacy-investigation-chatgpt-1.6854468, may 2023. (Accessed on 06/30/2023).

[344] Tim Donkers and Jürgen Ziegler. The dual echo chamber: Modeling social media polarization for interventional recommending. In *Fifteenth ACM Conference on Recommender Systems*. ACM, sep 2021.

[345] Iva Nenadić. Unpacking the "european approach" to tackling challenges of disinformation and political manipulation. *Internet Policy Review*, 8(4), dec 2019.

[346] Dorsaf Sallami, Rim Ben Salem, and Esma Aïmeur. Trust-based recommender system for fake news mitigation. In *Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*. ACM, jun 2023.

[347] M. Anne Britt, Jean-François Rouet, Dylan Blaum, and Keith Millis. A reasoned approach to dealing with fake news. *Policy Insights from the Behavioral and Brain Sciences*, 6(1):94–101, March 2019.

Annexe A

Appendix

Fig. A.1. About page

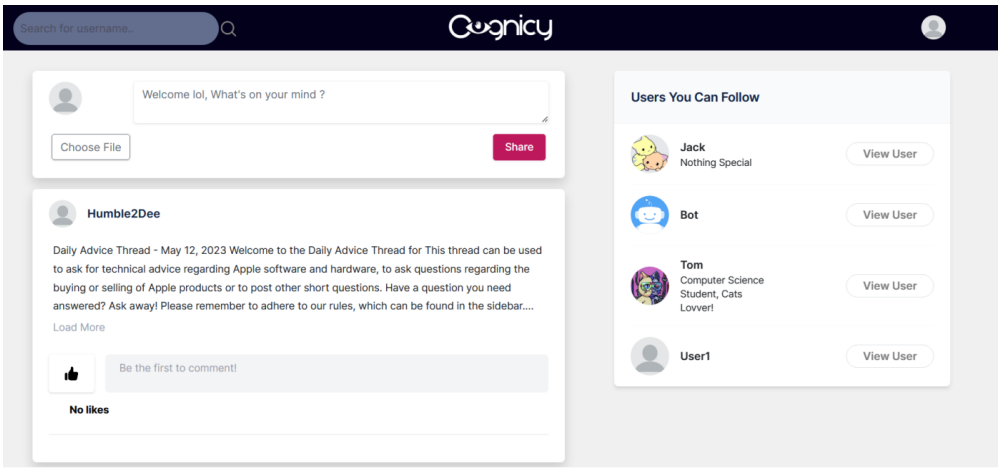

Fig. A.2. Cognicy home page

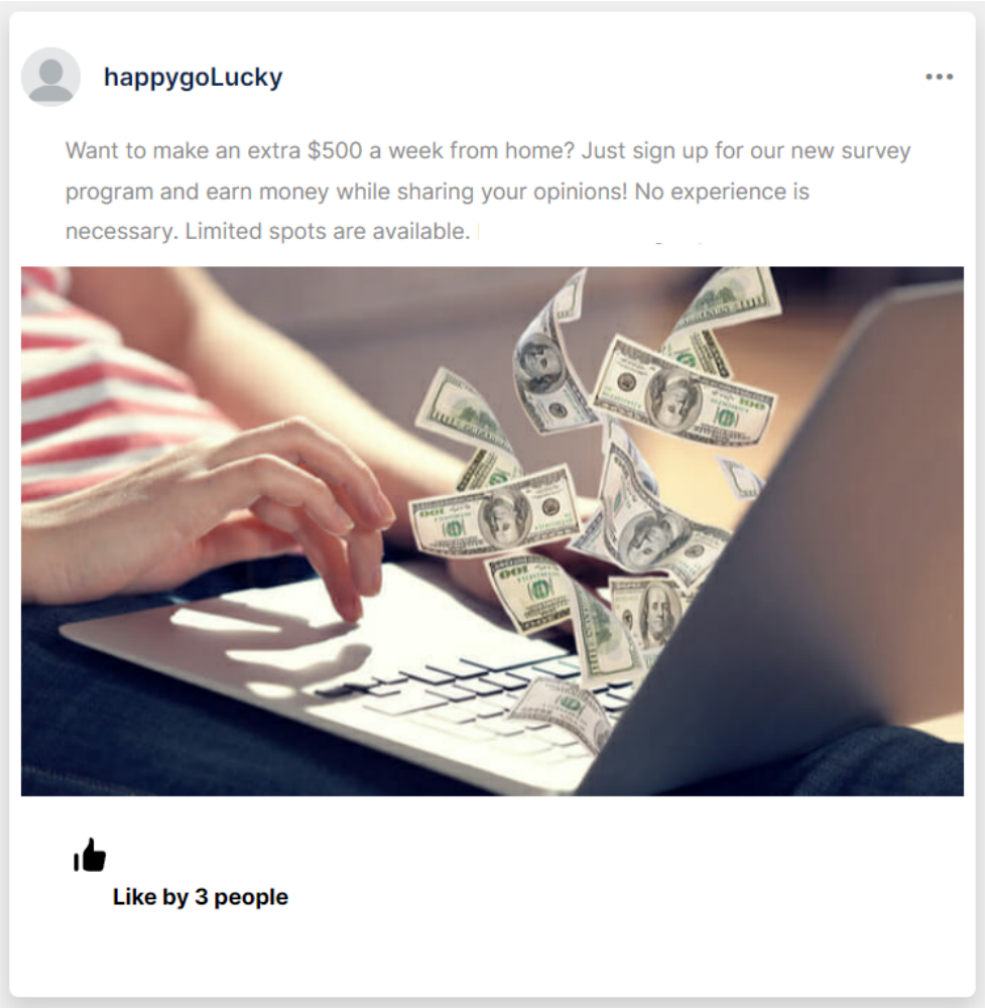

Fig. A.3. Financial gain post

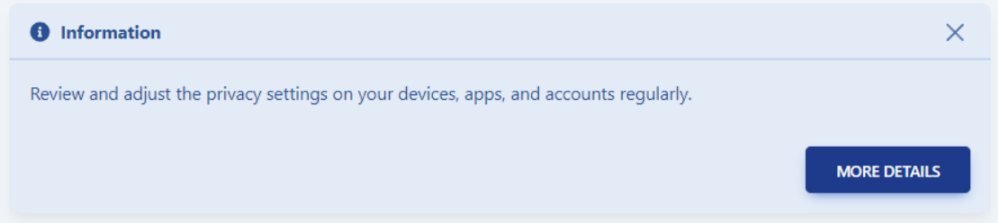

Fig. A.4. One size fits all nudge information offering information

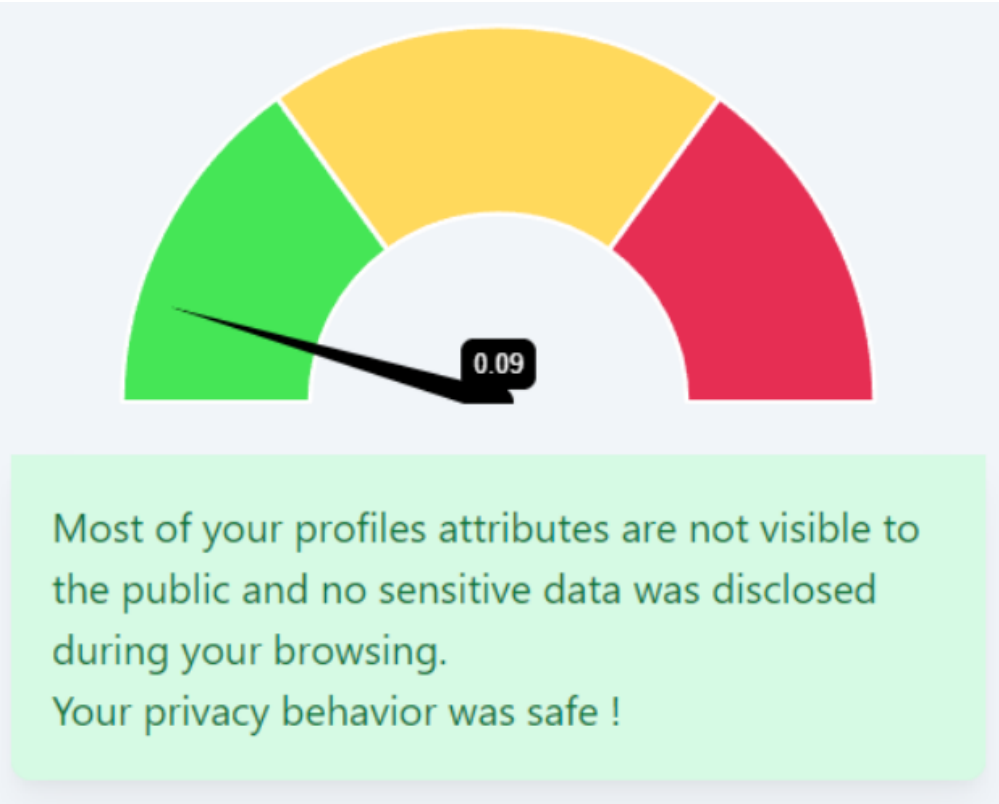

Fig. A.5. privacy report gauge: Example 1

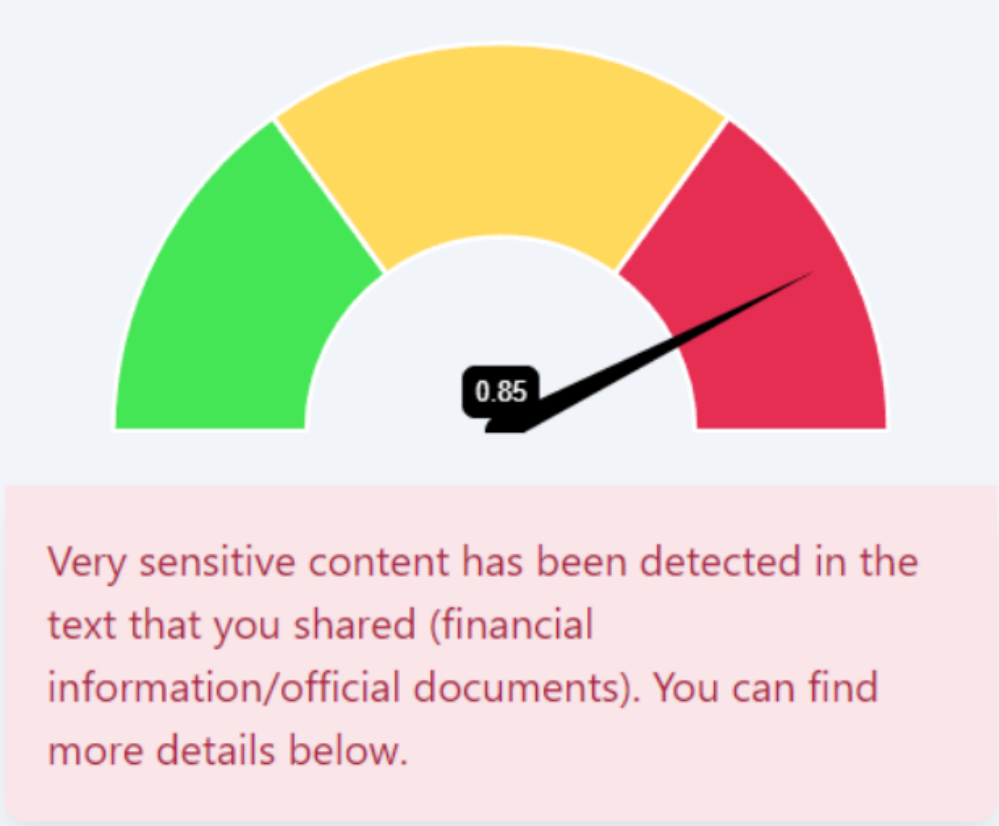

Fig. A.6. privacy report gauge: Example 2

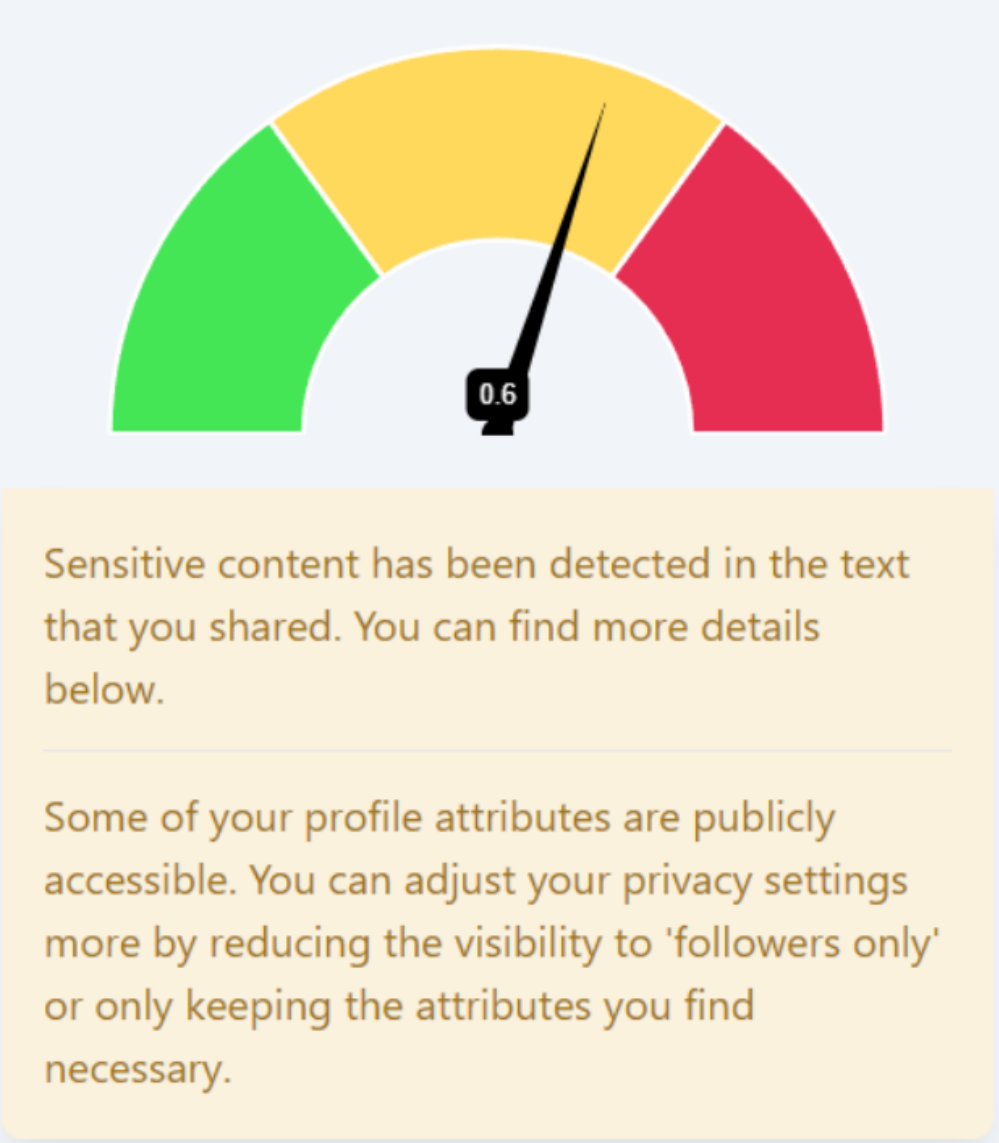

Fig. A.7. privacy report gauge: Example 3

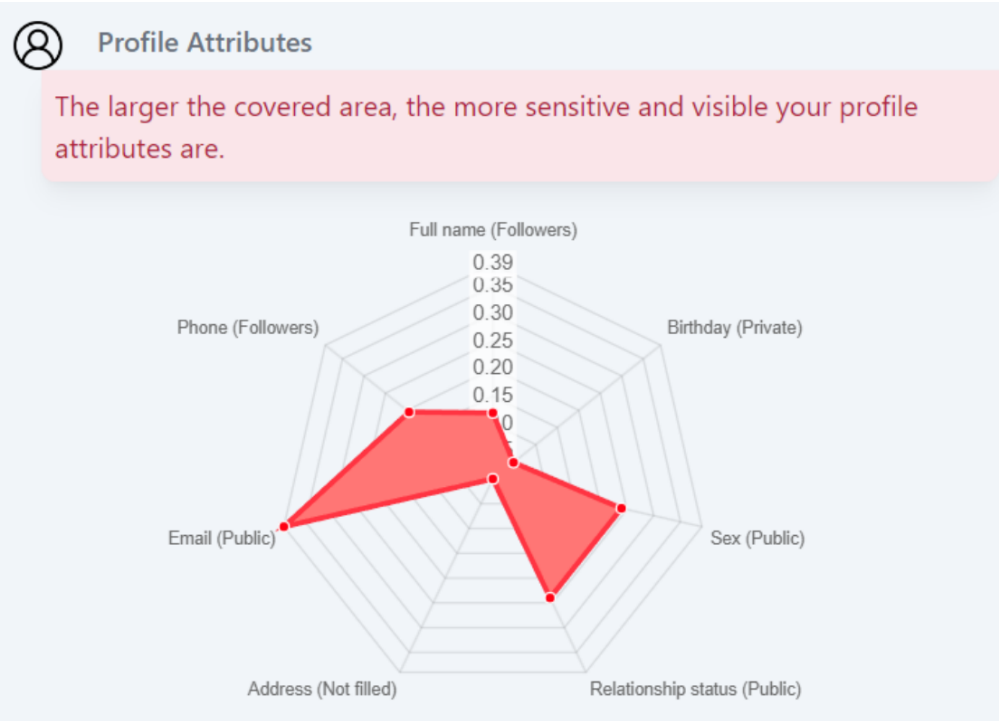

Fig. A.8. Profile attributes

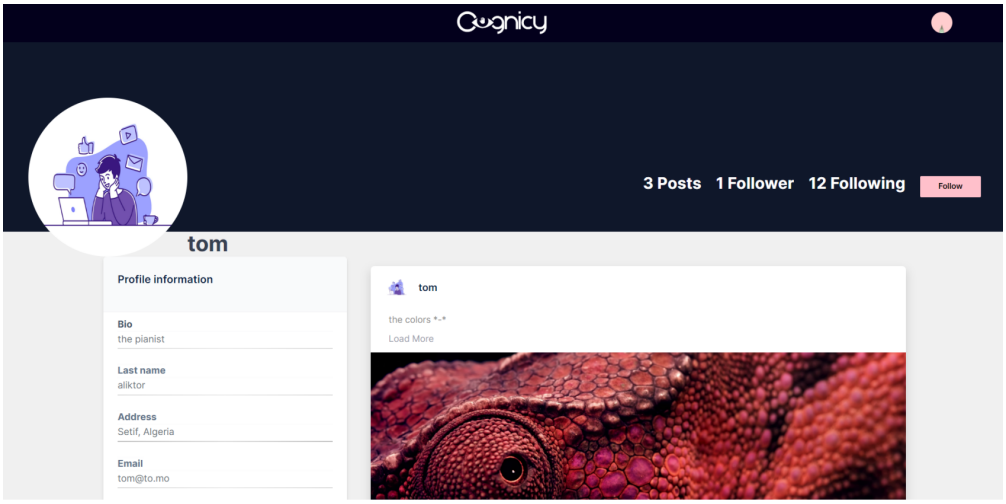

Fig. A.9. Profile page

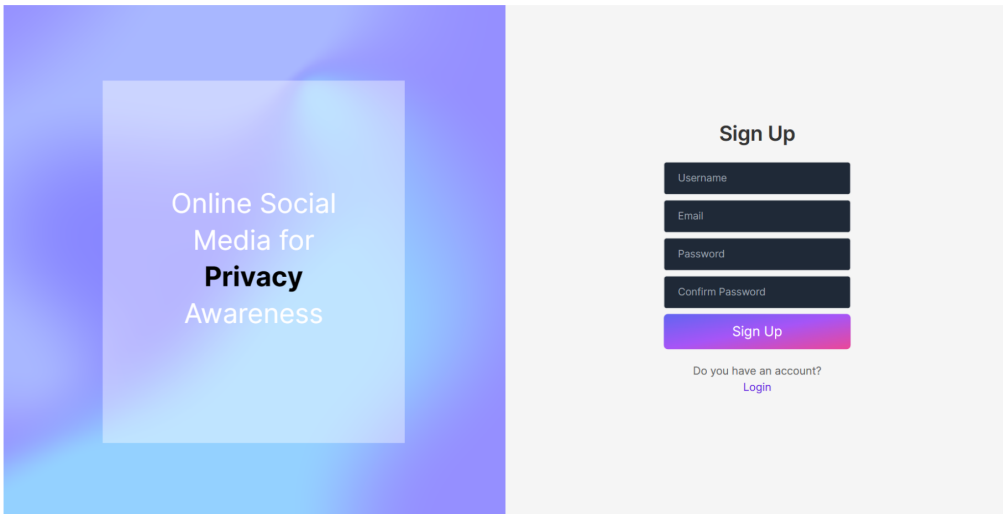

Fig. A.10. Sign up page

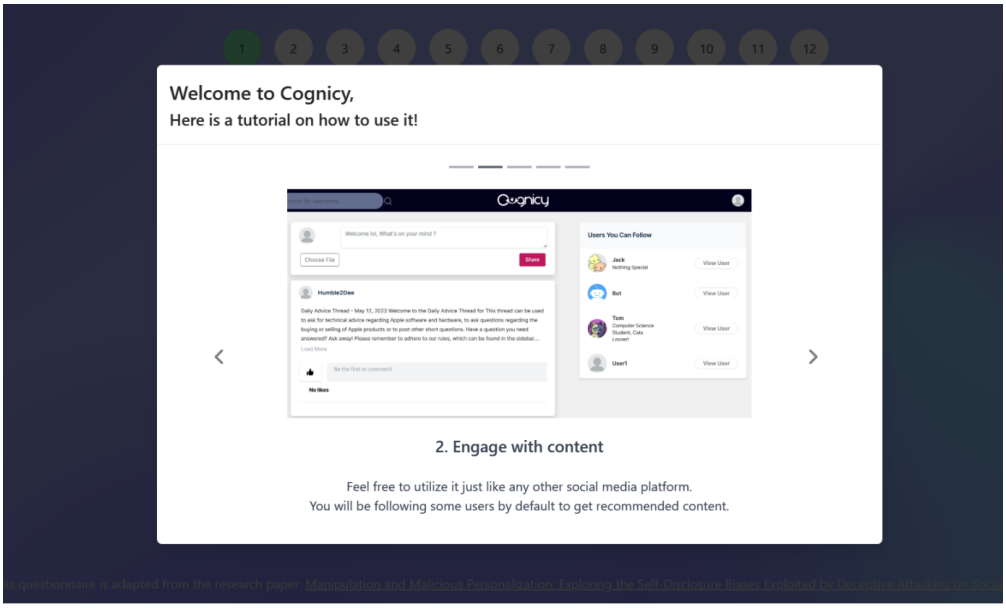

Fig. A.11. Tutorial after signing up