A User-centric and Agent-based Approach to Multi-user Privacy in Online Social Networks

Mosca, Francesca

Awarding institution:
King's College London

A User-centric and Agent-based Approach to Multi-user Privacy in Online Social Networks

By
Francesca Mosca

A thesis submitted in fulfilment for the
degree of Doctor of Philosophy

in the
Department of Informatics
School of Natural, Mathematical & Engineering Sciences
King's College London

2023

To the family that was,

that is,

and that will be.

Acknowledgements

This PhD has been a long journey in which I have learnt a lot, especially about myself. Despite it being a mostly lonely experience, I have enormously benefited from interacting with a large number of people, whose help, either personal or professional, has been vital for the completion of this work.

First of all, I would like to acknowledge the financial support that I have received from the Department of Informatics and the NMES Faculty at King's College London through their scholarship, and from J.P. Morgan & Chase through their PhD Fellowship in AI Research.

Second, I wish to thank Professor Jose Such, who, providing constant supervision and valuable feedback, has continuously guided my research work since my Masters' time. Also, I'd like to thank Professor Peter McBurney, who first introduced me to value-based reasoning, triggering my interest in this direction. A special thanks goes to my PhD examiners as well, Professors Pinar Yolum and Timothy Norman: their attentive comments have certainly improved this thesis, and their insightful questions have made my viva a very enjoyable experience.

Third, I am extremely grateful for both my long lasting and more recent friends, whose company, filled of lighthearted laughs, serious talks about the future and amazing memories overall, has made this challenging journey more bearable. Edoardo, Ilaria, Umberto, Diego, Yani, Xavi, Parisa, Moises, David, and many others: thank you!

Then, I will be forever thankful towards my family, who has always provided me with the best physical and mental space to recharge. They have empathised with me along all the way, celebrating my successes and comforting me when I needed it the most.

Finally, without Stefan this thesis would have not been the same. He has been the most precious partner in both my personal and professional life. His spontaneous approach and enthusiastic feedback have constantly encouraged me to improve myself and my work. From the bottom of my heart, thank you.

Abstract

Among the issues that arise from services on online platforms such as social networks (OSNs), there is an increasing concern about privacy and data protection. This is exacerbated when looking at multi-user privacy (MP), i.e. when the privacy decisions of an individual impact multiple stakeholders, which is an issue that has gathered little attention so far.

Following an iterative Value Sensitive Design approach and informed by the literature in Privacy and Autonomous Systems, in this thesis I investigate the design of autonomous systems that can effectively support OSNs users manage MP. This investigation culminates in ELVIRA, a user-centric multi-agent architecture that, by engaging in practical reasoning, recommends optimal collaborative solutions to MP conflicts. The optimality of the solutions is measured by considering not only the contextual privacy preferences of all the users involved, but also their moral values. Furthermore, ELVIRA justifies such privacy recommendations by producing tailored explanations, whose format has been investigated and validated with users.

Through software simulations and a user study, I demonstrate how the agent ELVIRA presents a combination of features that enables it to provide a more satisfactory support for users than alternative state-of-the-art approaches for managing MP in OSNs. In particular, ELVIRA's privacy recommendations are more acceptable across demographics, and ELVIRA's explanations nudge users to be more respectful of others' preferences and more appreciative of fair solutions to MP conflicts. Additionally, drawing from evolutionary game theory and simulating a word-of-mouth marketing strategy, I show how ELVIRA could be widely and stably adopted by OSNs users.

Finally, I outline possible extensions of the ELVIRA model, such as the definition of interactive explanations and the management of non-collaborative behaviour.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem statement and motivation

Our society is hyper-connected. People are constantly online, while engaging with an increasing amount of services offered through the internet. In particular, Online Social Networks (OSNs) are a pervasive phenomenon of our time. As of October 2021 (see Figure 1.1), billions of people worldwide access one or more social network platforms regularly (at least once per month). According to an online survey of 2500 OSNs users performed in February 2019 in the United States, the majority of users engage with visual content online (see Figure 1.2). It is estimated that in a single minute 240k photos are shared on Facebook and 65k on Instagram[i]. Generally, users manage their own privacy by specifying access control mechanisms that limit the audience of the content they share online, usually according to the characteristics of other users in the same network (e.g., whether the other user is a contact, a friend, a follower, etc.) and in line with the platform of choice.

Thanks to recent progress in data privacy legislation (see for instance the GDPR in Europe and the CCPA in California), users of online platforms are now generally more protected against misuse and misappropriation of personal data, but many other challenges need to be tackled before (if ever) users can be completely safe

[i] Estimate based on the internet usage in August 2021, see `https://www.domo.com/learn/infographic/data-never-sleeps-9`, accessed 30/01/2022.

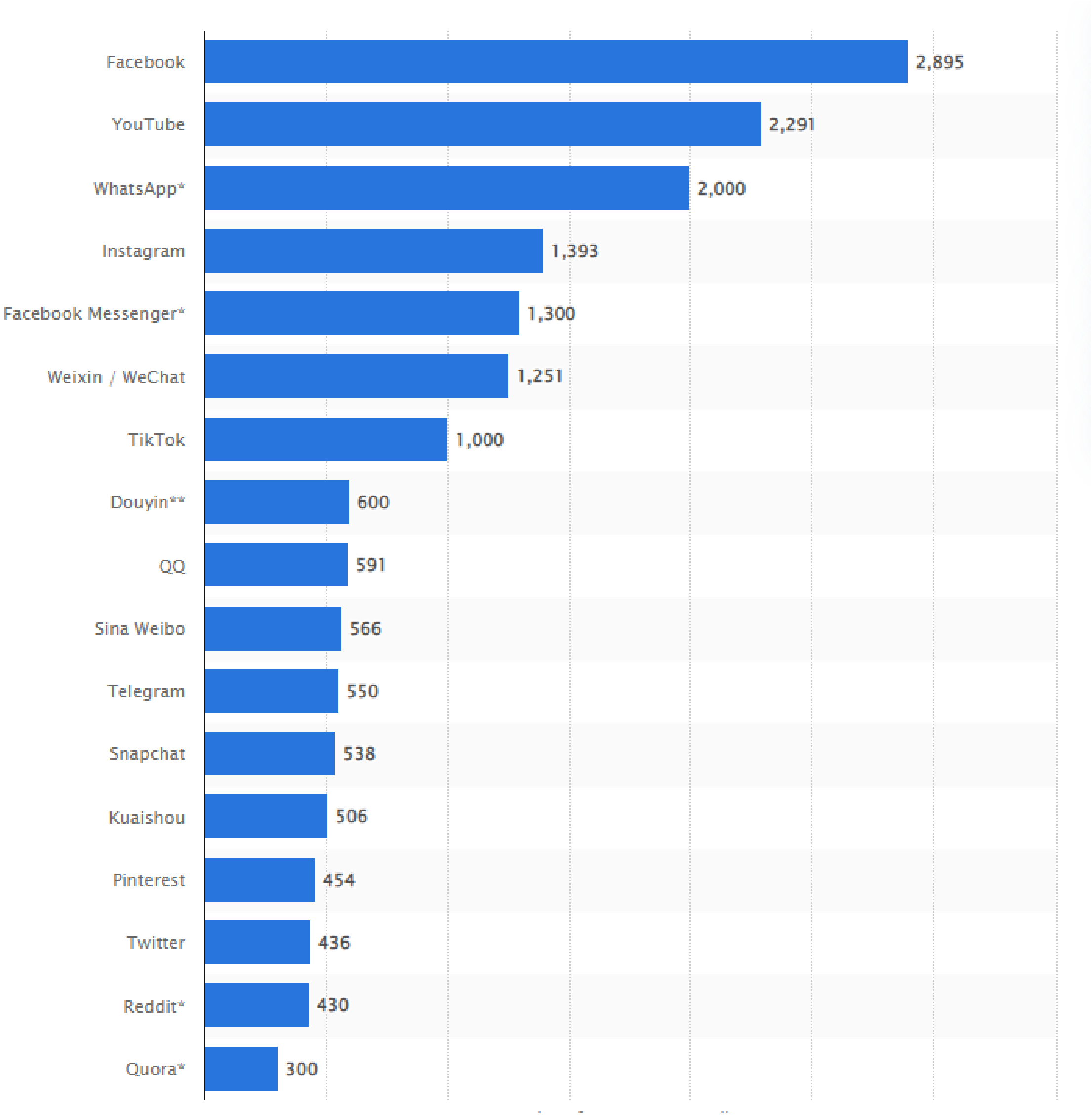
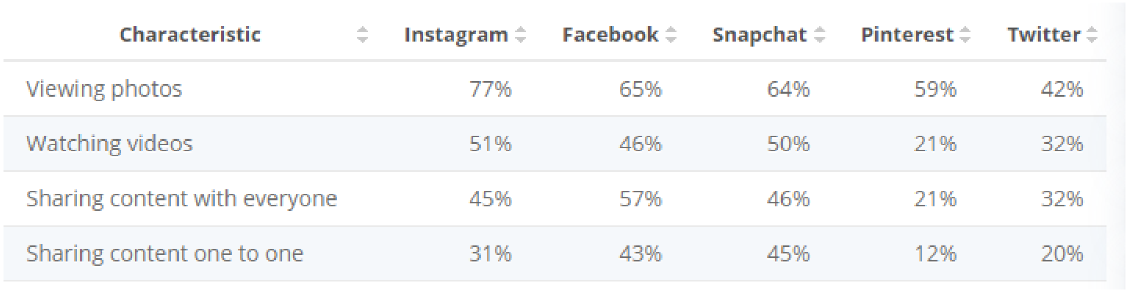


Figure 1.1: Most popular social networks worldwide as of October 2021, ranked by number of active users (in millions). Notes: * = Platforms have not published updated user figures in the past 12 months, figures may be out of date and less reliable; ** = Figure uses daily active users, so monthly active user number is likely higher. This diagram is taken from `https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/`, accessed 29/01/2022.



Characteristic	Instagram	Facebook	Snapchat	Pinterest	Twitter
Viewing photos	77%	65%	64%	59%	42%
Watching videos	51%	46%	50%	21%	32%
Sharing content with everyone	45%	57%	46%	21%	32%
Sharing content one to one	31%	43%	45%	12%	20%

Figure 1.2: Percentage of social media users (2500) in the United States who engaged in selected activities on OSNs in February 2019. This is taken from `https://www.statista.com/statistics/200843/social-media-activities-by-platform-usa/`, accessed 30/01/2022.

online. One of the outstanding issues in online privacy, which is the main motivation of my work and is critically relevant to OSNs, is the general lack of support for *multi-user privacy* (MP), often referred to as multi-party privacy or collective privacy [134, 83]. MP concerns all those situations where the privacy choices of an individual impact the privacy of other people. In fact, privacy does not regard only what we decide to disclose about ourselves, but also what others disclose about us [173].

Online collaborative platforms, and in particular OSNs, naturally represent a risk for multi-user privacy, because even though they allow users to collaboratively create, modify and interact with digital content, they usually do not allow to collaboratively manage the privacy settings of that digital content. In fact, an OSN user's privacy can be threatened not only by his/her misuse of the access control mechanisms made available by the platform, but also by their misuse (often unintentional) from other users. Whenever someone's privacy expectations are not aligned with those of the other users one interacts with on the OSN, *multi-user privacy conflicts* (MPCs) are likely to arise. As an example of MPC, consider the case where Alice, Bob and Charlie are all "friends" on an OSN, i.e., they are connected. Alice and Bob are attending a party which Charlie was not invited to. Bob would like to upload on the OSN a picture he took with Alice during the party. However, Alice knows that Charlie would be upset to see Alice and Bob attending a party together without him, so she prefers that no picture of the party would be uploaded. If Bob was actually to share the picture, he would compromise Alice's privacy. As Such and Criado well summarise in [174], OSN platforms currently support only *reparative solutions*, which result in unsatisfactory or awkward situations for the involved users. For instance, Alice could untag herself, but, apart from the fact that Charlie might have seen the photo as soon as it was posted, that is even before Alice could react, Alice and Bob's photo would still be available on the platform for Charlie to see it. Otherwise, if the photo was considered indecent (e.g., depicting nudity or violence), Alice could report the picture as inappropriate to the OSN, but without any guarantee of an efficient and timely action[ii].

MPCs are highly prevalent. In a recent user study which involved more than a thousand people [175], 30% of the participants reported to have experienced at least one MPC within one month of the survey date, 44% within six months, and 99% within their overall experience on OSNs. Even though not all conflicts can be classified as of high severity, too often they unnecessarily spoil the users' activities on the platforms.

All this evidence that I just presented urges researchers in Usable Privacy to intensify their efforts towards a better support for multi-user privacy management in OSNs. *Privacy assistants* have been argued to be a promising approach to help users manage their individual and collective online privacy: in fact, the widespread introduction of autonomous agents such as desktop and smartphone personal assistants may not be just another threat to privacy, but also a solution [177]. For instance, a privacy personal assistant may negotiate data sharing when interacting with online services [20] and learn and elicit the preferences of their users [21, 29, 57], who are often unaware of their preferences and oblivious of the privacy implications of their online behaviour [2]. Other types of privacy personal assistants have been suggested,

[ii] See, for instance, the Facebook policy `https://www.facebook.com/help/1753719584844061` (accessed 30/01/2022).

drawing from alternative common techniques in autonomous systems: agents that follow privacy norms (see [48, 5] among others), that engage in privacy-driven game theoretical dynamics (e.g., [166, 142, 187] etc.), that present arguments or semantic rules to justify privacy decisions (cf. [91, 111, 60] and others), etc.

However, when designing such privacy assistants, our attention should not be narrowly focused only on the privacy application: instead, we should also keep in mind the more general challenges that regard the safe deployment of autonomous systems in society, such as the *value alignment* problem and the *explainability* problem. In fact, in order to be truly supportive for users, privacy assistants should behave according to their users' expectations and recommend privacy decisions that do not contradict their human rights and values [56]. In addition, privacy assistants should be able to explain and justify their privacy decisions, which would bring about two main beneficial consequences: (i) users could fully understand the reasons and the consequences of the received recommendations and would be able to critically evaluate them (e.g., accept or reject); and (ii) privacy assistants could prove their value-alignment, i.e. their coherency with the users' values, potentially increasing their users' trust towards them [204].

1.2 Aim of this thesis

Given the need to provide a better support for multi-user privacy management in OSNs, I have dedicated my research work to the investigation, design, implementation and evaluation of autonomous agents that can *effectively* help OSN users manage their online privacy. In order to provide effective support, autonomous systems need to be user-centric, that is they need to recommend privacy actions and decisions that are consistent with the users' preferences, i.e. they need to be value-aligned, and clear to understand, i.e. they need to be explainable. In particular, I have aimed to answer the following main research question:

> **RQ:** How do we design autonomous systems that can *effectively* help users of online social networks manage multi-user privacy?

and the following three research sub-questions that have consequently emerged:

> **RQ-A:** Which *features* do autonomous systems need to present in order to effectively help users of online social networks manage multi-user privacy?
>
> **RQ-B:** How do we design *value-aligned* autonomous systems that can effectively help users of online social networks manage multi-user privacy?
>
> **RQ-C:** How do we design *explainable* autonomous systems that can effectively help users of online social networks manage multi-user privacy?

In order to answer the above questions, I have followed a Value Sensitive Design approach (VSD – see Section 2.4) [66], which is a theoretically grounded methodology that aims to include human values into the design, research and development of Information and Communication Technologies. VSD drives the technology designer through a three-phases iterative process: conceptual investigation, technical investigation and empirical investigation.

1.3 Overview

In this thesis I describe the scientific process, driven by the Value Sensitive Design approach (VSD), that I followed to answer the research questions above.

1.3.1 Conceptual investigation - I iteration

The initial part of my research focused on a first iteration of conceptual investigation with the aim of answering **RQ-A**. It consisted of an extensive and critical analysis of the previous literature on online privacy and autonomous systems (ASs) [2, 177, 94] and, more specifically, on both theoretical studies and empirical evidence on multi-user privacy (MP) on OSNs [134, 175, 207, 97]. From this, I gathered some insights regarding the features that ASs should present in order to be considered helpful and effective in managing MP.

First of all, ASs should aim to put all users involved in an MPC on an equal footing regardless of whether they are uploaders or co-owners of the content, so the perspectives of all the users are taken into account. This is because empirical evidence tells us that many of the MPCs are due to the available access control mechanisms, which only consider the perspective of one user, who tends to be the uploader [207]. Hence, for effectively solving MPCs, ASs should be *role-agnostic*.

Then, ASs should behave differently according to the users' subjective preferences, because different individuals manage privacy in different ways depending on the context [2, 129]. This means that ASs should be context-aware and able to map contexts onto privacy preferences, in order to adapt their behaviour to shifts in context. Hence, for effectively solving MPCs, ASs should be *adaptive*.

As part of the general need for ASs to be value-aligned with their users, ASs should take into account moral values when managing MP, because empirical evidence suggests that users do so [175]. For instance, some users go beyond their perceived personal gain to consider the consequences of their actions on others, or self-transcend to accommodate others' preferences. Hence, for effectively solving MPCs, ASs should be *value-driven*.

In summary, the first iteration of conceptual investigation identified role-agnosticism, adaptability and to be value-driven as the required features for an AS to effectively support MP.

1.3.2 Technical investigation - I iteration

Moving on to a first attempt of technical investigation, I designed an agent-based model, namely JIMMY, that is role-agnostic, adaptive and value-aligned with its users when recommending decisions w.r.t. MPCs in OSNs. I defined the agent's value-alignment drawing from the Schwartz Theory of Basic Values, which asserts that values drive human behaviour. Considering this, and as a preliminary answer to **RQ-B**, JIMMY aims to recommend, during the negotiation of privacy policies with other users, decisions and behaviours that are consistent with the moral and attitude-related preferences of its users.

1.3.3 Conceptual investigation - II iteration

Despite being in line with the empirical evidence that sees users sometimes going beyond their personal interest in order to accommodate others' privacy preferences, JIMMY seemed to be too naive, providing solutions that were aligned with the user's morality, but too far from their initial privacy preferences. For instance, to be benevolent towards others does not imply that the user will accept any sharing policy that other users may propose. Therefore, in a second iteration of the conceptual investigation, I have included the fact that ASs should consider solutions to MPCs according to the personal advantage or disadvantage that the users involved can face in terms of both: positively enjoying the benefits of sharing in OSN and maintaining relationships [94]; and negatively experiencing privacy violations [97, 207]. Hence, for effectively solving MPCs, ASs should not only be value-aligned, but also *utility-driven*.

Furthermore, as part of the second iteration of conceptual investigation, it emerged that the capability of an AS to provide an explanation of its processes [112], generally desirable for reasons of trustworthiness [204], accountability [45], and responsibility [56] is particularly crucial in the MPC context for allowing users to make informed choices, to know why a solution is suggested and its effects [134], and to align the differences between uploaders and co-owners [175]. Hence, for effectively solving MPCs, ASs should be *explainable*.

1.3.4 Technical and empirical investigation - later iterations

In a second iteration of technical investigation within the VSD, I have designed another agent-based model, namely ELVIRA, which presents all the mentioned features. In particular, whenever an MPC occurs, the agent first engages in a value- and utility-driven practical reasoning process to identify the best solution, and then it conveys and justifies such solution to the user. Several iterations of technical and empirical investigations have (a) confirmed the benefits of considering both utility and values when computing an MPC solution through software simulations; (b) informed a desirable design of explanations through a user study; and (c) shown how ELVIRA's recommendations and explanations were more frequently and better accepted than the ones generated by other state-of-the-art models in another user study. Such comparison with other state-of-the-art models, which present different subsets of the identified required features, confirmed that the five requirements together guarantee better chances for an effective MP support.

Finally, given this evidence, I concluded the VSD process that drove my research to answer my research questions was complete, because:

- the conceptual investigation suggested that autonomous systems, in order to effectively help users of OSNs manage multi-user privacy, should be **role-agnostic**, **adaptive**, both **value-** and **utility-driven**, and **explainable** (cf. **RQ-A**).

- the technical and empirical investigations I conducted suggested that ELVIRA, being value-aligned with its user (cf. **RQ-B**) and able to explain its solutions (cf. **RQ-C**), can effectively help users of OSNs manage multi-user privacy (cf. **RQ**).

Feature	Definition
Role-agnosticism	to offer equivalent support to all the users involved in the MPC, independently of them being uploaders or co-owners of the content
Adaptivity	to offer privacy recommendations that are consistent with the contextual preferences of the users involved in the MPC
Value-driven	to offer privacy recommendations that are consistent with the moral and attitudinal preferences of the user
Utility-driven	to offer privacy recommendations that are as close as possible to the user's privacy preference
Explainability	to offer usable explanations for the generated privacy recommendations

Table 1.1: The design features that enable autonomous systems to effectively support multi-user privacy in OSNs, as identified through an iterative conceptual investigation within the Value Sensitive Design approach.

Table 1.1 provides a summary of the features that answer RQ-A in order to facilitate references in later parts of the thesis.

1.4 Publications

Part of the work that I describe in this thesis has been previously published in peer reviewed venues.

- Journal article
  - **Mosca, Francesca** and Jose Such (2022). "An explainable assistant for multiuser privacy". In: *Autonomous Agents and Multi-Agent Systems* 36.1, pp. 1–45 [125]
- Conference papers

  - **Mosca, Francesca** and Jose Such (2021). "ELVIRA: an Explainable Agent for Value and Utility-driven Multiuser Privacy". In: *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 916-924 [124]
  - **Mosca, Francesca** (2020). "Value-Aligned and Explainable Agents for Collective Decision Making: Privacy Application". In: *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 2199–2200 [122]
  - **Mosca, Francesca**, Jose Such, and Peter McBurney (2020). "Towards a Value-driven Explainable Agent for Collective Privacy". In: *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 1937–1939 [126]

- Workshops and *Symposia* papers
  - **Mosca, Francesca**, Ştefan Sarkadi, Jose Such, and Peter McBurney (2020). "Agent EXPRI: Licence to Explain". In: *Proceedings of the 2nd International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems.* Springer, pp. 21–38 [123]
  - **Mosca, Francesca**, Jose Such, and Peter McBurney (2019). "Value-driven collaborative privacy decision making". In: *Proceedings of the AAAI Spring Symposium on Privacy-Enhancing Artificial Intelligence and Language Technologies*, pp.13-20 [121]

1.5 Structure of this thesis

In this chapter, I have introduced the motivations and the scope of the work I present in this thesis, namely the design of user-centric autonomous systems (ASs) to effectively support the management of multi-user privacy on online social networks.

In order to succeed in this direction, it is necessary to take into account the already known challenges that the deployment of ASs in society brings upon, such as their value alignment and their explainability.

In the next chapter, Chapter 2, I summarise the interdisciplinary background knowledge that represents the foundations of my work and that the reader should be familiar with in order to better appreciate my contribution. Then, in Chapter 3, I survey the research literature on ASs (i) that preserve privacy, (ii) that are value-aligned, and (iii) that are explainable.

In Chapters 4 and 5 I present two agent-based models, respectively JIMMY and ELVIRA, for collaboratively solving multi-user privacy conflicts (MPCs). JIMMY was the first attempt to design an agent that supports MP and it is adaptive, role-agnostic and value-aligned. ELVIRA, instead, satisfies all the required features, including being utility-driven and explainable, for a model to effectively support users to solve MPCs. ELVIRA is then empirically evaluated through software simulations, described in Chapter 6, and through a user study, which I report upon in Chapter 7. Furthermore, drawing from evolutionary game theory, I simulate in Chapter 8 the long-term adoption of ELVIRA in OSNs as a technology to manage MP under different conditions.

Finally, I conclude this thesis in Chapter 9 by summarising my contributions to the field of ASs for multi-user privacy management and by outlining possible future lines of research.

Chapter 2

Background

2.1 Introduction

In this chapter, I introduce the interdisciplinary background knowledge that the reader should be familiar with in order to better appreciate the work presented in this thesis.

First, in Section 2.2, I draw from the Privacy literature. I start presenting the multifaceted concept of privacy, with a spotlight on online informational privacy. Then, considering the main threats that users may encounter online, and specifically on online social networks (OSNs), I focus on the *insider threat* [87] and discuss its individual and multi-party components. To tackle the latter can be said to be the main goal of this thesis.

Next, in Section 2.3, I draw from the Social Sciences and Psychology. There, I illustrate the concept of *human value*, by discussing its definition and the influence of values over human behaviour. I also introduce the Schwartz Theory of Basic Values [154], upon which I define the value-aligned agent-based models in Chapters 4 and 5.

Then, in Section 2.4, I draw from Engineering and Technology design. There, I present *Value Sensitive Design* [66], an iterative methodological approach that guides the designer into accounting for appropriate values (here interpreted as properties or requirements, not in the Social Science meaning of moral values) into technology. This method has driven me through the definition of the requirements for solving multi-party privacy conflicts (MPCs) that I reported in Table 1.1, and through the design and evaluation of the agent-based models for managing multi-user privacy that I discuss throughout the entire thesis.

Furthermore, in Section 2.5, I draw from Artificial Intelligence, more specifically from Agent-based Modelling and Argumentation. There, I report on Atkinson and Bench-Capon's work on *practical reasoning* [14, 15], a process an autonomous agent can follow in order to reason about what to do. I adapt and apply their approach in Chapter 5, where I describe how an agent is able to reason about and explain the best solution for an MPC.

Finally, I draw again from the Social Sciences in Section 2.6, where I report on Miller's findings [112] regarding the nature of explanations in AI. The awareness that the user's idea of *explanation* when interacting with an artificial agent may differ from the one of the designer/engineer of that same agent motivates the theoretical and empirical work that I present in Chapters 5 and 7.

2.2 Theories of privacy

Privacy has been a concept of interest in the human society at least since Aristotle [53], when discussions about the distinction between public and private spheres first emerged. Nowadays, privacy is considered a fundamental right, as recognised in the UN Declaration of Human Rights[i]:

> "No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against

[i] See Art.12, `https://www.un.org/en/about-us/universal-declaration-of-human-rights`

such interference or attacks."

The definition of privacy has varied over time and across cultures, but, given the current "age of information" [2], in this thesis I mostly refer to *informational privacy*, i.e., privacy of personal data. In the following, I report the most well-known conceptualisations of privacy, as described by Such [177]:

- **Confidentiality**: as a security property of computer systems, privacy as confidentiality ensures the prevention of unauthorised reading of information [171], e.g., through encryption and authentication technologies.
- **Notice and Choice**: starting with Westin's definition of privacy as "the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated" [202], the concept of privacy has evolved into the *self-determination right* [144], which highlights the importance of providing notice for any collection and use of personal data, e.g., through privacy policies.
- **Boundary Regulation**: mostly related to social sciences, privacy is an interpersonal boundary regulation process [9, 137], where individuals manage the amount of information they disclose to others according to the social relationships in place.
- **Contextual Integrity**: privacy is contextual, meaning that to disclose or to conceal information can be more or less appropriate according to the context where the information flow happens [129] and is regulated, e.g., through social norms and rules.

In the recent years, there has been an attempt to combine all of the above definitions in order to encompass the *plurality of privacy* [2] when designing, engineering and operationalise privacy in technology [74].

2.2.1 Privacy in Online Social Networks

Nowadays, we engage with online services more than ever and for a variety of purposes, such as communication, information, entertainment, shopping, etc. However, all these services come with potential risks for our privacy, mainly related to personal data collection, processing, management and dissemination [177]: if our personal information is disclosed unwittingly, or is stored inappropriately, or is used for secondary purposes we are not aware of, alarming privacy violations can occur.

In particular, when considering OSNs, users can incur into some more specific privacy threats [62], such as identity theft, unauthorised access, misuse of personal information, stalking, and profiling, all of which are mostly due to third parties attacks. Yet, threats can emerge very commonly also within one's own personal network of contacts: the *insider threat*, defined as "inappropriately sharing content with members of the friend network" [87], has been described among the most worrisome threats for users, who are often unable to mitigate it even with an accurate use of the privacy settings of the online platform. To account for this threat brings forward two main challenges, somewhat complementary, regarding the definition of optimal access control mechanisms to regulate the disclosure of information: (i) how to help users manage self-disclosure of personal information, and (ii) how to help users manage disclosure of personal information from/about other people. While tackling these challenges, it is necessary to follow a user-centric approach, as different users have different needs and preferences in terms of privacy management [3].

Self-disclosure Traditionally, when managing access control for online services, there was a tendency to define *group-based* or *rule-based* access control mechanisms (see [134] for a review on the topic). More recently, the type of interactions available on OSNs has made necessary a new approach that focuses on the users interpersonal relationships, namely *relationship-based* access control (ReBAC) [68]. Ideally, when considering *fine-grained* ReBAC, access control is regulated in terms of nature and strengths of social relationships and can be specified for individual online contents, such as specific posts or pictures. This allows to better mimic the way people disclose personal information in real life, by distinguishing, for example, between relatives, colleagues, and close or distant friends. Even though the current OSN platforms have been mostly implementing ReBAC policies (e.g., friendship on Facebook, or followers on Instagram and Twitter), they generally offer poor granularity and flexibility, and no support to define access control considering the type of content to be shared [62]. This causes users to still encounter many difficulties when managing access control (e.g., on Instagram, it is possible to identify only 'close friends' among the followers, and followers can only be blocked altogether, denying access to all shared content[ii]) [115]. The lack of usability of OSNs access control mechanisms is certainly one of the main causes for inappropriate disclosure of personal information, but this can also be due to other reasons, such as negligence, unawareness or ignorance of the users, who may not be able to understand completely the consequences of certain privacy settings of the platforms they interact with. For this reason, it is paramount to design and develop privacy mechanisms that are able to support every user in every context. I summarise the progress of the research community in this direction in Section 3.2.1.

Multi-party disclosure To design better mechanisms to manage individual privacy on OSNs is not sufficient to guarantee protection against insider threats. In fact, users' privacy may be violated by content that is shared not only by themselves, but also by other people in the network. This is the case of *multi-user* [174] or *interdependent privacy* [83] mismanagement, where someone (the uploader) may share some content that refers to other users (the co-owners), who may have different

[ii] See `https://help.instagram.com/116024195217477/?helpref=hc_fnav` - last accessed on 29/01/2022.

privacy preferences w.r.t. that content. Traditionally, these situations are referred to as *multi-user privacy conflicts* (MPCs). Despite most of the content shared on OSNs being co-owned by multiple users [84], OSN platforms currently offer very inadequate support for multi-user privacy [174], allowing only the uploader of the co-owned content to manage its privacy settings. When MPCs occur, co-owners can rely only on reparative solutions, such as untagging themselves or reporting the content as inappropriate, which are clearly unsatisfactory, because the privacy violations have usually been immediately perceived. A recent large-scale user study [175] showed that MPCs happen often —almost the totality of the 1033 participants had experienced at least one conflict within a year from the survey date— and are mostly due to the uploaders' inability to identify appropriate sharing policies for co-owned items, which confirms previous findings [28, 97, 207]. Although some MPCs do occur intentionally, like in the case of cyberbullism or revenge-porn [197], the vast majority of MPCs happens in *non-adversarial settings* [175] and could be avoided if mechanisms that preventatively recommend optimal sharing policies for co-owned content were available. For this reason, researchers have been investigating how to better support users to collaboratively manage access control, as I report on in Section 3.2.2.

2.2.2 In this thesis

The work described in this thesis addresses the challenges related to the multi-party aspect of the insider threat, in particular w.r.t. how to help the user select the most appropriate access control policy when sharing content online that involves other people. Notice that the models that I describe in the following parts of this thesis rely on the assumption, often formulated in the literature (see Section 3.2.2), of non-adversarial behaviour for the users involved in the MPC.

From a theoretical point of view, my work mostly builds on the conceptualisation of privacy given by Nissenbaum [129] as *contextual integrity*. Contextual integrity sees people not like undifferentiated individuals in a homogeneous world, but as individuals with certain roles in distinctive social contexts, allowing for diversified privacy preferences and expectations. In a hospital, the patient may disclose personal information with the medical doctor, who is expected to be discreet; among close friends, if one shares confidential details with another, there may be an expectation of reciprocity; if a medical doctor was friendly sharing personal details with the patient, it would be considered inappropriate. The possibility of defining social norms and rules that regulate the contextual disclosure of personal data is particularly helpful when reasoning in terms of autonomous agents, as I do in this thesis, given the extended literature available on *normative systems* (see, e.g., [47]).

Furthermore, given the focus on multi-user privacy of my work, I also draw from the *communication privacy management theory* by Petronio [137]. Based on empirical evidence in a variety of contexts, Petronio argues that individuals define boundaries to distinguish what is public from what is private. Communication of private information triggers expectations of behaviour on the recipient, who is considered as responsible as the communicator for what regards the information protection. That is, privacy boundaries need to be held collectively, through prior negotiation and agreement on the access control for that information. Hence, thinking about autonomous agents, the collective approach necessary to manage multi-user privacy may be modelled according to well-studied decision-making techniques in multi-agent systems, such as negotiation [58], social choice [149], and argumentation [140].

As a final note, in the analysis of OSNs, interpersonal relationships are naturally represented using social graphs, where nodes and edges represent respectively the users and their relationships. If the edges are labelled, e.g., detailing the intimacy or strength of the relationship, ReBAC policies can be replaced by *topology-based* *policies* [134], where access control is determined according to topological properties of the social graph, e.g., degrees of separation and minimum intimacy over the connecting path. In the models to support multi-user privacy that I describe in Chapters 4 and 5, I rely on topology-based access control, which can be translated from and to ReBAC [176], because it allows for maximum flexibility and granularity.

2.3 Theories of values

Human values have been studied and defined by many theorists and researchers (see, among others, [156, 148, 8, 80, 85]), who generally agree on their main features, which I report in the following.

- Values are beliefs which influence the way we feel about aspects of life and ideas; we get passionate when talking about something that we value, i.e. that we care about.
- Values refer to desirable goals, intended both as states of existence or ways of acting.
- Values are trans-situational, element that distinguishes them from norms and attitudes; if we value honesty, we consider it independently of the context (towards friends or strangers, at work or with family, etc.).
- Values are the standards according to which we evaluate actions, people and events.
- Values vary in importance across individuals and their relative order influences actions and behaviour.

Some psychologists believe that values always have a direct and explicit influence on behaviour (see, e.g., [8, 148]), while others consider that values impact on behaviour only occasionally, according to the individuals and the situations (see, e.g.,

[110]). However, some more recent studies revealed substantial correlations between most values and their corresponding behaviours [24]. It is believed that people act according to their values because of a need of consistency [148] or self-reward [24], in the sense that who pursues what they value is more likely to get what they want. Also, interestingly, values are considered relatively stable for each person and change little during adulthood [24, 148].

Given the impact that values have on what is considered acceptable to act upon, the inclusion of values in the design of an autonomous agent, whether this would act on our behalf or provide recommendations for decision-making, is crucial in order to identify actions that are coherent with our intentions and expectations.

2.3.1 The Theory of Basic Values

The Theory of Basic Values [155], first introduced by Shalom Schwartz in 1992, aims to measure universal values that are recognised across all major cultures. According to Schwartz, values are seen as socially desirable concepts that allow humans to interact between themselves, representing mental goals and the way used to describe and communicate such goals.

Schwartz identifies ten main values, fully described in Table 2.1 [24], which are interconnected and influence each other. These values can be organised in a circular structure (see Figure 5.3), where two dimensions summarise the main tendencies [154], defining four directions, or *hypervalues*, which pull apart while defining the behaviours. On one axis, *openness to change* is opposed to *conservation*, representing dynamic and independent ways of acting versus conservatory and self-restraining attitudes. On the other axis, *self-transcendence* reflects tolerant and altruistic behaviours in opposition to *self-enhancement*, which characterises authoritarian and image-conscious conducts.

Furthermore, this theory provides a framework that relates the ten values to be-

Value	**Definition (specific value items)**
Power	Social status and prestige, control or dominance over people and resources (social power, authority, wealth)
Achievement	Personal success through demonstrating competence according to social standards (successful, capable, ambitious, influential)
Hedonism	Pleasure and sensuous gratification for oneself (pleasure, enjoying life)
Stimulation	Excitement, novelty, and challenge in life (daring, a varied life, an exciting life)
Self-direction	Independent thought and action-choosing, creating, exploring (creativity, freedom, independent, curious, choosing own goals)
Universalism	Understanding, appreciation, tolerance and protection of the welfare of all people and of nature (broadminded, wisdom, social justice, equality, a world at peace, a world of beauty, unity with nature, protecting the environment)
Benevolence	Preservation and enhancement of the welfare of people with whom one is in frequent personal contact (helpful, honest, forgiving, loyal, responsible)
Tradition	Respect, commitment and acceptance of the customs and ideas that traditional culture or religion provide the self (humble, accepting my portion in life, devout, respect for tradition, moderate)
Conformity	Restraint of actions, inclinations, and impulses likely to upset or harm others and violate social expectations or norms (politeness, obedient, self-discipline, honoring parents and elders)
Security	Safety, harmony and stability of society, of relationships, and of self (family security, national security, social order, clean, reciprocation of favors)

Table 2.1: Description of the ten basic values in the Schwartz theory.

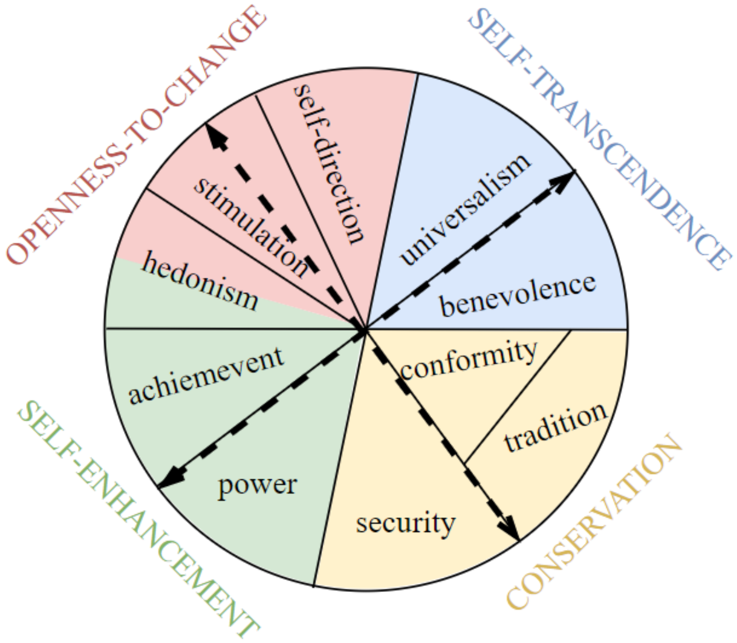


Figure 2.1: Values and value-dimensions in the Schwartz Theory of Basic Values.

haviours and that enriches analysis, prediction, and explanation of value-behaviour relations [157]. People take daily decisions according to the values they believe into. Even though values can compete with each other, the individual realises the dissonance and decides what to do by giving priority to some values over the others, that is by promoting a trade-off between competing values.

Schwartz suggests two main methodologies to measure the basic values: the Schwartz Value Survey (SVS) and the Portrait Values Questionnaire (PVQ).

Schwartz Value Survey The SVS presents two lists of value items: first, 30 nouns which describe desirable end-states, then 27 items which describe desirable ways of acting. This distinction between end-states and ways of acting follows from Rokeach [148], but it seems not to have substantive importance [155]. Respondents need to rate the importance of each item "as a guiding principle in their life" according to a 9-point non-symmetrical scale (7: "supreme importance", ..., 0: "not important", -1: "opposed to my values"), which reflects the way people think about values [154]. Each value is represented by 3 to 8 items and obtains as final score the average rating given to the corresponding items.

Portrait Values Questionnaire The PVQ is available in two versions and reports either 40 or 21 short verbal portraits of people. These portraits highlight a person's goals, aspirations or behaviours that implicitly correspond to one of the main values. Respondents are asked "How much like you is this person?" on a 6-point scale from "very much like me" to "not like me at all". Asking to compare themselves to a portrait forces the respondent to focus only on the similarities between themselves and the presented profile [154]. The PVQ manages to elicit a person's values without explicitly referring to the values. Each value is represented by 3 to 6 portraits and obtains as final score the average rating given to the corresponding portraits.

The PVQ, which implicitly elicits values through behavioural situations, requires from the respondents a more concrete and less cognitively complex task than the SVS, which demands more familiarity with abstract thinking [154]. For this reason, the PVQ is considered more appropriate for studies which consider a diverse population in terms of schooling background, and for studies that are completed online without direct guidance from an interviewer [156]. The PVQ-21 has been successfully used in many studies to date, including in the European Social Survey[iii], a biennial cross-national survey of attitudes and behaviour that has involved 38 different countries since 2002.

2.3.2 Other Values Scales

In the last century, many psychologists and social scientists attempted the definition of a formal framework of values that define human behaviour. In the following, I

[iii] https://www.europeansocialsurvey.org/

report an overview of the most known theories of values and of the tools that have been suggested to elicit human values. I include also a personality model, namely the Big Five, given the influence that personality traits can have on behaviour, despite their conceptual and empirical difference from values [147].

Allport and Vernon In 1914 Eduard Spranger theorised six ideal types of people, characterised by their predominant value attitudes, meant as beliefs, ways of thinking and patterns of living [130]. The six main values considered by Spranger are: *theoretical* (discovery of truth), *economic* (what is most useful), *aesthetic* (form, beauty, and harmony), *social* (seeking love of people), *political* (power), and *religious* (unity). Basing on Spranger's evaluation of the personality, in 1931 Allport and Vernon designed the Study of Values (SoV) as a psychological tool to measure personal values on the basis of declared behavioural preferences [8]. The SoV consists of 45 multiple-choice questions referring to alternative activities or occupations from which the respondent chooses the ones that are most appealing. Despite being for long among the most known studies to measure personality, the archaic content and inappropriate wording of the questionnaire have hindered its application in most recent decades [93]. In 2003 Kopelman et al. revised the questionnaire in order to make it more appropriate for the 21st century [93]. However, the refreshed version of the SoV is available to use only with explicit permission from the authors.

Rokeach In 1973 Milton Rokeach [148] distinguished two types of values: *terminal values*, such as happiness, equality, freedom, etc., which refer to desirable end-states of existence; and *instrumental values*, such as ambition, honesty, obedience, etc., which refer to preferable modes of behaviour. According to Rokeach, people generally reflect upon values considering them as absolute; however, human behaviour is driven by the relative importance of competing values in a given situation. Rokeach proposed the Value Survey[iv] in order to elicit the relative ordering of individual values: this presents 18 terminal values and 18 instrumental values and instructs the respondent to "arrange them in order of importance to YOU, as guiding principles in YOUR life"[148]. However, this does not seem the most appropriate way to elicit human values: in fact, the values as reported in the survey can be interpreted in a subjective way [71], leading to inconsistent rankings across participants.

Hartman In 1967 Robert Hartman [76] introduces the concept of formal axiology as a precise science to measure people's ability to value things in terms of appreciation of their properties. There are three hierarchical value dimensions, namely *systemic*, *extrinsic* and *intrinsic* values, that can be specified systematically with an objective scale of valuational richness, the Hartman Value Profile (HVP). The different combinations of the three value dimension codify emotions, motivations, and behaviours. The HVP consists of two parts with 18 phrases each, that the respondent needs to rank from good to bad according to their subjective sense of goodness and badness[v]. The maximum score in the HVP corresponds to the best and most rational value-vision, which most favours adaptation, survival and flourishing. More recently, Pomeroy [138] provided empirical validation for Hartman's formal theory.

Inglehart Ronald Inglehart was the founder and first president (1981-2013) of the World Values Survey[vi], an international research program devoted to the scientific and academic study of social, political, economic, religious and cultural values of people in the world. The project, which aims to analyse how people's values, beliefs and norms change across nations and over time, runs waves of representative comparative social surveys[vii] every five years. Basing on the data collected from

[iv] Complete survey available at http://faculty.wwu.edu/tyrank/Rokeach%20Value%20Survey.pdf

[v] For the complete survey see Figure 1 and 2 in [43]

[vi] https://www.worldvaluessurvey.org/wvs.jsp

[vii] Complete survey from the last wave (2017-2020) is available at https://www.worldvaluessurvey.org/WVSDocumentationWV7.jsp.

the World Values Survey, Inglehart hypothesised that socio-economic development brings systematic changes in basic values. In order to validate this hypothesis, he defined two major dimensions of cross-cultural variation: *traditional* versus *secular-rational* and *survival* versus *self-expression* [86]. For reasons of convenience such as data availability and consistency, only ten items from the survey are considered to score the two dimensions [85]. The Inglehart-Welzel World Cultural Map [85] shows how scores of societies are located on these two dimensions and, when comparing data from different survey waves, allows to track the socio-economical evolution of each society.

Hofstede Geert Hofstede started developing the Cultural Dimensions Theory (CDT) in the 80s while working in IBM and kept updating it until 2010. The CDT is a framework for comparing cultures by analysing six value dimensions: *Power Distance*, *Uncertainty Avoidance*, *Individualism/Collectivism*, *Masculinity/Femininity*, *Long/Short Term Orientation*, and *Indulgence/Restraint* [80]. Respondents answer six demographic questions and 24 content questions (four per dimension)[viii] based on individual perceptions of each one's own private life, job, organisation and society. The answers to the content questions show systematic differences across nationalities, but these should not be reflected on the personality of the individuals [80]. Hofstede's work focuses more on the anthropological aspect of values than on the psychological one, and therefore is less appropriate to be taken into consideration when designing value-aligned autonomous agents with the intention of supporting the individual's decision making process.

Big Five The Big Five model is a taxonomy commonly used for representing the human trait structure. Since Cattell in the Forties, several investigators have

[viii] Complete survey available at https://geerthofstede.com/wp-content/uploads/2016/07/VSM-2013-English-2013-08-25.pdf.

focused on applying factor analysis to personality survey data, highlighting the semantic associations people use when describe themselves or other people with nouns and adjectives [72]. Different markers have been identified, but there is a general traditional consensus on the following five basic factors: Neuroticism, Openness to Experience, Extraversion, Agreeableness, and Conscientiousness. Among the many tools that operationalise the Big Five framework, Goldberg [73] and the Oregon Research Institute present the *International Personality Item Pool*[ix], which collects over 250 scales of different length and accuracy; Costa and McCrae suggest the *Revised Neuroticism-Extraversion-Openness Peersonality Inventory* (NEO PI-R) [44]. Despite its popularity, the Big Five model has been object of serious criticism [135, 31], in particular regarding the lack of an underlying theory, the inappropriate exclusive use of factor analysis, and the disregard of other more private elements of personality (e.g., honesty, religiosity, etc.).

2.3.3 In this thesis

Given the user-centric objective of this thesis, I rely on the Schwartz Theory of Basic Values to design *value-aligned* autonomous agents for managing multiuser privacy online.

Although other alternative value theories present some desirable properties, in my opinion the Schwartz Theory seems the most suitable value theory for my purpose because is the only one that provides both (i) a structured theoretical framework, which describes how the similarity or diversity of values can influence behaviour, and (ii) reliable tools to elicit the values that have been previously applied in a variety of disciplines and empirical studies beyond the Social Sciences and Psychology. Other AI researchers before me have designed their autonomous systems based on the Schwartz values (see Section 3.3), because these allow decision-making systems

[ix] https://ipip.ori.org/

to recommend decisions more in tune with the preferences of the decision-maker [190], especially when considering the individual's priorities between perspectives in a certain decision context.

It is important to notice that, while the Schwartz values have been recognised across different cultures, the relationship between values and behaviour in privacy contexts that I have identified and applied in my models may not be universally valid and the design of my models may reflect a Western cultural bias (mine) [32, 212]. However, if a better value theory or more universal insights on mappings of values and behaviours were made available, the architectures of the value-aligned agent-based models that I describe in Chapters 4 and 5 could be effortlessly adapted without losing any of their properties.

2.4 Value Sensitive Design

In this section I introduce Value Sensitive Design (VSD) as the methodological approach I have applied throughout the development of the research work presented in this thesis.

The second half on the twentieth century has seen an intense technological progress in Information and Communication technologies (ICT). Van der Hoven [189] identifies three phases corresponding to different focuses and goals of technology: in the 60s-70s, technology was aiming mainly to solve problems; in the 80s-90s, developers realised the importance of usability and started including user requirements in their design; then, more recently, there has been an increasing interest in making technology positively impacting on the society, by including also human and social values in the design process. In fact, "The first question which should therefore be asked with respect to technology is whether it actually delivers the goods, whether it really contributes to the good life, however conceived" [189].

With this purpose of making technology more inclusive and aware of its users values and contribute to social well-being, VSD was initially developed in the early 90s by Batya Friedman and Peter Kahn. VSD is a theoretically grounded methodological approach that aims to include human values into the design, research and development of Information and Communication Technologies [66, 189, 49, 206]. Despite being applied mainly in ICT, VSD has driven works also more widely in all engineering and design disciplines (see [49] for an overview of VSD applications in the past twenty years), including intelligent autonomous agents [188].

In [67], Friedman and Kahn recommend a list of 12 values that have a particularly relevant ethical status and should be discussed w.r.t. the design of ICT: Human Welfare, Ownership and Property, Privacy, Freedom From Bias, Universal Usability, Trust, Autonomy, Informed Consent, Accountability, Identity, Calmness, and Environmental Sustainability. However, not only this is not considered to be an exhaustive list, as Friedman and Kahn point out, but new perspectives have recently emerged when discussing the inter-cultural validity and impact of values [32] (see Section 2.3 for a discussion of the state-of-the-art work on human values in the Social Sciences).

VSD provides specific instructions to follow during the design process [66] and this enables everyone, also those who do not have a philosophical or ethical background, to proactively integrate ethical reflections in the design of their work [49]. In particular, VSD recommends three steps [66]:

1. *Conceptual investigation*: this step concerns the identification of (i) the stakeholders, both direct and indirect, that are affected by the design at hand; and of (ii) the values that are relevant within the design. One should ask whether the values of interest are compatible, i.e., they can all be guaranteed equally for all the involved stakeholders, or whether trade-offs between competing values need to be specified (e.g., anonymity vs trust, usability vs privacy, autonomy vs control, etc.).

2. *Empirical investigation*: this step informs the designer about the user's perception of the technology at hand through both qualitative and quantitative methods. One should ask how the stakeholders, individually or at some organisational level, understand the values of interest and what they expect from the technology, e.g., in terms of priority over values. This is helpful to define the success of a particular design.

3. *Technical investigation*: this step presents a two-fold goal, namely (i) understanding how the design properties of existing technologies can hinder or promote values, and (ii) proactively design systems which support the values identified during the conceptual investigation. Regarding (i), the technical investigation differs from the empirical one in terms of the scope, which is now on the technology itself and no longer on the stakeholders.

These three investigations are interrelated and interdependent. Hence, VSD requires an *iterative approach* where each step informs the designer about the others and may redirect the overall design process.

2.4.1 In this thesis

While developing the project described in this thesis, I have followed a VSD approach and I have particularly benefited from its iterative methodology. In fact, given the user-centric focus of my work, it has been crucial to follow a research approach that would enable me to include some properties, identified as desirable by users and/or in the related literature, in the design of autonomous systems to support the collaborative management of multi-user privacy. Previously, VSD has been successfully applied by researchers in agents and multi-agent systems (e.g., see [33]).

Informed by the prior literature on privacy in online social networks, in a first iteration of conceptual investigation I have identified *adaptivity*, *role-agnosticism* and *value-alignment* as the main features to be embedded in a model for managing multi-user privacy online. Hence, during a first technical investigation, I have designed a preliminary model, that I describe in Chapter 4, where a negotiating agent supports equally all the users involved in a privacy conflict according to their contextual preferences, i.e., it is adaptive and role-agnostic, by recommending actions which are coherent with the user's morality, i.e., it is value-aligned. However, further technical investigation and additional insights from empirical studies in the literature highlighted some limitations of that model and the necessity for other features to be included in the model design.

Hence, in a second iteration of conceptual investigation, I have revisited the model requirements by adding the need to be also utility-driven and explainable (cf. Table 1.1 in Section 1.3). This has led to the design of a new agent-based model, which I present in Chapter 5, that presents all the requirements identified during the conceptual investigation. New technical and empirical investigations (see Chapters 6, 7 and 8) have shown the goodness of the new design in comparison with other state-of-the-art approaches for multi-user privacy management in OSNs, confirming the successful feature selection during the conceptual investigation and concluding the VSD process.

2.5 Practical reasoning

In this section, I summarise the approach to practical reasoning presented in the work of Atkinson and Bench-Capon [14, 13, 15].

Practical reasoning regards the deliberation of the best action to perform in a given circumstance. An argumentation scheme ArgS that describes the values that are promoted by performing an action can be considered as a prima-facie justification for performing that action [17].

ArgS: "In the current circumstances R, I should perform action A, to

CQ1: Are the believed circumstances true?
CQ2: Assuming the circumstances, does the action have the stated consequences?
CQ3: Assuming the circumstances and that the action has the stated consequences, will the action bring about the desired goal?
CQ4: Does the goal realise the value stated?
CQ5: Are there alternative ways of realising the same consequences?
CQ6: Are there alternative ways of realising the same goal?
CQ7: Are there alternative ways of promoting the same value?
CQ8: Does doing the action have a side effect which demotes the value?
CQ9: Does doing the action have a side effect which demotes some other value?
CQ10: Does doing the action promote some other value?
CQ11: Does doing the action preclude some other action which would promote some other value?
CQ12: Are the circumstances as described possible?
CQ13: Is the action possible?
CQ14: Are the consequences as described possible?
CQ15: Can the desired goal be realised?
CQ16: Is the value indeed a legitimate value?
CQ17: Is the other agent guaranteed to execute its part of the desired joint action?

Table 2.2: The critical questions (CQs) that can attack the argument scheme ArgS.

> bring about new circumstances S, which will achieve goal G and promote value V.”

However, ArgS can be intended only as presumptive justification for action and its soundness can be challenged by *critical questions* (CQs) [198, 14]. Negative answers to the critical questions, which I report in Table 2.2 as they are presented in [14], represent arguments attacking the original argument, i.e., the instantiation of ArgS, when considering the representation of the world, the desirability of the action, or the feasibility of the outcome.

In particular and among other elements, the feasibility of the outcome can be influenced by the actions of other agents involved in the same circumstances (see CQ17). The concept of *joint actions* [15], which are complex actions corresponding to a set of actions individually performed by a set of agents (with no requirement of cooperation or common purpose), is useful to represent this type of situation.

Definition 1. *A* ***joint action*** j_{Ag} *is a tuple* $\langle \alpha_1, ..., \alpha_n \rangle$*, where each* $\alpha_j \in Ac_j, j \leq n$ *represents the action of the agent* $ag_j \in Ag$ *(*$|Ag| = n$*). A joint action contains one, and only one action, for every agent in Ag. The set of all joint actions for the set of agents Ag is denoted by* J_{Ag}*.*

Example 1. *In the Ultimatum Game [75], the outcome of the game, i.e., whether the players keep the money, depends on both the action selected by the Player 1, i.e., to offer amount* $x \leq N$*, and on the action of the Player 2, i.e., to accept or reject the offer. In this case,* $J_{Ag} = \{\langle offer_x, accept_x \rangle, \langle offer_x, reject_x \rangle, \forall x \text{ s.t. } 0 \leq x \leq N\}$*, where Player 1 makes an offer x and Player 2 accepts/rejects it.*

In order to reason rigorously about joint actions, *Action-based Alternating Transition Systems* (AATS) were introduced by Wooldridge and van der Hoek [208] as a natural choice to represent an open agent system as a set of states, and actions as the transitions between them, where the outcome of an action can be influenced by what other agents decide to do. Atkinson and Bench-Capon [14] augmented the original AATS, which are formally based on alternating-time temporal logic, and defined AATS+V by considering how each transition could promote or demote certain values.

Definition 2. *An* ***Action-based Alternating Transition Systems with Values*** *(AATS+V) is a* $(2n + 8)$*-tuple*

$$S = \langle Q, q_0, Ag, Ac_1, ..., Ac_n, \rho, \tau, \Phi, \pi, Av_1, ..., Av_n, \delta \rangle,$$

where:

- Q *is a finite, non-empty set of states;*
- $q_0 \in Q$ *is the initial state;*
- $Ag = 1, ..., n$ *is a finite, non-empty set of agents;*

- *Ac_i is a finite, non-empty set of actions, for each $i \in Ag$ where $Ac_i \cap Ac_j = \emptyset$ for all $i \neq j \in Ag$;*

- *$\rho : Ac_{Ag} \rightarrow 2^Q$ is an action pre-condition function, which for each action $\alpha \in Ac_{Ag}$ defines the set of states $\rho(\alpha)$ from which α may be executed;*

- *$\tau : QJ_{Ag} \rightarrow Q$ is a partial system transition function, which defines the state $\tau(q, j)$ that would result by the performance of j from state q – note that, as this function is partial, not all joint actions are possible in all states (cf. the pre-condition function above);*

- *Φ is a finite, non-empty set of atomic propositions;*

- *$\pi : Q \rightarrow 2^\Phi$ is an interpretation function, which gives the set of primitive propositions satisfied in each state: if $p \in \pi(q)$, then this means that the propositional variable p is satisfied (equivalently, true) in state q;*

- *Av_i is a finite, non-empty set of values $Av_i \subseteq V$, for each $i \in Ag$;*

- *$\delta : Q \times Q \times Av_{Ag} \rightarrow \{+, -, =\}$ is a valuation function which defines the status (promoted (+), demoted (−) or neutral (=)) of a value $v_u \in Av_{Ag}$ ascribed by the agent to the transition between two states: $\delta(q_x, q_y, v_u)$ labels the transition between q_x and q_y with one of $\{+, -, =\}$ with respect to the value $v_u \in Av_{Ag}$.*

After providing the syntax for representing the world as states, and the actions of the agents as transitions between the states, I can proceed now to formally detail the three steps of the practical reasoning process: (i) problem formulation; (ii) epistemic stage; and (iii) choice of action.

Problem formulation The problem formulation involves the identification of the propositions and values which are relevant in the situation, and the consequent construction of the AATS+V. In this step, the soundness of AS can be contested by eight CQs, which highlight potential discrepancies between the agents' AATS+V representations (CQ2-CQ4, CQ12-CQ16).

Epistemic stage The epistemic stage regards the determination of the current state and of the joint action that will follow the agent's choice of a particular action. In order to identify the joint action, some assumptions may be required about the beliefs and the expected behaviour of the other agents involved in the situation. In this step, CQ1 and CQ17 need to be resolved.

Choice of action During the choice of action step, by instantiating ArgS and CQ5-CQ11, the agent develops arguments and counter-arguments in support of an action or another and evaluates them according to its value ordering. In particular, the agent considers alternative ways to realise the same consequence (in terms of reached state, achieved goal or promoted value), other consequences of the same action that can be desirable (other promoted values) or undesirable (demoted values or impediment for other positive consequences).

By completing the three steps of the practical reasoning and by collecting only negative answers for the CQs, the agent identifies the best action to perform.

2.5.1 In this thesis

After exploring alternative approaches for enabling autonomous systems to reason over values (see 3.3.3 for an overview of techniques employed in the related literature), I focused my work on value-based argumentation, and in particular on Atkinson and Bench-Capon's contributions, as these allowed me to define an agent-based model that could not only reason over potentially conflicting values, as it often happens in several domains including privacy decision-making, but also take into account the actions of others while deliberating.

In the model that I introduce in Chapter 5 and that I evaluate in Chapters 6 and 7, I define an agent, ELVIRA, that by performing practical reasoning is able to identify the sharing policy that is most coherent with the preferences of all the users involved in the MPC.

As I will detail later, I model the resolution of an MPC as a *joint action*, where the agent with the role of uploader can offer a sharing policy and the agents with the role of co-owners can either accept it or reject it. I adapt the definition of AATS+V and of the argument scheme AS so that the uploader can reason about the effects of offering a particular policy; by discussing the critical questions, these too adapted to the context, the uploader selects the policy that most satisfies both the sharing and moral preferences of all the involved users and that is believed to be accepted by the co-owners (epistemic assumption).

2.6 The social nature of explanations

In this section, I summarise Miller's work [112] on the insights from the Social Science that any AI researcher should consider when facing explainability challenges. Miller [113] alerts that, if AI researchers build explanations for themselves rather than for their users, they are going to fail in their task of making their system "explainable".

As depicted in Figure 2.2, eXplainable Artificial Intelligence (XAI) should not be tackled from a pure AI perspective, as it is mostly a human-agent interaction problem. Instead, it could benefit from the vast and mature body of work in Philosophy, Cognitive Science and Social Psychology, which highlight the following elements:

- Explanations are *contrastive*: people are generally interested in counterfactual cases, i.e. they do not wonder why some event X happened, but rather why the event X happened instead of some other Y.

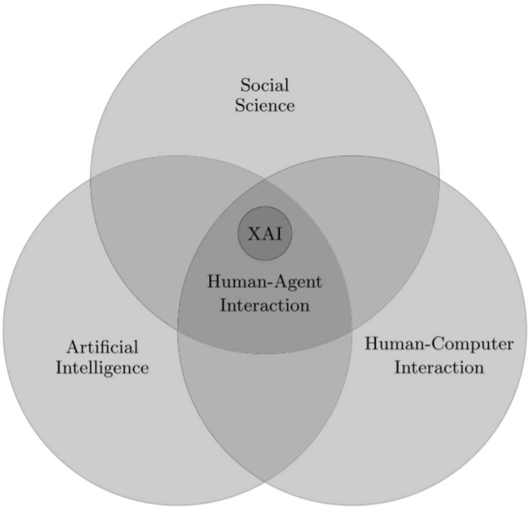


Figure 2.2: Scope of explainable artificial intelligence [112].

- Explanations are *selected*: when considering a causal chain of events that led to X, people are generally not interested in the complete list of preconditioning events, but only in a few that are subjectively considered more relevant.

- Explanations are *social*: explanations are best delivered as part of an interaction, where the beliefs of the explainee are to be taken into account by the explainer.

- *Probabilities* are not as important as *causal links*: to use statistical generalisations is not as effective as to refer to the causes of an event.

Miller [112] defines an explanation both as a product and as a combination of two processes:

1. the **cognitive process** is an abductive inference process to "determine an explanation for a given event, called the *explanandum*, in which the causes for the event are identified, perhaps in relation to a particular counterfactual case, and a subset of these causes is selected as the explanation (or *explanans*)";

2. the **social process** is the "process of transferring knowledge between explainer and explainee, generally an interaction between a group of people, in which

> the goal is that the explainee has enough information to understand the causes of the event".

In conclusion, to provide an explanation is not sufficient to identify the most likely chain of events that resulted in the event to be explained, but it is necessary to *tailor* it according to the context in which the explanation is delivered, the subject that will receive the explanation, and the purpose of sharing that explanation [170].

2.6.1 In this thesis

In Chapter 5 I introduce an agent-based model, ELVIRA, that is explainable according to Miller's definition.

In fact, as a consequence of applying practical reasoning, the agent is provided with a *cognitive process*, i.e., the ability of gathering all the necessary information to explain an event. Then, its *social process* is designed and validated empirically, by considering the needs, preferences and feedback of the users who are supposed to interact with the agent in the context of multiuser privacy management. In particular, ELVIRA is able to generate explanations, eventually contrastive or counterfactual ones, by providing arguments according to the argument scheme AS and the critical questions (as introduced in the previous section) that justify or attack the selection of each possible sharing policy as a solution to the conflict.

2.7 Conclusion

In this chapter, I have reported an overview of the background knowledge or techniques that I assume, apply or adapt in the following parts of this thesis.

First, I have described the multifaceted conception of *privacy* and its online threats. The insider threat in OSNs, that is the inappropriate disclosure of personal information within one's network, gives the underlying motivation to the work presented in this thesis. The agent-based models I will present to tackle the multi-party aspect of the insider threat build on the privacy theories of Nissenbaum (contextual integrity) and Petronio (communication privacy management).

Second, I have introduced the concept of *values* according to the Social Sciences and I have discussed alternative theories of values that can be embedded into the design of *value-aligned* autonomous agents. I have focused in particular on the Schwartz theory, because it provides (i) a useful overall structure of values, where values can compete or combine with each other when influencing the human behaviour, and (ii) reliable tools to elicit the value preferences from users.

Then, I have presented Value Sensitive Design as the methodological approach that has supported the development of my entire PhD work. By iteratively identifying the values, or properties, that are crucial to successfully solve MPCs in OSNs, and by investigating both technically and empirically the models I have designed to embed them, I have managed to design a user-centric autonomous agent that has collected positive feedback from real users.

Furthermore, I have summarised the practical reasoning approach of Atkinson and Bench-Capon, which represents a cornerstone for the definition of the main agent-based model that I introduce in this thesis. I have reported the most relevant definitions and notations from Atkinson and Bench-Capon's work in order to support the understanding of the adaptations and simplifications that I make later in Chapter 5.

Lastly, I have detailed the expectations for explainable artificial agents, according to the vast body of work in the Social Sciences. The importance of considering explanations as the combination of both a cognitive and a social process fits perfectly with the user-centric focus of my work, which is driven by the understanding of *what* is useful for the user and *how* to best deliver it.

In the next chapter, I will present an overview of the literature related to autonomous agents in the context of privacy, value-alignment and explainability.

Chapter 3

State of the Art

3.1 Introduction

Managing online privacy is problematic in our hyper-connected society, as I introduced in Chapter 1. Agent-based systems, where artificially intelligent entities support users while making their privacy decisions, have been suggested as a potential solution. However, embedding artificial intelligence (AI) in the society in a safe and trusted manner brings upon other challenges. In fact, in order to truly assist their users in whatever application we consider, autonomous systems (ASs) should be user-centric: they should act on the users' behalf according to their will and desires, i.e., *value-alignment* of ASs, and users should be able to understand their behaviour, i.e., *explainability* of ASs. In other words, to be effectively helpful for people, ASs need to cater for the users as a whole, by taking into account their human, fallible nature, while acknowledging the specific needs of each individual user and tailoring the support they offer. This means, on the one hand, that ASs should be aligned both with the values of humanity and with their users' expectation. On the other hand, ASs should be able to meaningfully engage with their users, by keeping them in the loop and explaining their processes and outcomes in a satisfactory way.

In this chapter I offer an overview of the approaches to manage privacy, value-alignment and explainability in AI, focusing in particular on the agent-based models and solutions. Research efforts in these directions have been prolific, but too often they have tackled only individual aspects of the overall challenge, i.e. there are plenty of models that either support privacy, or are value-aligned, or are explainable. First, in Section 3.2, I describe the agent-based models that have been presented in the recent years in order to help users of online services, mostly in online social networks (OSNs) contexts, manage their privacy. After outlining the models to preserve both individual and multi-user privacy, I analyse how the latter compare with the features required for an effective multi-user privacy support (cf. Table 1.1 in Section 1.3). Then, in Section 3.3, after introducing the value alignment problem and the related design challenges for ASs, I discuss recent studies that prove the increasing awareness of the community towards value alignment. In particular, I present research regarding (i) the elicitation of moral values, (ii) the automated reasoning over values, and (iii) the application of value-aligned agent architectures in several scenarios. Finally, in Section 3.4, after introducing the main concepts related to explainability in AI and the related design challenges for ASs, I report on a number of promising approaches towards explainable agents, with a particular focus on the models that, similarly to the work I present in Chapter 5, are argumentation-based.

3.2 Privacy-enhancing Agents

As introduced in Section 2.2, OSNs users can encounter a variety of privacy threats and have reported to be mostly worried by the *insider threat* [87], that is the inappropriate sharing of personal data within the one's network.

In order to tackle this, a prolific line of research has been developed towards helping users better manage their online privacy, independently of their privacy understanding and experience. In particular, scholars have advocated for more usable access control mechanisms, which would help users prevent inadvertent disclosure on OSNs. In Section 3.2.1, I present an overview of these mechanisms, focusing in particular on the agent-based models. In fact, if the social rules that regulate the disclosure of personal information [129] are modelled as norms, then autonomous agents defined in normative multi-agent systems could efficiently contribute to solving some of the privacy problems that users encounter online in general, and in OSNs in particular. The definition and implementation of *privacy personal assistants* [177] could, for instance, simplify the configuration of the privacy settings of the platforms the users want to interact with, by dynamically suggesting when and what to share according to the user's privacy preferences. Notice that, by learning the user's desired behaviour, an autonomous agent may enable everyone, including the more inexpert and risks-unaware users, to overcome the "privacy paradox" [2], i.e., the empirical inconsistency between desired and actual behaviour in privacy management.

Furthermore, an agent-based approach to the insider threat can be particularly helpful when tackling its multi-party component, i.e., when the user's privacy is violated by content inappropriately shared by other users in the network. In these situations of multi-user privacy conflict (MPC), the collective management of access control to personal information [137] is crucial to avoid privacy violations. Agents that are able to identify and recommend group-based optimal privacy policies have been presented in the literature, as I detail in Section 3.2.2.

3.2.1 Individual privacy

Most research in the area of usable access control mechanisms has focused on the automated recommendation of flexible and appropriate sharing policies. For instance, through active learning mechanisms that adapt to the user's behaviour [29, 57] it is possible to predict the desired set of privacy settings according to the content to be shared. Specifically in the context of photo sharing, some works recommend sharing policies based only on image features [168, 210], some consider also social graphs properties [146, 160, 118, 63], similar characteristics among users [10, 117], or the user's sharing history [169, 7, 100].

I refer the interested reader to extensive literature reviews on the topic (such as [62]) and focus next on agent-based approaches for individual privacy management. In fact, these could help users manage their online privacy [177] by recommending appropriate privacy settings or by directly acting on the user's behalf.

Kurtan and Yolum [96] introduce PELTE, an agent that recommends individual privacy decisions for images using tags. When the user's sharing history is not sufficient for predicting the correct policy for a new input, the agent considers the tags of all the images available in the user's network, modelling the users' tendency to mimic their peers in absence of clear preferences, which is a common dynamic described by the social learning theory.

Similarly, in [89] Kepez and Yolum present an approach that suggests privacy configurations by considering the user's previous posts and configurations. In this case, when not enough information is available, the agent relies on a multi-agent system architecture to aggregate the trust-weighted recommendations of other users' agents.

Misra and Such [116, 119] introduce REACT as a personal assistant agent that recommends customised access control decisions based on relationship type, relationship strength and content. This model achieved high accuracy in a user study, succeeding at minimising the user's effort in expressing their preferences.

Criado and Such [48] present a computational model of Implicit Contextual Integrity, where an agent uses the information model to learn implicit contexts, relationships and the information sharing norms in order to help users avoid sharing undesired data, while minimising their burden.

Finally, Ruiz-Dolz et al. [150] propose a preliminary argumentation-based approach to identify optimal sharing policies and generate explanations that help users understand the consequences of their privacy decisions. Starting from the user's behaviour on the network and the nature of the content, positive or negative arguments related to privacy, trust, risk and content are automatically generated and evaluated in an argumentation graph; after the acceptable arguments have been identified, an explanation in favour or against sharing the content is presented to the user.

3.2.2 Multi-user privacy

Given the high incidence of multi-user privacy conflicts (MPCs) [175] and the uploaders' inability to identify appropriate sharing policies for co-owned items [28, 97, 207], researchers have worked on designing collaborative models to support multi-user privacy management in OSNs: I discuss next the main approaches suggested so far, but refer the interested reader to reviews on the topic for more details and references [174, 134, 83]. Similarly to the work I present in this thesis, the models that I discuss in this section often do not detail how to detect MPCs and mostly focus on how to identify a solution after the MPC is detected. Notable exceptions are [81, 211, 176], whose detecting mechanisms could be preliminarily applied in combinations with other resolutive models.

Most of the proposed methods to solve MPCs are based on preference-aggregation techniques: in [182, 37, 81, 82, 145] the solution is identified mostly by majority voting; [161, 6, 173] introduce fuzzy rules for decision making, where factors such as content sensitivity, trust between co-owners and concession behaviour play a role; Xu et al. [209] describe a voting system where the co-owners' trust values, which are updated according to privacy loss, are used to weight the users' preferences.

Squicciarini et al. [166] suggest a system based on the Clarke-Tax mechanism, where users are incentivised to express truthful sharing preferences and are rewarded for promoting co-ownership when being truthful. Ulusoy and Yolum [185, 187] present a similar auction system, enriched with an abuse control feature and with agents that can learn the users' bidding strategies. In [141, 142] Rajtmajer et al. study the convergence of users' access control policies in multi-round and one-shot games, when assuming full or bounded rationality in the players.

In [60], Fogues et al. present a model where users are supported by learning agents which recommend sharing policies while considering contextual and preference-based features. The same authors suggest also another recommendation engine in [59], where different argument schemes prove to be very influential when identifying the optimal sharing solution. Ruiz-Dolz et al. [151] propose a model similar to the one for individual privacy mentioned earlier [150], where conflicts are solved by eventually persuading the uploader not to share the content through arguments extracted from the context and the involved users' preferences. In [91], Kökciyan et al. design agents which represent their users' sharing preferences through semantic rules and reach common sharing decisions using assumption-based argumentation.

Mester et al. [111] introduce an iterative negotiation mechanism where, through semantic rules, the co-owners can justify the eventual rejection of sharing offers to help the uploader suggest an acceptable policy. Kekulluogle et al. [88] extend this model by introducing different utilitarian strategies which reduce the uploader's disadvantage and consider social reciprocity. Utilities of a deal are also explicitly considered in the one-step negotiation protocol suggested by Such and Rovatsos [176].

Ajmeri et al. [5] introduce a value-aligned component in the context of data sharing, where a normative system allows agents to aggregate the users' value preferences to select appropriate actions. Other approaches based on normative systems are by Calikli et al. [36], where privacy norms based on the social identity theory are learnt adaptively for different contexts, and by Ulusoy and Yolum [186], where privacy decisions are made according to social and individual norms emerged from previous activities.

Not agent-based models Even though are not based on autonomous systems, there are some other interesting approaches, such as using cryptography [26, 131] or obfuscation through image processing techniques [84, 195, 131], that provide more fine-grained solutions to MPCs. In these cases, only specific authorised viewers have access to the content, which can be, eventually, altered for unauthorised users by cropping or blurring parts. Given that these mechanisms do not require an intentional collaboration among the involved users —that is, the users do not need to explicitly agree on a commonly acceptable solution—, if implemented in real OSNs, they would represent a promising answer also for those MPCs that occur in malicious contexts, such as revenge-porn and cyberbullism.

3.2.2.1 Comparison with requirements

All the approaches that I described so far present some strengths and show the community's interest and progress in making up for the insufficient support that OSN users currently receive when dealing with multi-user privacy. However, if we consider the required features for an effective MP support (cf. Table 1.1 in Section 1.3), then all these models reveal evident weaknesses and none of them presents all the required properties, as I summarise in Table 3.1.

Role-agnosticism is the requirement more commonly fulfilled in the literature. Most of the aggregation-based, the game theoretic, the learning and fine-grained approaches disregard the users' roles in the conflict and look at their preferences only. In the negotiation systems there is usually a clear distinction between the actions available to the uploader or the co-owners, but they still aim to identify a solution that is commonly acceptable.

The fine-grained approaches are clearly the most *adaptive* ones, allowing extreme flexibility for each privacy decision. The game-theoretic models, the learning-based approaches and the normative systems also permit to reach decisions which are very

Approaches	role-agnostic	adaptive	utility-driven	value-driven	explainable
game-theory	[141, 142, 166, 176, 185, 187]	[141, 142, 166, 176, 185, 187]	[141, 142, 166, 176, 185, 187]	-	-
aggregation	[6, 37, 81, 82, 145, 161, 173, 182, 209]	[6, 81, 161, 173, 209]	[81, 209]	-	-
human values	[5]	[5]	-	[5]	-
learning	[36, 60, 186]	[36, 60, 186]	-	-	-
argumentation	[59, 91]	[59]	-	-	[151], *
semantic rules	[88, 111]	[88, 111]	[88]	-	*
norms	[5, 36, 186]	[5, 36, 186]	-	[5]	*
obfuscation	[84, 131, 195]	[84, 131, 195]	-	-	-
cryptography	[26, 131]	[26, 131]	-	-	-

Table 3.1: Summary of the properties satisfied by previous approaches in the literature; * marks partial fulfilment of the property.

context-dependent. Some aggregation-based models, such as [37, 82, 145, 182], are generally not adaptive because of their rigid and static way of aggregating the users' preferences. Argumentation approaches [91, 151] tend to solve the conflicts following an "all-or-nothing" approach, that is by persuading a user to accept the requests of the other one, without looking for a middle ground solution.

The *utility-driven* requirement is surely fulfilled by the game-theoretical approaches, and by some other specific models [81, 209, 88] where the solutions are identified with the effort of maximising the users' utility, or to minimise their privacy loss.

Regarding the solutions which are *value-driven*, there is only [5]. However, there have been efforts of modelling real-world dynamics, where the users often concede and try to accommodate each other's preferences [175], such as reciprocity [88] and bounded rationality [141].

Finally, approaches based on argumentation [59, 91, 151], or that use semantic rules [111, 88] or normative systems [5] have the potential to support some type of *explainability* of the system, but none of these works autonomously generates explanations for their outputs and shares them with the users. There is one exception [151], where explanations are explicitly defined in the model, but in a static way, offering limited information, and without any empirical evaluation.

3.3 Value-aligned Agents

The deployment of artificial agents in the society presents a number of challenges, among which the value alignment is prominent, as I introduce in Section 3.3.1. *Machine Ethics* [183] is a field of AI research concerned with the issues of enabling autonomous intelligent systems to "behave ethically" when operating within our society. Fully ethical agents, such as average human adults, can make explicit ethical judgements and generally are competent to reasonably justify them [120]. But how can we design machines to be fully ethical agents? Researchers in AI have looked at this questions from different, complementary perspectives: in Section 3.3.2, I describe strategies that aim to identify the moral values at stake in different contexts; in Section 3.3.3, I summarise methodologies that enable artificial agents to reason about competing values in order to decide how to behave; and, finally, in Section 3.3.4, I report some examples of value-aligned agent architectures which showcase contextual reasoning on given sets of moral values.

3.3.1 Value-alignment problem

Given the pervasiveness of ASs in our day-to-day life and the increasing role that AI techniques play in supporting our daily decisions, concerns have been raised about the beneficiality humanity gets out of such technologies. There are innumerable scenarios where humans, after providing ASs with what seems to them very clear instructions, would be at minimum surprised, and in some cases seriously harmed, by the behaviour of the AS. For example, when we ask a self-driving car to bring us to the airport as fast as possible, we would not expect to reach our destination suffering from motion-sickness and chased by the police for not respecting the speed limit. When we tell our virtual home assistant that we want to cut on food to lose weight, we would not expect to find the smart fridge locked at our baby's meal time. If we ask our virtual personal assistant to book the most amazing holiday trip, we would not expect to overdraw our bank account. While these examples depict possibly exaggerated situations and can compare to those tales such as the Sorcerer's apprentice and King Midas [203], where magical agents fulfil wishes in literal ways leading to unsatisfactory and damaging results, they are helpful to understand some of the potential risks involved in the human/autonomous-agent interaction and cooperation.

Already for some years, the scientific community has been interested in understanding how to ensure that ASs are not only beneficial for humanity, but also that they do not develop destructive methods when pursuing their goals [34]. In 2017, during the 2017 Asilomar Conference on Beneficial AI, an interdisciplinary group of researchers defined the *Asilomar AI Principles*[i], a list of 23 issues that should guide the future development of autonomous systems, with a special focus on ethics and values. In particular, the Asilomar Principles include the *Value Alignment Problem* (Asilomar Principle 10):

> "Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation."

Related to this, the 11th Asilomar Principle on *Human Values* also recommends that:

> "AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity."

[i] https://futureoflife.org/ai-principles/ (last accessed on 03/06/2021).

Challenges for designing value-aligned agents Value-alignment of ASs can be realised at different levels, such as the architectural level and the actionable level. In the architectural level, I refer to the set of values that are *by design* acknowledged and explicitly considered when designing ASs. Achieving value-alignment would then mean to *include in the AS design all the human values that anyone could consider relevant* to the context. But how do we identify such an unanimous value domain? How do we translate the effect of each of these values onto the selection of possible alternative actions? In the actionable level, I refer to the subset of values that the user expects to influence the contextual decision process of the AS. Achieving value-alignment would then mean that *the values that are the most relevant to a user have the most influence on the AS decision process.* But how do we guarantee such a correspondence between individual preferences and the outcomes of automated reasoning processes? What do we do when the most preferred values of an individual would fire actions with contrasting effects?

3.3.2 Selecting moral values

The moral values to be embedded in artificial agents can be provided in a number of ways, for instance (i) by drawing from ethical theories of values, (ii) by learning from demonstration or imitation learning, or (iii) by directly interrogating the society.

The first approach relies on studies in Social Sciences, Psychology and Philosophy like the ones I discussed in Section 2.3. Ethical value frameworks can grant not only a definition of moral values, but also a description of the interrelation and influence between values. I speculate that this element in particular has pushed many researchers to rely on theories such as the Schwartz Theory of Basic Values, because the availability of a formal and complete system of values somewhat facilitates the modelling of reasoning layers for artificial agents.

The second approach aims to learn from labelled training data what values are relevant in given contexts or to given users. Despite multiple issues arise when considering value-learning through data [165], scholars have presented a few strategies for learning values from stories and, more generally, natural language corpora. For instance, Nahian et al. [127] show how to learn proper or improper behaviour from children's literature and allegorical tales. Even though values are not learnt explicitly, this work suggests that value-alignment can be learnt and generalised to new situations. As usual, special attention must be paid when selecting the corpora, and cultural bias must be acknowledged. In fact, when moving away from educational children literature, even just reaching consensus on labelling the data becomes challenging. For this purpose, Liscio et al. [103] propose a hybrid (human-AI) methodology to engage humans in identifying context-specific values, i.e., values that, differently from the ones defined in general theories of values (see Section 2.3), are or become relevant only in selected circumstances.

The last approach relies on mostly empirical studies which aim to understand which values are considered relevant by humans in different circumstances. To identify relevant values and gain consensus between ethicists, engineers and users is crucial to "educate" artificial agents to act according to humans' expectations. Moral dilemmas, which consist of comparing two options none of which is unambiguously morally preferable to the other, have been traditionally studied as a way of evaluating and comparing ethical decision-making. Bjorgen et al. [30] suggest to adapt this strategy for advancing Machine Ethics research and recommend the creation of a curated repository of moral dilemmas that can represent a benchmark for estimating the ethical performance of an autonomous system. This approach has already been proven successful to spark conversation about the ethical behaviour of AI in society: just think about the classic *trolley dilemma* [64] and its in-depth analysis w.r.t. self-driving cars in the Moral Machine experiment [18].

3.3.3 Reasoning over values

After defining what is a value and what are the values that are relevant in a context, an artificial agent needs to have a strategy to identify the best action to perform, where "best" is evaluated in terms of consistency with the values at hand. Research in this direction has proliferated mostly in the fields of argumentation and normative systems.

Bench-Capon [27] introduces the concept of value-based argumentation frameworks (VAFs), where arguments can be considered admissible by an audience but be rejected by another, according to their partial order over values. VAFs are crucial in multi-agent systems to identify actions that are consistent with the moral preferences of multiple stakeholders. Together with Atkinson [16, 14] (cf. Section 2.5), Bench-Capon also discusses moral problems through practical reasoning. In fact, an artificial agent can decide what is to be done according to its value preferences, which allow the evaluation of arguments and the consequent emergence of norms regarding what is morally correct, morally praiseworthy and morally excusable.

VAFs require assumptions about the values to be considered and about the influence of these values on the possible actions. Van der Weide et al. [191] introduce several argument schemes to reason about values in order to assign meaning to them, which allows to determine whether they are promoted or demoted in different circumstances.

Verheij [193] proposes to combine existing qualitative and quantitative methods to compare values and study their impact on decision-making. His suggestion is not based on abstract argumentation as Bench-Capon's work, but stays close to classical logic and standard probability theory. Information modelled from contexts, preferences and rules allows to generate presumptively and conclusively valid arguments with a conditional form.

Sierra et al. [162] discuss value alignment in the sense of alignment between the values (with a cognitive and consequentalist view) that one holds dear and the norms of the socio-technical system one is situated in. The authors first suggest how to aggregate individuals' and groups' preferences over single or set of values. Then, the alignment of the behaviour of an agent with the values at hand depends on the degree of preference for the state resulting from a norm.

Serramia et al. [158] model ethical decision-making in normative systems as a single-objective optimisation problem. Assuming a set of relevant values and a total order over them, the definition of a utility function supports the identification of the optimal norm system, i.e., the norm system with the maximum value support.

Liao et al. [102] suggest that normative systems and formal argumentation can be used for implementing a "moral council" that supports the decisions of an artificial agent while characterising the moral preferences of all the stakeholders involved, who can have contrasting values or even same values but contrasting arguments. Their agent architecture is also able to provide explanations for moral decisions in terms of justification and dialogue.

Without relying on argumentation or normative systems, Loreggia et al. [104] present a model to make decisions that are consistent with subjective preferences of the decision-maker and with exogenous preferences such as ethical principles. Both preference systems are represented through CP-nets, whose distance determines whether the deviation from the ethical principles is acceptable.

3.3.4 Contextual applications

Following the increasing awareness about the importance of designing systems that uphold the moral, societal and legal values of individuals and societies, scholars have presented value-aligned or value-driven agent architectures in several application scenarios.

Cranefield et al. [46] introduce in the context of elderly care robots a new plan selection mechanism in reactive planning. Following a Value Sensitive Design approach, they identify the Schwartz values that are relevant to the context and translate them into more concrete values that corresponds to actionable goals. As an example, the Schwartz value dimension 'self-enhancement' is translated into 'follow user's desires' which can be act upon by 'serving the desired meal'. This allows the authors to define plans based on a Belief-Desire-Intentions (BDI) architecture that can reason over alternative goals and competing values.

Still in the context of eldercare, Anderson et al. [11] present a case-supported principle-based behaviour paradigm where a number of ethical requirements are defined by experts and then actions are performed by the robot in order to maximise the compliance to those ethical requirements. According to its perceptions (e.g., 'low battery', 'no interaction', etc.), the robot can decide to implement different actions (e.g., 'charge batteries', 'engage with patient', etc.) in order to satisfy some overall duties (e.g., 'maximise good to patient', 'minimise non-interaction', etc.), which are treated as ethical values.

Following a Value Sensitive Design approach, Boshuijzen-van Burken et al. [33] describe an agent-based model of values in humanitarian logistics for refugees, where Schwartz values are operationalised and considered in terms of trade-off between the wellbeing of the refugees and the public opinion. In particular, they define sets of possible actions for each of the involved roles (e.g., newcomer, governmental, institutional or non-government institutional agents) which can temporarily promote one of Schwartz value-dimensions. Each agent is given individual thresholds defining their relative preference over the values; when the promotion of some value falls below its threshold, the agent is motivated to increased it again. This model is accessible through an interactive website, where decision makers can explore the impact and the consequences of governmental policies.

Heidari and Dignum [78] attempt to bridge the gap between the abstract Schwartz values and the definition of corresponding behaviours by presenting a framework to model values into decision systems. Specifically, they define some metrics that allow to compare alternative actions in terms of the values they promote, by taking into account the influence that values have on each other, as illustrated in the circular structure suggested by Schwartz. In order to translate the abstract Schwartz values into actionable ones, they use the value trees defined by [190], where values become more concrete when moving from the roots to the leaves. For example, in the context of a fishery village, which they consider to showcase their theoretical model, when considering whether to make a donation, 'to donate to council' promotes 'sustain village', which is a concrete value referring to the Schwartz 'tradition'.

3.3.4.1 Hybrid architectures

The increasing attention on Machine Ethics [183] and the inclusion of moral values into the design of autonomous agents should not lead researchers to ignore the utilitarian tradition of automated decision systems. In fact, utility-based optimisation provides an elegant and theoretically sound mechanism to identify the most convenient decision. By combining a careful and contextual definition of a utility function with a value-aligned architecture, the resulting agents may not only better mimic the human behaviour and better respect individual and social moral norms, but they may also recommend decisions that would be better than the ones identified by humans, thanks to their enhanced computational abilities.

Within this scope, Dehghani et al. [54] present a cognitively motivated model for decision-making which combines both a utilitarian and a deontological approach. When only secular values are involved in the decision, the agent aims to maximise utility. However, whenever moral values play role, the agent becomes less sensitive to the utilitarian outcome and prefers to act in a more ethical way. The combination of first-principles logical reasoning and analogical reasoning allows the agent to better mimic the complex human reasoning, as the authors show by reproducing previous psychological findings.

Santos et al. [152] introduce a multi-agent system to facilitate group decision-making by incorporating affective characteristics of the participants in the negotiation process. Each agent represents its user by reasoning about their mood, personality (coded with the Big Five factors) and emotions. The negotiation strategy can be adapted according to the interlocutors and their evolution throughout the interaction, strategically selecting whether to appeal, threaten or reward. Even though this model does not include values in a strict sense, it is interesting to see how these affective agents manage to reach an agreement much faster than non-affective ones, suggesting that architectural layers beyond the utilitarian ones can be very beneficial when seeking agreement in multi-user scenarios, which is one of the focuses of this thesis.

3.4 Explainable Agents

Explainable Artificial Intelligence (XAI) concerns the ability of artificial intelligent systems to generate solutions that can be understood by humans. Together with value-alignment, it is one of the crucial challenges for the safe deployment and use of AI systems in society.

After introducing the explainability problem in Section 3.4.1, when reviewing the state-of-the-art of explainable autonomous systems, in Section 3.4.2 I focus on the agent-based ones that explain their outputs through argumentation, similarly to what I do later in the thesis. In fact, argumentation is a transparent technique particularly suitable for providing explanations, justifying decision-making outcomes, and more. Then, in Section 3.4.3, I briefly survey other promising approaches to explainability. For more general reviews of explainability in AI, I refer the reader to [4], which summarises motivations, trends and research approaches to XAI; [101], which categorises existing works on data-driven and knowledge-aware XAI; and [172], which highlights the relevance of contrastive and counterfactual explanations and introduces a taxonomy in this regard.

3.4.1 Explainability problem

Explainability has been argued as a necessary component for the appropriate deployment of artificial agents in the society. In fact, explainability is believed to increase people's trust towards AI systems [204], enabling users to better discern whenever the model may be wrong or imprecise, especially in those contexts where people need more than a binary prediction to follow an automated recommendation. Other implications that follow from making AI explainable regard *accountability* [45] and *responsibility* [56], in the sense that the understanding of the reasons why an autonomous agent behaved in a certain way, or provided a certain output, simplifies the process of identifying liabilities. From a legal perspective, the right to an explanation has been acknowledged[ii] especially in those contexts where an automated decision-making process significantly affects an individual, such as the financial or legal ones.

Furthermore, explainability is essential for appropriate user-AI interaction and cooperation. As Kraus et al. [95] point out, explanations are relevant in situations where a user interfaces an AI system whose goal is known (e.g., a classification task in a neural network), but are even more crucial when the system's goal is not evident, i.e. when the system's output depends on multiple users' or agents' preferences not known a priori, such as in the management of online multi-user privacy. In these situations, explanations could increase user's satisfaction, for instance proving the alignment of the system with the expectations of the user.

Let me recall (see Section 2.6) that an explanation is the output of two pro-

[ii] See, for instance, General Data Protection Regulation (GDPR) and California Consumer Privacy Act (CCPA).

cesses [112]: (i) a *cognitive process*, which aims to gather all the information that is necessary for the explainer in order to explain something to the explainee, and (ii) a *social process*, which regards the way of transferring the information from the explainer to the explainee. Too often AI researchers focus on the cognitive process only and design explanations that are suitable for themselves but may fail to satisfy the users of their AI systems [113]. Hence, AI practitioners and developers should bear in mind that AI users may have needs and expectations regarding what an explanation looks like that are different (i) from their own and (ii) between each other. For instance, thinking of an AI-supported medical diagnosis, for an expert physician an MRI scan would suffice as a justification, while a patient may require further verbal elaboration in simple terms. However, what the designer of the hypothetical medical-support AI believes to be a good explanation is not necessarily helpful for that patient. For this reason, explanations need to be *tailored* to their recipient and to the context and their evaluation should focus more on the users' feedback than on technology [113].

Yet, despite the common urge to develop explainable autonomous systems, progress in this direction has been sometimes hindered by the interchangeable misuse of some terminology, as pointed out well by Arrieta et al. [12] and summarised in Table 3.2. In this thesis, I refer to explainability as a *design requirement* that improves the usability of AI tools rather than a post-hoc solution concept to 'black-box' models (e.g., deep learning).

Challenges for designing explainable agents In order to maximise the explainability of ASs, several aspects need to be considered during the design phase. First and foremost, the explanation of an autonomous decision should provide enough information to *allow the inference and comprehension of the process that has led to such decision.* This implies that the explanation should enable the user

- **Understandability** denotes the characteristic of a model to make a human understand its function – how the model works – without any need for explaining its internal structure or the algorithmic means by which the model processes data internally.
- **Comprehensibility**: When conceived for ML models, comprehensibility refers to the ability of a learning algorithm to represent its learned knowledge in a human understandable fashion. Given its difficult quantification, comprehensibility is normally tied to the evaluation of the model complexity.
- **Interpretability** is defined as the ability to explain or to provide the meaning in understandable terms to a human.
- **Explainability** is associated with the notion of explanation as an interface between humans and a decision maker that is, at the same time, both an accurate proxy of the decision maker and comprehensible to humans.
- **Transparency**: A model is considered to be transparent if by itself it is understandable.

Table 3.2: Nomenclature of most common XAI-related concepts as defined by Arrieta et al. [12].

to understand the space of possible outcomes and the relations between them. But how do we design such transparent cognitive processes? How do we facilitate the generation of contrastive and counter-factual explanations? Furthermore, the explanation should be *accessible and usable* by any recipient, for instance in terms of adopted technical language and background knowledge. But how do we design and implement explanations that are tailored to the contexts and to the explainees?

3.4.2 Argumentation-based explainable agents

Argumentation is having a strong positive impact on XAI. From a cognitive process perspective, by modelling an automated decision-making process as an argumentation procedure, it is possible (i) to analyse and justify every single step that led to its outcome, (ii) to reason under uncertainty and (iii) to take into account different points of views (audiences) who can present conflicting information. Moreover, when considering the social process, argumentation is naturally appropriate to convey the explanation, as different types of dialogues and interactions can be defined [109]. For an overview of how argumentation enables explainability in AI, the reader can refer to [192, 52].

Considering *AI planning* applications, which have been among the first to discuss and present explainable approaches [65], Cyras et al. [51] exemplify the use of abstract argumentation for modelling scheduling problems. Starting from a problem instance, they represent it through an argumentation framework, whose stable extensions are in one-to-one correspondence with schedules that are feasible, efficient and satisfying fixed decisions. The argumentation framework then allows for the extraction of argumentative explanations. The authors describe an implementation ('Schedule Explainer') of this model in [50] and showcase possible interactive explanations in makespan scheduling. The empirical evaluation of the model is planned as future work.

Shaheen et al. [159] use argumentation-based explanatory dialogue games to explain the decisions of a Satisfiability Modulo Theories solver in the context of treatment plans for multiple chronic conditions. The recommended plan is the search output of an optimal path across multiple graphs (one for each chronic condition to be considered). The dialogue protocols allow to identify supporting reasons for a given treatment, while justify the exclusion of others, by highlighting any potential adverse drug interaction. A user validation, not yet performed, will provide useful insights in terms of usability, credibility and dependability of the proposed dialogue game.

Regarding *recommendation systems*, models traditionally presented in the literature are often black-boxes that hardly present justification for their outputs. Rago et al. [139] describe a hybrid recommendation system for movie recommendations that is able to provide adaptive and interactive argumentative explanations. Considering the predicted user's rating, they build a tripolar argumentation framework (where neutralising relationships between arguments are considered in addition to supporting and attacking ones) giving a dialectical interpretation of the factors influencing a recommendation. Thanks to scaffolding argumentation, several types of explanation, e.g., textual, visual/tabular, interactive, etc., can be automatically generated with different levels of detail and presented to users. The quality and content of these argumentative explanations have been evaluated empirically through two user studies.

The researchers involved in the CONSULT project[iii] (see [41] and [92] among others) have suggested an argumentation-based *decision support system* to help patients suffering from chronic conditions self-manage their treatments. Their system can recommend appropriate actions after reasoning over a meta-level argumentation framework that includes different inputs such as sensors data, health records, and clinical guidelines. Users can interact with a dashboard or, dialogically, with a chatbot in order to gather justification and further explanation for the recommended treatment. A pilot study has been performed to evaluate the functionality and usability of the implemented model, but the quality of the explanations and of the interactions has not yet been analysed appropriately.

Looking at *AI and Law* applications, there have been a number of studies aimed at explaining and justifying legal decisions or procedures, to support either end-users or specialised workers when engaging with some regulations. Unfortunately, also due to the lack of available sources given the sensitive application contexts, these explainable model have not been evaluated in large-scale studies, and are mostly showcased in limited scenarios. For instance, Burgeemestre et al. [35] present a value-based-argumentation approach to reason about compliance with laws and regulations. Their argumentation framework links the abstract regulatory goals to concrete control measures, which can be explained and justified under the consideration of corporate values. In a case-study regarding EU custom regulations, the appropriateness of the obtained compliance decisions (more than their explanation) was evaluated by a group of experts.

[iii] `https://www.kcl.ac.uk/research/consult`

Wardeh et al. [200] propose two interactive tools for the purpose of policy consultations. In particular, by instantiating an argument scheme such as the practical reasoning (see Section 2.5) and some appropriate critical questions, arguments for the actions to be (or not) performed can be formulated. By engaging with the tools, citizens can efficiently seek justifications for and understand the policy proposals, eventually objecting them and making counter-proposals.

Collenette et al. [42] present an abstract dialectical framework to reason about legal cases within the European Convention on Human Rights. By navigating a tree representation of the legal case from the leaves (the relevant factors in the case) to the root (the verdict) through appropriate questions and answers, the autonomous system is able to accurately predict its outcome, and in a more adaptive way than similar machine learning approaches. Most importantly, the system complements the identification of the verdict with strong explanatory features which report the details of the performed reasoning.

3.4.3 Other explainable models

Despite argumentation being one of the most promising approaches to explainable AI [192], there are alternative lines of research that have shown interesting results.

Regarding planning applications, Winikoff et al. [205] model the agent's deliberation as a BDI process on goal trees, where behaviour is determined and explained according to desires, beliefs and valuings. An explanation for the action in a given node N of the tree is generated by traversing the relevant parts of the goal tree and includes the conditions for selecting the node at hand and the conditions for not selecting all the other options. The explanations can include details on preparatory actions, motivations and failure handlings. The authors provide an accurate empirical evaluation of usability (through user study) and efficiency (through software simulations) of their model.

Chakraborti et al. [39] argue that an AI system, rather than explaining the correctness of its plan and the rationale for its decision in terms of its own model, should aim to explain its process in terms of the human's model, who may have a different representation and interpretation of the problem at hand. This 'model reconciliation problem' has been validated with human-in-the-loop as described in [38]. For a (not planning oriented) approach to model reconciliation with BDI agents, see also the dialogue games presented by Dennis and Oren [55], where two interlocutors are able to identify if, and where, a divergence of views exists between them with regards to a BDI agent's operation.

Kim et al. [90] study the problem of inferring specifications that describe temporal differences between two sets of plan traces and present BayesLTL, a Bayesian probabilistic model for inferring contrastive explanations as linear temporal logic specifications. The generation of contrastive explanations was evaluated in a simulated realistic scenario and, even though they were somewhat simplistic, the explanations were consistent with an expert's interpretation of the events.

Sukkerd et al. [178] propose a method to automatically generate verbal explanation of multi-objective probabilistic planning. By comparing the generated plan with the possible alternatives in terms of trade-off between competing objectives, their algorithm reports contrastive explanations that justify the goodness of the identified plan. The authors show through a user study [179] that the generated explanations help users determine whether the agent's planning solution is aligned with their objectives.

Moving away from AI planning, Ghazimatin et al. [70] propose an explainable recommender system, namely Prince, that explores counterfactual evidence for discovering causal explanations in a heterogeneous information network, while protecting users' privacy. In particular, Prince is a provider-side mechanism that produces tangible explanations for end-users, where an explanation is defined to be a set of minimal actions performed by the user that, if removed, changes the recommendation to a different item. The authors justify formally the correctness of their algorithm and show empirically how users considered Prince's explanations more useful than the ones generated by other models.

The models described in this subsection, even though they are different from the argumentation-based approach I follow in this thesis to design explainable agents, share similar objectives as mine and present aspects that are crucial for explainability. First, they all aim to provide *tailored* and *contrastive* explanations, by considering the user's mental model of the problem or their values, preferences and experience. Then, they validate empirically their explainable mechanisms, appreciating that the feedback of end-users is paramount to evaluate explanations.

3.5 Conclusion

In this chapter I have reported the most relevant results presented in the literature towards automated management of online privacy, the value-alignment problem and the explainability challenge in AI in general, and in agent-based models in particular.

Regarding privacy-enhancing agents, the models outlined in Section 3.2.1 are complementary to the work I present in this thesis. They help the users identify their *individual* privacy preferences; then, if conflicts are detected, the models I introduce in Chapters 4 and 5 could support them to identify the optimal collective sharing policy. The models I discuss in Section 3.2.2 share the same objective as this thesis, namely to support the management of multi-user privacy (MP). However none of them is successful in presenting all the features necessary for an effective support of MP that I have identified through conceptual investigation within the Value Sensitive Design approach. Specifically, the properties that more rarely are satisfied in the literature are to be value-driven and explainable: these are exactly the features that I aimed to promote in the models I introduce later in this thesis.

Value-alignment is crucial to guarantee that the agent's behaviour is consistent with the user's desire and expectation. In Section 3.3 I have summarised the research efforts in this direction, in particular regarding (i) the elicitation of relevant moral values, (ii) reasoning about competing values to identify appropriate actions, and (iii) contextual applications of value-aligned agents in different scenarios. In this thesis, I introduce two privacy-enhancing agent-based models that are value-aligned (see Chapters 4 and 5). In order to do so, I draw upon existing theories of values, namely Schwartz's theory (illustrated in Section 2.3.1), to select the relevant moral values, which I then map onto actions and behaviours in the context of online privacy according to insights from the literature and common sense. The subjectivity of this mapping process, common to the majority of value-aligned systems, however, is among the main limitations of this type of agent architectures.

Finally, in Section 3.4 I have provided an overview of the explainable agent-based models, with a particular focus on the ones that rely on argumentation. In fact, argumentation guarantees several properties, such as transparency, the ability of reasoning under uncertainty and to consider contrasting information, which simplify the design of intrinsically explainable architecture. Moreover, argumentation-based dialogue games can be defined in order to enable the interaction that is deemed as necessary by the definition of explanations as social processes. Yet, many XAI researchers may fail to design explainable systems that cater for the needs of their users, as the lack of empirical validation of these models highlights. In order to avoid this, the explanations generated by the agent model I describe later in Chapter 5 are not only designed being informed by users' feedback, but are also empirically validated to assess their helpfulness for users.

Chapter 4

JIMMY: A Value-aligned Negotiating Agent

4.1 Introduction

While following a Value Sensitive Design approach (VSD, see Section 2.4), through conceptual and technical investigations of the literature on online multi-user privacy (MP) I noticed a discrepancy between the way users manage MP in the real world and the way scholars have tried to support users with their theoretical models. Empirical studies [175, 199, 97] showed that often users are not interested only in their own social benefit when sharing content online, but that they care, eventually too late, about the consequences that their decisions may have on others; in particular, users reported to regret not to have asked for permission before sharing co-owned content, and many were willing to compromise and find a solution that was acceptable for all the involved parties. However, many of the mechanisms that have been suggested in the literature (see Section 3.2.2) fail to consider this type of behaviour, where users may disregard their own advantage in favours of others, and mainly recommend the sharing action which guarantees the best outcome in terms of sharing utility.

For this reason, I aimed to bridge the gap between the MP literature and the empirical evidence by designing a *model that*, instead of focusing solely on the sharing outcome, *supports the interaction* of the users involved in a multi-user privacy conflict (MPC). Informed by a first iteration of VSD conceptual investigation, I concluded that an autonomous system, in order to effectively support MP, should present the following features:

- *Role-agnosticism*: the model should treat in the same way all the parties involved in the conflict; this is because the asymmetricity of the access control mechanisms currently available to uploaders and co-owners is considered one of the main causes of MPCs [207].

- *Adaptivity*: the model should generate solutions to MPCs that are consistent with the contextual preferences of the involved users; this is because privacy is highly subjective and managed in different ways according to the scenarios [2, 129].

- *Value-alignment*: the model should provide recommendations that are consistent with the moral and behavioural preferences of the users; this is because, as empirical evidence shows, users are often willing to transcend their own interest in order to favour others and avoid conflicts [175].

Hence, I designed JIMMY, an agent-based model that acts like the *Jiminy Cricket* from *Pinocchio* and helps its users, when interacting with each other, pick the actions that are most aligned with their moral values. The users' interaction for the resolution of an MPC is modelled as an audience *negotiation*, where JIMMY equally supports each user by recommending whether to accept/reject an offer or make a counter-offer, in a consistent way with their moral values.

In Section 4.2, I introduce some preliminary concepts that are necessary to define the possible actions that the JIMMY agent can recommend. In Section 4.3, I describe the dynamics of the negotiation, including the details for generating value-aligned offers, and I report a complete example of a negotiation. In Section 4.4

I show some formal properties of this agent architecture, namely termination in a finite time and soundness. In Section 4.5 I discuss the advantages and disadvantages of JIMMY, specifying how these have influenced the next steps of my research, and I conclude the chapter with Section 4.6.

4.2 Preliminary concepts

In order to recommend negotiating actions that are coherent with the users' desires, the agent JIMMY needs to be informed about (i) the user's preferred privacy level for the content in discussion and (ii) the user's preferred order over a set of moral values. Both these factors can be elicited with minimal user intervention as detailed below.

4.2.1 Sharing policies

I represent a collaborative platform (e.g. an OSN) where users are supposed to interact with each other and share content as a graph $G = (V, R)$, where V is the set of the OSN users, and R describes all their relationships $(v_k, v_j, i_{kj}) \in R$; $i_{kj} \in \{1, ..., i_{max}\} \subset \mathbb{N}$ represents the closeness or intimacy of the relationship between two users v_k and v_j (1 denotes a superficial connection, i_{max} denotes the maximum intimacy) and can be elicited by using predictive techniques as presented in [61].

Content is shared in the platform according to sharing policies, which define the criteria users must satisfy in order to access such content[i].

Definition 3. *A **sharing policy** is a tuple $p = \langle d, i \rangle \in \mathcal{P}$, where $d \in [0, d_{max}]$ represents the maximum allowed distance a user must be from the owner of the content, meant as the length of the shortest path of the social graph connecting the*

[i] Although new relationships can emerge and old ones evolve or dissolve, i.e. links and intimacy between users may vary over time, this is irrelevant w.r.t. the definition of the sharing policy: some user who today has access to some content, may lose it in the future and vice-versa.

two users, and $i \in [0, i_{max}]$ represents the minimum required intimacy over each edge of the path connecting the two users.

Note that the policy definition used in this chapter could be translated to and back from the usual group-based access control policies of social media sites like Facebook [173]. Also, note that individual privacy preferences for each item, i.e. the sharing policy that each of the involved users would select if they were to decide alone about the item, can be provided directly by the user or can be elicited following data-driven AI techniques as in [167, 116], to minimise user effort, or can be derived from suitable defaults based on approaches like [201].

4.2.2 Schwartz Basic Values

Based on the Schwartz Theory of Basic Values (see Section 2.3.1), I identify five values, or groups of similar values, which are relevant in the context of collective decision-making processes. The individual relative preference over these values is proven to be relatively stable over the lifetime [24]; this suggests that their elicitation, for instance using the Schwartz Value Survey or the Portrait Values Questionnaire [154] (see Section 2.3.1) may not be a constant or repeated burden for the user.

For each of the identified values, I interpret its influence on the negotiation behaviour as follows:

- *self-direction (sd):* the user is open-minded and ready to change the negotiating strategy during the decision-making process to suggest new solutions;
- *power (po):* the user holds his/her initial sharing preference with higher regard than the one of the other user, leaving little space to accommodate the others' preferences;
- *benevolence, universalism (be):* the user is willing to accommodate the preference of another user;

- *security (se):* the user prefers the safest sharing policy, meant as the most restrictive one in terms of publicity;
- *conformity, tradition (tr):* the user's choice is highly influenced by the society's expectations.

As I mentioned, JIMMY recommends actions during the negotiation according to how the user would like to behave in that context. Given the influence that moral values have on behaviour [24], JIMMY considers the relative importance of each value for the user as the factor that most influences the user's negotiation strategy. I formalise the relative importance of the values for each user as follows.

Definition 4. *A **value order** is a particular order $o \in \mathcal{O}_{\mathcal{V}}, o = v_1 \succeq v_2 \succeq v_3 \succeq v_4 \succeq v_5$ over the values $v_i \in \mathcal{V}$, where $\mathcal{V} = \{be, po, se, sd, tr\}$, that determines the relative influence the user believes each value has on his/her behaviour, where $\mathcal{O}_{\mathcal{V}}$ is the space of all the possible total or partial orders over $\mathcal{V}$.*

Example 2. *Emma has a total order over her values $o_E : po \succ se \succ sd \succ tr \succ be$, meaning that she considers power as the most important value to guide her behaviour, followed by security as the second most influential and so on.*

Note that, while the example shows a total order, JIMMY could work also with partial orders, e.g., users having some preferences of values over the others but not for all values, as I will detail later.

4.3 The negotiation process

The communication between the users involved in the MPC follows a number of steps (or negotiation rounds) until a decision is agreed upon. For simplicity, I consider an alternated proposals negotiation framework [58], as depicted in Figure 4.1, and a MPC that involves two users, one with the role of *uploader* (A), the other with

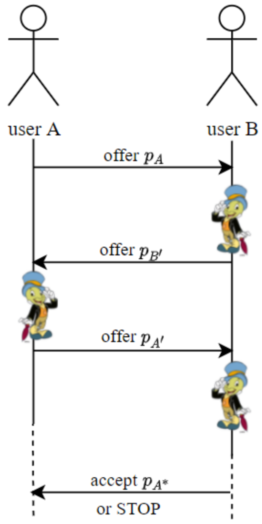


Figure 4.1: Message Sequence Chart showing the negotiation process between the uploader A and the co-owner B.

the role of *co-owner* (B). Each user is supported by one independent instance of the JIMMY agent during the negotiation.

In the first iteration of the negotiation, the uploader A starts the dialogue offering his/her preferred policy p_A to the co-owner B. In the second iteration, B evaluates the received offer: if B accepts, then the negotiation is concluded and the content is shared with the policy p_A; if B is not satisfied by the offer, then B can make a counter-proposal, which is evaluated by A in the third iteration and so on. However, an agreement may be impossible to reach, for instance if both parties keep offering the same policy without trying to accommodate each other. In this case, which I detail at the end of the next subsection, the negotiation is considered as failed and no content is shared.

In order to support the user while interacting with the other party during the negotiating process, at every iteration JIMMY suggests to its user whether to accept the offer or to make a counter-offer, i.e. the agent always recommends the negotiating action that is the most coherent one with their moral values.

4.3.1 Generation of a new offer

I now introduce the elements that allow the agent JIMMY to identify the negotiating action that is the most coherent with its user's values at each negotiation round. In a nutshell, each moral value (self-direction, power, benevolence, security, and tradition – see Section 4.2.2) is mapped onto a negotiating behaviour, which is translated in a function that influences either the overall negotiation strategy or each single iteration of the negotiation.

First, I model the behaviour related to *self-direction* as a function that influences the negotiating strategy of the user. This is achieved by modifying the user's value order, with the aim of reflecting the influence of being creative and open-minded while looking for a solution.

Definition 5. *The* ***sd-function*** $f_{sd} : \mathcal{O}_{\mathcal{V}}, T \longrightarrow \mathcal{O}_{\mathcal{V} \smallsetminus \{sd\}}$ *defines the influence of the value self-direction over the entire user's strategy. Considering the order* $o \in \mathcal{O}_{\mathcal{V}}$ *provided by the user* X*, it returns another order* $\tilde{o} \in \mathcal{O}_{\mathcal{V} \smallsetminus \{sd\}}$ *where, according to the negotiation turn* $t \in T$ *of* X*, either the value self-direction is removed or the value self-direction is removed and the two following values, if any, are swapped, in order to generate a dynamic negotiation strategy.*

$$\tilde{o} = f_{sd}(o, t) = \begin{cases} removeSD(o) & \textit{if } t \textit{ is odd} \\ removeSDswap(o) & \textit{if } t \textit{ is even.} \end{cases}$$

The sd-function allows the agent, and therefore its user, to be flexible during the negotiation process by eventually employing two strategies alternatively, i.e. two different orders over benevolence, power, security and tradition.

Example 3. *Let's recall that the preferred order over the values for Emma is* o_E : $po \succ se \succ sd \succ tr \succ be$. *The first time* f_{sd} *is activated, i.e., the first time JIMMY needs to recommend an action for Emma,* f_{sd} *generates* $\tilde{o} = removeSD(o_E) = po \succ se \succ tr \succ be$*, where self-direction has been removed from Emma's original value* *order. The next time f_{sd} will be activated, it will produce $\tilde{o} = removeSDswap(o_E) = po \succ se \succ be \succ tr$, where self-direction has been removed and benevolence and power, which were the two values following self-direction in the original order, have been swapped.*

Regarding the behaviours corresponding to the other values, i.e., *power*, *benevolence*, *security* and *tradition*, instead of impacting the overall negotiation strategy, these values influence each negotiation step by collectively altering the generation of a new policy offer. In order to track the sequential influence that each value-function has on the definition of the final new policy to be offered, the value-functions take in input a "placeholder" policy p_v, which memorises the outcome of the other value-functions previously combined (see Definition 7 for clarification on this).

Definition 6. *A **value-function** $f_v : \mathcal{P}^3 \longrightarrow \mathcal{P}^3$ defines the cumulative effect of the value v on the generation of the new sharing policy to be suggested. It takes as inputs the preferred policies p_A, p_B of the two users (or their last offered policies), and p_v, a placeholder policy which keeps track of the influence of the other value-functions that have already been considered. The co-domain of f_v is a subset of its domain: i.e. given three policies p_A, p_B, p_v and defining*

$$p_1 = \langle \min(d_A, d_B, d_v), \max(i_A, i_B, i_v) \rangle = \langle d_1, i_1 \rangle$$

$$p_2 = \langle \max(d_A, d_B, d_v), \min(i_A, i_B, i_v) \rangle = \langle d_2, i_2 \rangle$$

as the tuples having respectively the most and the least restrictive components over p_A, p_B, p_v, then

$$f_v(p_A, p_B, p_v) \in ([d_1, d_2] \times [i_2, i_1])^3. \tag{4.1}$$

A possible instance of the f_v functions is reported as follows, but alternative definitions could be suggested:

- $f_{po}(p_A, p_B, p_v) = (p_A, p_B, avg(p_A, p_v))$: *power* influences the new sharing policy by averaging it with the user's own preferred policy;

- $f_{be}(p_A, p_B, p_v) = (p_A, p_B, avg(p_B, p_v))$: *benevolence* influences the new sharing policy by averaging it with the interlocutor's preferred policy;

- $f_{se}(p_A, p_B, p_v) = (p_A, p_B, avg(\langle \min_{p_A, p_B} d, \max_{p_A, p_B} i \rangle, p_v))$: *security* influences the new sharing policy by averaging it with the most restrictive distance and intimacy, computed over the two users' preferred policies;

- $f_{tr}(p_A, p_B, p_v) = (p_A, p_B, avg(p_\sigma, p_v))$: *tradition* influences the new sharing policy by averaging it with the socially preferred policy p_σ, i.e., the policy that a majority of people would select for sharing some content with the same sensitivity, which can be automatically elicited (see, e.g., [60]).

Note that, since the distance and the intimacy must be integer numbers, rounding is performed towards the one own policy in general, and towards the other user's policy when the value-function is f_{be}.

Finally, I can now introduce the crucial part of the model, that is the generation of a new proposal. At each step of the negotiation after the first one, a user receives an offer that needs to be evaluated. To do so, JIMMY first generates for the user what would be his/her optimal next counter-offer, given his/her preferred policy and order over values.

Definition 7. *A proposal generator $g : \mathcal{P}^3 \longrightarrow \mathcal{P}$ is a function which provides the policy $\tilde{p}$ that the user should offer according to his/her preferred value order, his/her policy preference (which is either the initial one or the last generated one), and the last received offer. In particular, it is a composition of all the value-functions $f_v \in \mathcal{F}$, where the order of the composition is given by the output order $\tilde{o} : v_1 \succ v_2 \succ v_3 \succ v_4$ of the sd-function, and the projection operator Π_3, which selects only the third policy from the last obtained tuple:*

$$\tilde{p} = \Pi_3 \circ f_{v_1} \circ f_{v_2} \circ f_{v_3} \circ f_{v_4}(p_A, p_B, null).$$

Notice that the third policy of the last obtained tuple corresponds to the resulting effect of the last value-function applied (f_{v_1}) on the last placeholder policy p_v, which in turn represents the aggregated effects of all the previously applied value-functions ($f_{v_4}, f_{v_3}, f_{v_2}$) when considering the users' preferred policies p_A and p_B. In case the value order is a total order, then the composition order of the value functions is trivial (see a complete example in the next section). If the value order is a partial order, different solutions can be applied: for instance, a possibility could be to pick randomly only one of the value-functions for each group of equally preferred values.

After identifying the counter-offer that would be most coherent with the user's desired behaviour, JIMMY computes the distance between the newly-generated potential counter-offer and the last received offer.

Definition 8. *The distance ϵ_{p_A,p_B} between two policies $p_A = \langle d_A, i_A \rangle$ and $p_B = \langle d_B, i_B \rangle$ is defined as the Manhattan distance:*

$$\epsilon_{p_A,p_B} = |d_A - d_B| + |i_A - i_B|.$$

If such distance is within a reasonable range, which can be left as a default value or decided by the user, e.g., it is equal to 0 so that the policies are the same, then JIMMY recommends the user to accept the offer and the negotiation ends. Otherwise, JIMMY suggests to offer $\tilde{p}$ as a counter-proposal and the dialogue proceeds until either the convergence is reached or until it is recognised as impossible, i.e. when both users have tried out all their strategies (that is, different composition orders generated by the sd-function) and cannot help but keep suggesting the same policies. If an agreement is found, then the item in discussion is shared according to the last offered policy; alternatively, the content remains private.

Recall that, at each step of the negotiation, the choice of the action to perform is left to the user, in respect of his/her autonomy [194]; JIMMY simply nudges the user to dialogue with the other in a way which is consistent with his/her moral values, similarly to the Pinocchio's Jiminy Cricket.

4.3.2 Example of a JIMMY-supported Negotiation

Let us examine a situation where users Alice (A) and Bob (B) discuss about sharing some content on an online collaborative platform. Their preferred policies are respectively $p_A = \langle 5, 8\rangle$ and $p_B = \langle 1, 1\rangle$ and a MPC occurs. According to the sensitivity of the content, in general people would suggest the policy $p_\sigma = \langle 2, 6\rangle$.

While negotiating a sharing policy that is acceptable for both, both Alice and Bob are supported by an instance of the JIMMY agent, respectively JIMMY$_A$ and JIMMY$_B$, which each has access to its user's preferred order over the values:

$$o_A : se \succ_A be \succ_A tr \succ_A po \succ_A sd$$
$$o_B : be \succ_B tr \succ_B se \succ_B sd \succ_B po.$$

Here, both users have a single strategy, because *self-direction* is in the last or second-last position in the order and therefore f_{sd} has no values to eventually swap.

Let us consider the instances of the value-functions listed earlier, where $p_v = null$ at the beginning of each negotiation step s. Let us assume that in order to accept a received offer, the distance between the received offer and the newly generated offer must be $\epsilon = 0$.

Finally, for the purpose of this example and for the sake of brevity, let us assume that each user accepts his/her agent's recommendation at every step.

At $s = 0$, JIMMY$_A$ suggests to offer Alice's original preferred policy $p_0 = p_A = \langle 5, 8\rangle$.

At $s = 1$, Bob needs to evaluate whether to accept or reject the offer. To do this, JIMMY$_B$ computes the best (according to Bob's values) policy that Bob would eventually counter-offer and then, if this coincides with p_0, JIMMY$_B$ would recommend to accept:

$$\begin{aligned}g(p_A, p_B; o_B) =&\Pi_3 \circ f_{be} \circ f_{tr} \circ f_{se} \circ f_{po}(p_A, p_B, null)\\ =&\Pi_3 \circ f_{be}(f_{tr}(f_{se}(\langle 5, 8\rangle, \langle 1, 1\rangle, \langle 1, 1\rangle)))\\ =&\Pi_3 \circ f_{be}(f_{tr}(\langle 5, 8\rangle, \langle 1, 1\rangle, \langle 1, 4\rangle))\\ =&\Pi_3 \circ f_{be}(\langle 5, 8\rangle, \langle 1, 1\rangle, \langle 1, 5\rangle)\\ =&\Pi_3(\langle 5, 8\rangle, \langle 1, 1\rangle, \langle 3, 7\rangle)\\ =&\langle 3, 7\rangle.\end{aligned}$$

Since $\epsilon_{p_A,p_B} = |5 - 3| + |8 - 7| > 0$, JIMMY$_B$ suggests Bob to reject the offer and to propose $p_1 = \langle 3, 7\rangle$.

At $s = 2$, it's Alice's time to evaluate Bob's offer:

$$\begin{aligned}g(p_A, p_1; o_A) =&\Pi_3 \circ f_{se} \circ f_{be} \circ f_{tr} \circ f_{po}(p_A, p_1, null)\\ =&\Pi_3 \circ f_{se}(f_{be}(f_{tr}(\langle 5, 8\rangle, \langle 3, 7\rangle, \langle 5, 8\rangle)))\\ =&\Pi_3 \circ f_{se}(f_{be}(\langle 5, 8\rangle, \langle 3, 7\rangle, \langle 4, 7\rangle))\\ =&\Pi_3 \circ f_{se}(\langle 5, 8\rangle, \langle 3, 7\rangle, \langle 3, 7\rangle)\\ =&\Pi_3(\langle 5, 8\rangle, \langle 3, 7\rangle, \langle 3, 8\rangle)\\ =&\langle 3, 8\rangle.\end{aligned}$$

Since $\epsilon_{p_A,p_1} = |3 - 3| + |8 - 7| > 0$, JIMMY$_A$ recommends to reject the offer and to propose $p_2 = \langle 3, 8\rangle$.

At $s = 3$, JIMMY$_B$ reasons about the last offer received from Alice:

$$\begin{aligned}
g(p_2, p_1; o_B) =& \Pi_3 \circ f_{be} \circ f_{tr} \circ f_{se} \circ f_{po}(p_2, p_1, null) \\
=& \Pi_3 \circ f_{be}(f_{tr}(f_{se}(\langle 3, 8\rangle, \langle 3, 7\rangle, \langle 3, 7\rangle))) \\
=& \Pi_3 \circ f_{be}(f_{tr}(\langle 3, 8\rangle, \langle 3, 7\rangle, \langle 3, 7\rangle)) \\
=& \Pi_3 \circ f_{be}(\langle 3, 8\rangle, \langle 3, 7\rangle, \langle 3, 7\rangle) \\
=& \Pi_3(\langle 3, 8\rangle, \langle 3, 7\rangle, \langle 3, 8\rangle) \\
=& \langle 3, 8\rangle.
\end{aligned}$$

Since $\epsilon_{p_1,p_2} = 0$, that is JIMMY$_B$ has obtained as newly generated offer the same policy that was last offered by Alice, JIMMY$_B$ recommends Bob to accept and the content is shared with policy $p = \langle 3, 8\rangle$.

4.4 Properties of the model

Given that dealing with MPCs on OSN is a real and practical problem, it is crucial for the model to present some properties that allow its implementation on real systems, such as termination in a finite time and soundness.

Lemma 1. ***Termination*** *On the assumption that neither party withdraws, in a finite time, the offers p_t suggested alternatively by the two instances of the JIMMY agent always converge towards an agreement $\widehat{p}$*

$$d(p_t, \widehat{p}) \to 0 \textit{ for } t \to N < +\infty \tag{4.2}$$

or the impossibility of reaching an agreement is recognised.

Proof. During the negotiation process, a user can either maintain his/her position or accommodate the other user's preference. Let us consider each case separately:

(a) Both users want to accommodate each other: let us prove this by contradiction and let us assume that (4.2) is false; this means that the distance between the new suggestion and the final agreement may increase at each iteration or that the convergence may happen in an infinite number of iterations. By the definitions of the functions f_v, the output of each f_v is always within the range defined by the most and the least restrictive tuples of each iteration (see Equation (4.1)). If the users are both trying to accommodate each other's preference, it means that the new suggestion is a tuple whose at least one element is internal to the domain; i.e. the domain of $f_v(p_\alpha, p_\beta, p_v)$ becomes one of the following:

$$\begin{array}{ll} [d_1, d_2) \times (i_2, i_1] & (d_1, d_2] \times (i_2, i_1] \\ [d_1, d_2) \times [i_2, i_1) & (d_1, d_2] \times [i_2, i_1) \\ (d_1, d_2) \times (i_2, i_1). & \end{array}$$

Noting that these are all subsets of $\mathbb{N}^2$, it follows that the width of the domain of the new suggestion decreases at every iteration. Therefore, the distance between the new suggestion and the final deal can only decrease, as they both are elements of the domain; this contradicts our initial hypothesis. Also, given the fact that the domain is a finite and bounded subset of $\mathbb{N}^2$, the convergence happens in a finite number of iterations. Therefore, (4.2) is valid.

(b) One user wants to accommodate, the other user holds his/her position: the reasoning is similar to the previous case, but now the contraction of the domain happens only at every other iteration, i.e., whenever a user makes an accommodating offer. In fact, when a user sticks to his/her preference, he/she keeps offering a tuple whose elements are on the border of the domain. Eventually with a slower speed than in the previous case, the domain does get contracted and, given its finite dimension and its boundary, it converges in a finite time.

So, (4.2) is valid.

(c) At some point, both users start holding their positions: if both users stick to their preferences, it means that both of them keep suggesting policies whose elements are on the border of the domain

$$f_v(p_\alpha, p_\beta, p_\sigma) \in [d_1, d_2] \times [i_2, i_1] \in \mathbb{N}^2;$$

therefore, the domain cannot contract. The system interrupts the negotiation whenever both users have tried all their strategies after receiving the same inputs. Since every user has at most two strategies, this happens at latest at the 5th iteration of the algorithm with no changes in the offers: that is, the termination of the algorithm is realised in a finite time. □

Lemma 2. ***Soundness*** *Assuming that the users always follow the suggestions of their agent JIMMY, the outcome of the negotiation is optimal, i.e., it is consistent with their initial preferences and orders over the moral values.*

Proof. This can be proven by contradiction. Let us assume that the outcome of the negotiation is not consistent with the users' initial preferences over the sharing policies and their preferred order over the values, even though both users always followed their agent's suggestions. This implies that, in at least one step of the negotiation process, JIMMY recommended a sharing policy which was not consistent with the users' inputs. A new policy proposal is computed according to the function g (see Definition 7) as the composition of value-functions f_v. Hence, if a new suggestion is not consistent with the user's preferences, it means that either (i) the composition order or (ii) the value-functions f_v are not consistent with the inputs. However, (i) the order for composing the value-functions f_v is defined exclusively by the sd-function (see Definition 5) which, according to the order over the values provided by the user, assures that different relevance is given to different value-functions f_v

according to the priority that the user assigns to the values that each f_v function represents. Therefore, the composition order is, by definition, consistent with the user's preference. On the other hand, (ii) the value-functions f_v are defined in such a way that they reflect the interpretation of each value in the negotiation context given the initial policy preferences. So, by definition, the value-functions are consistent with the initial preferences of the users. In both (i) and (ii) we reached a contradiction, therefore we can say that, given a coherent behaviour from the users' side, every step of the negotiation must be consistent with the initial preferences of the users. Every outcome of the model must be reached through a sequence of such consistent steps; therefore every outcome, whether it is a deal or not a deal, must be consistent with the initial preferences of the users. □

4.5 Discussion

JIMMY is my first attempt to support users when collaboratively managing multi-user privacy online. On the one hand, its design satisfies the requirements for solving MPCs in OSNs that I identified in the first iteration of conceptual investigation and that I have described at the beginning of this chapter. In particular, the agent is *role-agnostic*: in fact, JIMMY equally supports both participants during the audience negotiation process in an equal way. Then, JIMMY is *adaptive*: whenever a new negotiating action needs to be recommended, the agent generates a new policy starting from the original (in the first iteration) or updated (later iterations) policy preferences of the users (see Definition 7); this also contributes to the soundness of the model, as proven in Lemma 2. Finally, JIMMY is *value-aligned*: new negotiating proposals are generated through the composition of value functions, which directly map behavioural preferences either on the overall negotiation strategy (see Definition 5) or on individual negotiation steps (see Definition 6), and whose composition order is defined by the user's value order (see Definition 4). The value-alignment of

JIMMY also contributes to its soundness.

On the other hand, however, this preliminary model presents some limitations which have directed and influenced the subsequent steps of my research work.

First, even though I properly define its components and it presents some interesting formal properties, this model lacks some other formal definitions, such as the ones for the syntax and the semantics of the negotiation protocol [109]. Also, the notion of optimality, which in this chapter refers to “consistency with the user's moral values”, would benefit from a better definition, that would help distinguish and compare different degrees of goodness of a solution.

Second, this model assumes the user-agent interaction, but does not include any element to support it. Theoretically, JIMMY may present the user with a value-based justification of its recommendations, but it is not enabled to reason over alternative solutions. This would, for instance, make impossible for the agent to provide contrastive explanations of its outputs. This leads to two distinct but complementary research lines: (i) to design agents with more complex reasoning abilities, and (ii) to investigate further how to successfully convey the model's output to the users.

Lastly, by relying only on the moral values during the generation of a new policy, JIMMY fails to represent some other motivations that drive the users in the real-world, such as their perception of individual advantage or disadvantage when sharing content with some desired or undesired audience. For this reason, I argue that a hybrid agent architecture, which combines both a value-driven component and a utility-driven component, may provide recommendations that are more acceptable and useful for users.

4.6 Conclusion

Following a VSD approach (see Section 2.4), a preliminary conceptual investigation of the related literature identified *role-agnosticism*, *adaptivity* and *value-alignment* as design features to be embedded in models to effectively manage multi-user privacy on OSNs. In this chapter I have presented the output of a first technical investigation in this direction, that is the design of JIMMY, an agent that aims to support users when collaboratively defining privacy policies for co-owned content. JIMMY's design is such that it is *role-agnostic*, i.e. it supports equally all the users, *adaptive*, i.e. its recommendations depend on the users' contextual preferences, and *value-aligned*, i.e., it recommends actions according to the user's morality, while supporting the users *interaction* during the collaborative deliberation of a solution for an MPC.

This interaction is modelled as a *negotiation*, where alternatively one user offers a sharing policy and the other user either accepts it or rejects it. The agent, by considering the user's moral values, recommends the action which is most coherent with the user's desired behaviour, similarly to what the Jiminy Cricket does in the Pinocchio novel. Specifically, I mapped some relevant Schwartz values onto behaviours which are common during a negotiation: each user (i) may want to hold onto their own preference (when *power* is dominant), or (ii) may be willing to accommodate the interlocutor (when *benevolence* is dominant); or (iii) may prefer to act according to what the society would generally do in a similar circumstance (when *tradition* is dominant); or (iv) may lean always towards the safest, i.e., the most private, option (when *safety* is dominant). Also, I considered that when *self-direction* influences the behaviour, then open-mindness and creativity may lead the user to alter their negotiating strategy in order to identify new solutions. According to the relative preference of the user over these values, JIMMY recommends the most appropriate action at each negotiation step, that is either to accept the received offer or to make a counteroffer. If the users always follow the agent's recommendation and an agreement is recognised, then that agreement is guaranteed to be the most consistent with the users' initially preferred sharing policy and values.

Although JIMMY satisfies the properties stated in the Introduction to this chapter, namely it is role-agnostic, adaptive and value-aligned, it also presents some drawbacks. For instance, the user-agent interaction, crucial for *explaining* the agent's outputs to the user, is not modelled, and the identified solutions, being only value-driven, risk not to fully satisfy the users. For instance, despite being mainly driven by benevolence, a user may perceive the potential privacy violation caused by some undesired person accessing the content as too serious to just accommodate the co-owner's preference.

Still following a VSD approach, in the next chapter I will describe the output of my updated conceptual and technical investigations, namely a new agent-based model which overcomes most of JIMMY's deficits: it is value-driven but it also considers the individual's utility of a sharing policy; it is able to reason about alternative solutions; and it presents an explainable architecture, which aims to help users understand the underlying processes of the model.

Chapter 5

ELVIRA: A Value-aligned, Utility-driven and Explainable Agent

5.1 Introduction

Aware of the limitations of the JIMMY model described in the previous chapter and in line with an iterative Value Sensitive Design approach (VSD, see Section 2.4), I have performed another round of conceptual investigation regarding the effective resolution of multi-user privacy conflicts (MPCs) on OSNs. As a result, as I have introduced in Section 1.3 (cf. Table 1.1), I have included two new requirements: a model that effectively manages online multi-user privacy should not only be adaptive, role-agnostic and value-aligned, but also utility-driven and explainable. In this chapter I describe ELVIRA, another agent-based model that can support the collaborative resolution of MPCs, as the output of a new technical investigation stage, that had the aim of designing a system that would fulfil all these properties.

The design of ELVIRA represents a refinement of JIMMY's: its strengths, such as being role-agnostic, adaptive and value-aligned, are maintained, and its drawbacks, namely not being explainable and utility-driven, are compensated. In fact, when computing the solution to the MPC, ELVIRA explicitly considers not only the user's *moral values*, again modelled based on the Schwartz theory [154], but also the user's individual *utility*, interpreted as the gain or loss the user can perceive when desired or undesired audiences access the content subject of the MPC. Then, ELVIRA is also *explainable* in the way that it presents both a cognitive process, given by the practical reasoning approach described in Section 5.3, and a social process, defined and empirically evaluated in Section 5.6. Finally, ELVIRA is *role-agnostic* and *adaptive*, as guaranteed by the formal properties of soundness, completeness, anonymity and neutrality (Section 5.4). Therefore, ELVIRA's design is such that it complies with all the desired requirements described in Table 1.1.

5.2 The ELVIRA architecture

ELVIRA is an agent that could be used on top of a social network, as an extra service that could be offered to the users, or directly embedded in the OSN architecture. Therefore, I define all the components of the agent basing on a graph-based representation of the social network the agent may interact with.

I represent an OSN as a graph $G = (V, R)$, where V is the set of the OSN users, and R describes all their relationships $(v_k, v_j, i_{kj}) \in R$, where $i_{kj} \in [0, i_{max}]$ represents the intimacy or closeness of the relationship, which can be elicited automatically [61].

Among other activities, users can engage with the network by sharing online content[i] that they own offline. While in certain circumstances ownership is clear (e.g., when a user takes a selfie, that picture belongs to her/him), there are situations when ownership can be more challenging to define [175]: in a group picture, all the depicted people would co-own the photo; in a picture depicting kids, it is likely that the parents, despite not being depicted, would own the photo; etc. We consider everyone whose privacy may be impacted by a picture to be an *owner* of that picture.

[i] In this thesis I mostly focus on photographic content, but similar solutions can be applied also to other types of content.

Definition 9. *Given a set of digital content X and the function **ownership**, $own : V \rightarrow X$, a user $v \in V$ owns the item $x \in X$ if $x \in own(v)$.*

Ownership is not an injective function and the same item $x \in X$ could be co-owned by multiple users. E.g., when both $v_1, v_2 \in V$ own the item x, we denote the *co-ownership* as $x \in own(v_1) \cap own(v_2)$ and the *co-owners* as $Ag = \{v_1, v_2\}$.

In line with previous work [176], but noting that this is equivalent and can be translated to and back from the group-based access control models used in OSN platforms [173], I define a *sharing policy* as follows:

Definition 10. *A **sharing policy** for an item $x \in X$ from user $k \in V$ is $sp = \langle d, i \rangle$, where d is the length of the shortest path connecting a user with k, and i is the minimum intimacy that each link of the path connecting the user with k must satisfy for the user to have access to the item.*

I assume that every user has a *preferred sharing policy* for each content they are involved in (e.g., they own), and that it can be elicited automatically (e.g. see Section 3.2.1). I denote with sp_k the user's k preferred sharing policy. In addition, each sharing policy sp defines for the user k an individual *audience* $aud_{sp,k}$, i.e., a set of users who satisfy the conditions of sp from user k. An MPC occurs when users that are involved in the same item, namely the *co-owners* Ag of the item, have contradictory preferred sharing policies which lead to different preferred audiences.

Definition 11. *An **MPC** regarding an item $x \in X$ co-owned by users $k \in Ag$ and $j \in Ag$, i.e., $x \in own(k) \cap own(j)$, occurs when k and j's preferred audiences do not coincide, i.e. $aud_{sp_k,k} \neq aud_{sp_j,j}$.*

Definition 12. *When considered from all the involved users' points of view, a sharing policy sp' grants access to the item to the **collective audience** $aud_{sp'}$, which is* *the intersection of the individual audiences generated by sp' for each involved user:*

$$aud_{sp'} = \bigcap_{k \in Ag} aud_{sp',k}.$$

In the remaining part of this chapter, I will refer to *candidate solutions* for an MPC as collective audiences.

Furthermore, I consider that the item can be shared in its original form (*as-it-is*) or in its pre-processed version (*modified*), e.g., where some parts are blurred or cropped (similarly to the type of fine-grained solutions like [84] that I mentioned in Section 3.2.2). In fact, empirical evidence [175] suggests sharing modified content is sometimes an acceptable compromise among co-owners. Generally, the candidate solution audience guarantees access to the original item; in addition, if specified with $aud_{sp',mod}$, the solution allows also to share the *modified* content with the users in $\bigcup_{k \in Ag} aud_{sp'_k,k} \setminus aud_{sp'}$ that are excluded from the solution audience.

5.2.1 The utility-driven layer

Users are known to benefit from sharing in social media [94], e.g. gaining utility if an appealing picture is shared, but they also lose utility if a compromising picture is seen by the wrong people. These effects are amplified with people having closer/more intimate relationships, as they usually generate more utility gain/loss if included or excluded from the preferred audience [175].

A compromising solution to a MPC may generally moderate the gain of utility of some users in order to alleviate the loss of utility for others, according to the portions of the individual preferred audiences that are included in the solution. Finally, I also consider that each user may eventually prefer to under-share or over-share the item, that is to make it visible to a smaller or broader audience than the preferred one.

Following the rationale above in order to define the utility of a suggested solution audience, I first define the following sets with respect to the user k and his/her

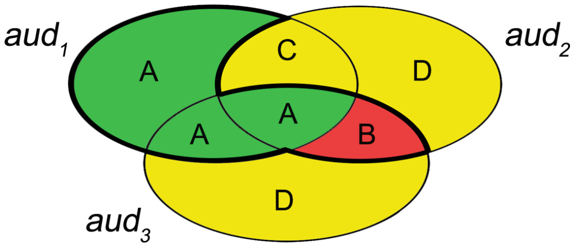


Figure 5.1: MPC between 3 users, a possible solution aud' (represented with bold borders), and the $\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}$ sets for user 1.

preferred audience $aud_{sp_k,k}$, considering the collective audience aud' as a potential solution to a MPC where k is involved (see Figure 5.1 for a graphical representation), then the appreciation function capturing the tendencies to under/over-share, and finally the utility function.

Definition 13. *The* ***allowed audience*** $\mathcal{A}$ *is the set of users who* k *desires to grant access to* $x \in X$ *and that are part of the solution audience, i.e.,* $\mathcal{A} = aud_{sp_k,k} \cap aud'$*. The* ***allowed extra audience*** $\mathcal{B}$ *is the set of users who* k *desires to forbid access to* $x \in X$ *but that are part of the solution audience, i.e.,* $\mathcal{B} = aud' \setminus aud_{sp_k,k}$*. The* ***excluded audience*** $\mathcal{C}$ *is the set of users who* k *desires to grant access to* $x \in X$ *but that are forbidden to access or allowed to access only a modified version, i.e.,* $\mathcal{C} = aud_{sp_k,k} \setminus aud'$*. The* ***excluded extra audience*** $\mathcal{D}$ *is the set of users who* k *desires to forbid access to* $x \in X$ *and that are either forbidden to access or allowed to access only a modified version of the item, i.e.,* $\mathcal{D} = \bigcup_{l \neq k} aud_{sp_l,l} \setminus aud'$*.*

Definition 14. *Given a set of digital content* X*, the function* ***appreciation***, $app : X \longrightarrow [-1, 1]$*, maps an item* $x \in X$ *into a positive value if the user is happy to overshare, and to a negative value if the user prefers to under-share.*

Notice that the elicitation of appreciation will require further study: for instance, the user may define it for each item or for sets of similar items (e.g. based on their sensitivity), or it may be learnt over time.

$\Delta utility$		Domain
$+\frac{i_j}{d_j}$	$\forall j \in \mathcal{A}$	allowed audience
$app(x)\frac{i_j}{d_j}$	$\forall j \in \mathcal{B}$	allowed extra audience
$-\alpha\frac{i_j}{d_j}$	$\forall j \in \mathcal{C}$	excluded desired audience
$app(x)\beta\frac{i_j}{d_j}$	$\forall j \in \mathcal{D}$	excluded extra audience

Table 5.1: Variation of the individual utility for item x, considering audience sets, appreciation and mode of sharing.

Definition 15. *Given an audience aud, its* ***utility*** *for user k is:*

$$u_{k,aud} = \sum_{j \in \mathcal{A}} \frac{i_{kj}}{d_{kj}} - \alpha \sum_{j \in \mathcal{C}} \frac{i_{kj}}{d_{kj}} + app(x) \left(\sum_{j \in \mathcal{B}} \frac{i_{kj}}{d_{kj}} + \beta \sum_{j \in \mathcal{D}} \frac{i_{kj}}{d_{kj}} \right), \tag{5.1}$$

where d_{kj} represents the length of the shortest path between user k and any user j (i.e., their distance) and i_{kj} represents the sum of the intimacy over that path.

For the sake of clarity, Table 5.1 shows the individual contributions of each audience set to the variation in utility. Note that the components for the sets $\mathcal{C}$ and $\mathcal{D}$ depend on the selection of α and β, system parameters which determine whether to share the content only *as-it-is* ($\alpha = 1$ and $\beta = 0$) or also *modified* ($0 < \alpha, \beta < 1$). However, experiments showed (see Experiment IV in Section 6) that the optimal choice of these two parameters does not seem critical, as no significant impact was found on the differences between individual utilities achieved under different values for the parameters.

Example Let us consider the simplified OSN in Figure 5.2. Alice wants to upload on an OSN the picture x, where she appears with her friends Bob and Charlie ($Ag = \{A, B, C\}$). Their preferred sharing policies for x are respectively $sp_A = \langle 2, 2 \rangle$, $sp_B = \langle 1, 3 \rangle$ and $sp_C = \langle 3, 4 \rangle$, and generate the following individually preferred audiences: $aud_{sp_A,A} = \{A, B, C, D, E, F, G, I\}$, $aud_{sp_B,B} = \{A, B, C, D, G\}$ and $aud_{sp_C,C} = \{A, B, C, G, I\}$. A conflict occurs, because the three individual

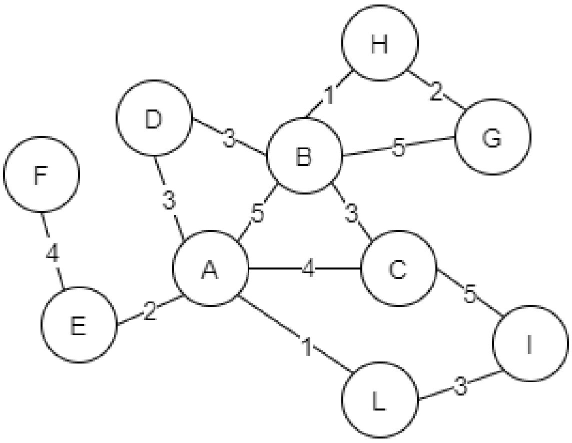


Figure 5.2: The simplified online social network discussed in the example.

user	$\mathcal{A}$	$\mathcal{B}$	$\mathcal{C}$	$\mathcal{D}$
A	{B,C,D,G,I}	∅	{E,F}	∅
B	{A,C,D,G}	{I}	∅	{E,F}
C	{A,B,G,I}	{D}	∅	{E,F}

Table 5.2: Detail of the audience sets for each user when considering $aud_{sp'}$.

preferred audiences do not coincide. Furthermore, Alice and Charlie prefer to eventually under-share the content x ($app_A(x) = app_C(x) = -1$), while Bob prefers to overshare it ($app_B(x) = +1$).

Let us consider $sp' = \langle 2, 3 \rangle$ as a possible solution to this conflict. This generates the solution audience $aud_{sp'} = \{A, B, C, D, G, I\}$; if we consider $aud_{sp',mod}$, then sharing the modified content is allowed ($0 < \alpha, \beta < 1$) and $\{E, F\}$ will access the pre-processed content.

Then, considering the individual audience sets as reported in Table 5.2, Alice, Bob and Charlie would perceive the following variation in utility (with some values of α

and β:

$$
\begin{aligned}
u_{A,aud_{sp'}} &= \sum_{j\in\{B,C,D,G,I\}} \frac{i_j}{d_j} && -\alpha \sum_{j\in\{E,F\}} \frac{i_j}{d_j} \\
&= \frac{5}{1}+\frac{4}{1}+\frac{3}{1}+\frac{10}{2}+\frac{9}{2} && -\alpha\left(\frac{2}{1}+\frac{6}{2}\right) \\
u_{B,aud_{sp'}} &= \sum_{j\in\{A,C,D,G\}} \frac{i_j}{d_j} && +1\cdot\left(\sum_{j\in\{I\}} \frac{i_j}{d_j}+\beta\sum_{j\in\{E,F\}} \frac{i_j}{d_j}\right) \\
&= \frac{5}{1}+\frac{3}{1}+\frac{3}{1}+\frac{5}{1} && +\frac{8}{2}+\beta\left(\frac{7}{2}+\frac{11}{3}\right) \\
u_{C,aud_{sp'}} &= \sum_{j\in\{A,B,G,I\}} \frac{i_j}{d_j} && -1\cdot\left(\sum_{j\in\{D\}} \frac{i_j}{d_j}+\beta\sum_{j\in\{E,F\}} \frac{i_j}{d_j}\right) \\
&= \frac{4}{1}+\frac{3}{1}+\frac{8}{2}+\frac{5}{1} && -\frac{6}{2}-\beta\left(\frac{10}{3}+\right).
\end{aligned}
$$

5.2.2 The value-driven layer

The *Theory of Basic Values* by Schwartz [154] (refer to Section 2.3.1 for more details on this) is one of the most well-known and established socio-cultural theories of human values, according to which values are socially desirable concepts that represent the mental goals which drive human behaviour and influence any people's decision.

Schwartz identifies ten main values and orders them in a circular way, considering reciprocal similarities and influences. As depicted in Figure 5.3, two dimensions emerge overall and define four directions that represent higher order values $\mathcal{V}$, or *hypervalues*, which pull apart while influencing the human behaviours: *openness-to-change* (OTC), *self-transcendence* (ST), *conservation* (CO), and *self-enhancement* (SE).

Driven by the Schwartz hypervalues, which for simplicity I refer to as just *values*, I identified four main behaviours that humans follow when interacting in group deliberations such as when resolving a MPC:

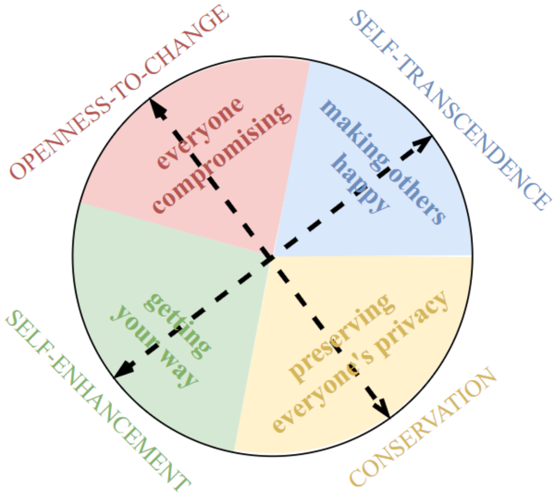


Figure 5.3: The Schwartz hypervalues and their interpretation within the MPC resolution context.

- *self-transcendence* pushes towards making the others happy, by accepting their ideas and preferences;
- *self-enhancement* pushes towards getting the one's own way, by maintaining or increasing one's own utility;
- *openness-to-change* pushes towards appreciating compromises which differ from everyone's initial preference;
- *conservation* pushes towards preserving individual and social security.

Studies showed that the preferred order of individuals over the values is relatively stable over their lifetime [24]. This suggests that it should not be necessary to elicit it –through validated tools [154]– from the users for every MPC, but it may be sufficient to do it just once initially or every some fixed time interval (e.g., yearly). It is important that ELVIRA is accurately informed about its user's preference over the values, in order for it to recommend sharing solutions that are coherent with the user's desired behaviour in the circumstance of the MPCs.

In fact, by comparing the candidate solutions and preferring one audience over

Value		Sharing Condition	Behaviour
OTC	+	with aud_f	everyone compromising
	-	with some user's preference	
CO	+	with more private option	preserving everyone's privacy
	-	with a more public option	
ST	+	with the other's preference	making others happy
	-	ignoring the other user's preference	
SE	+	with own preference	getting your way
	+	gaining better utility	
	-	gaining worse utility	

Table 5.3: Promotion and demotion of the values for a user when comparing different sharing options.

the others, it is possible to promote or demote the values, as I specify in Table 5.3; note that, in the case of conservation and self-enhancement, the comparison is mainly made with the user's initial preference. For example, if a candidate audience coincides with the user's initial preference or would generate the maximum utility, then its selection would allow the user to get his/her own way and would promote self-enhancement; if a candidate audience is the most private one, or more private than the user's preference, then its selection would allow to better preserve everyone's privacy and would promote conservation, and so on.

Definition 16. *Given an audience aud, its* ***value promotion*** *for user k, who has a preferred order over the values $o \in \mathcal{O}_{\mathcal{V}}$, is:*

$$v_{k,aud} = \sum_{i=1}^{|\mathcal{V}|} (I - i) \cdot prom_{aud}(o_i), \tag{5.2}$$

where $I = |\mathcal{V}| + 1$, and $prom_{aud}(o_i) = 1$ if the i-th preferred value is promoted by selecting aud, $prom_{aud}(o_i) = -1$ if the i-th preferred value is demoted, and $prom_{aud}(o_i) = 0$ otherwise.

Running example Considering the same MPC as in the previous example (see a summary of the user's preferences in Table 5.4), let us discuss how Alice may

Users k	sp_k	Values	$app(x)$
Alice	$\langle 2,2\rangle$	$ST \succ OTC \succ CO \succ SE$	-1
Bob	$\langle 1,3\rangle$	$CO \succ SE \succ OTC \succ ST$	+1
Charlie	$\langle 3,4\rangle$	$OTC \succ CO \succ ST \succ SE$	-1

Table 5.4: Users' preferences in the MPC discussed in the running example.

promote and demote her values by selecting different candidate solutions.

By selecting $aud_{\langle 2,2\rangle}$ as solution audience, Alice would promote SE, because $\langle 2,2\rangle$ is her own preference, but would demote OTC and ST. By selecting $aud_{\langle 1,3\rangle}$, Alice would promote CO, because $\langle 1,3\rangle$ is the most restrictive policy, and ST, because she is selecting another user's preference; but she would demote OTC and SE, because she would gain a lower utility than with her preferred audience (for brevity, I do not report all the individual utilities for each audience). By selecting $aud_{\langle 2,3\rangle}$, Alice would promote OTC, because $\langle 2,3\rangle$ is different from every user's preference, CO, because $aud_{\langle 2,3\rangle}$ is more restrictive than her preference, and SE, because she would gain a higher utility than with her preference. For a complete view of the value promotion for all the involved users, see later Figure 5.4.

5.2.3 Collaborative behaviour

Every user involved in the MPC is represented by an agent and all the agents work together collaboratively to resolve the MPC –recall that the focus of this thesis is on the majority of MPCs which happen in non-adversarial settings [28, 97, 207, 175]. That is, for each MPC involving n users, there will be a set Ag of n agents, with one *uploader* agent and $n-1$ *co-owner* agents. The agents perform most activities similarly, regardless of their role; however, the uploader has more responsibilities and presents a more complex reasoning process, as I detail in the next section. Therefore, I will present ELVIRA from the perspective of the uploader agent, and I will specify when and how the agent behaviour differs for the co-owners.

I report below an overview of the collaborative process to solve an MPC:

1. a user (the uploader) wants to share some content, but an MPC is detected for a group of users;

2. each agent elicits its user's preference, both sharing-wise and value-wise;

3. the candidate solutions for the MPC are identified;

4. each agent individually evaluates every possible candidate solution;

5. the uploader agent collects all the individual evaluations and identifies the best solution;

6. the uploader agent shares the optimal solution with the co-owner agents;

7. every agent presents to its user a justified recommendation for selecting the optimal sharing solution;

8. every involved user decides whether to accept the recommendation.

Whenever an MPC is detected in an OSN (**1**), all the agents of the involved users get activated. As I explained in Sections 5.2.1 and 5.2.2, each agent already knows or can elicit its user's preferred order over the values and preferred sharing policy for each item (**2**), or group of items. In fact, in the following I describe how to resolve an MPC over a single item for simplicity but without loss of generality, as one could define a preferred audience over a collection of items too and solve a set of MPCs at the same time. Next, the uploader agent collects from the other agents the sharing preferences of all the users and defines the set of candidate audiences $\mathbb{A}$ (**3**). $\mathbb{A}$ is a finite set which includes the n collective audiences $aud_1, ..., aud_n$ deriving from the users' preferred sharing policies, and aud_f, where f is some function identifying a subset of the union of all the individually preferred audiences, such that $aud_f \neq aud_k \quad \forall k \in Ag$. Since each audience can be selected either *as-it-is* or *modified*, there are $|\mathbb{A}| \leq 2(n+1)$ possible solutions to the MPC ($\leq$ because two or more co-owners may have the same preferred audience). In the remaining part of this chapter, we do not specify whether the audience is selected *as-it-is* or *modified*, because all the candidate solutions are considered equally, as we show later in Lemma 6.

For each audience $aud \in \mathbb{A}$, each agent k computes its *individual score* (**4**), which represents its appreciation of the particular option in terms of utility and value promotion:

$$s_{k,aud} = \begin{cases} -u_{k,aud} \cdot v_{k,aud} & \text{if } u_{k,aud} < 0 \text{ and } v_{k,aud} < 0 \\ u_{k,aud} \cdot v_{k,aud} & \text{otherwise} \end{cases} \tag{5.3}$$

where the utility $u_{k,aud}$ is computed as in Equation (5.1) and the value promotion $v_{k,aud}$ is computed as in Equation (5.2). u and v are multiplied for assigning equal weight to utility and values regardless of their range.

Then, all the co-owners share their individual evaluation of each audience, including utility, value promotion and score, with the uploader (**5a**), who aggregates them in an *overall score* for each audience $aud \in \mathbb{A}$:

$$s_{aud} = \sum_{k \in Ag} s_{k,aud}. \tag{5.4}$$

Notice that, given the assumption of collaborative behaviour among users and agents [175], the ELVIRA agents are expected to share their truthful preferences in terms of candidate solution and evaluations, because the underlying shared goal is not to harm anyone by sharing content inappropriately. Also, in order to falsify the evaluations of the solutions, an agent should either manipulate its own value order, which is not in its interest and goes against the principles of value-aligned AI, or manipulate its own utility, by accentuating the perceived loss or gain generated by the selection of each candidate solution. In this model, the utility evaluation is strictly computed according to Definition 15, but further work may look into adding other reasoning layers that would enable a more strategic decision making process.

At this point, the uploader agent has gathered all the necessary information that allows it to identify the best collective solution (**5b**) by applying argumentation techniques, as I explain in the next section. Notice that, while the problem of identifying the best collective solution could be framed simply as a multi-objective (utilities and values) optimisation problem, the use of argumentation enables the agent not only to find the optimal solution, but also to track all the information that is necessary for justifying such outcome to the end-user in the form of attacking or supporting arguments, which, in turn, can be easily reported to the user as part of the explanation.

Then, the solution is shared with all the agents (**6**), who present it to their users together with a tailored explanation describing its optimality (**7**); in Section 5.5 I discuss possible designs for these explanations, but the use of argumentation (and of the practical reasoning process in particular) would facilitate the generation of diverse types of explanations, including dialogical, that can be adapted to the recipient's need and preferences. Finally, every user is free to decide whether to accept the recommendation (**8**): if any of the users rejects it, then the content is not shared and a manual negotiation can eventually take place among the users.

5.3 Practical reasoning

I describe the practical reasoning process performed by the ELVIRA agent, adapting the work by Atkinson and Bench-Capon that I summarised in Section 2.5. The agent, by completing the abductive reasoning process that I detail below, not only identifies the most desirable audience, but it also gathers all the necessary information to discuss its causal attribution, which represents the *cognitive process* required for providing an explanation [112]. I specify how ELVIRA uses this information to generate the explanations in Section 5.6.

First, I consider that an agent can propose, attack and defend justifications for a given action by relying on an argument scheme (ArgS) and its associated critical questions (CQs) [14, 15]. From the point of view of the uploader agent, ArgS can be expressed as: *"I should offer the audience aud', that will lead to an agreement, that will generate the score $s_{aud'}$ and that will promote the values V"*.

In order to identify the best solution to offer, ELVIRA uploader follows a practical reasoning process (PR)[14]: (1) it identifies a desirable outcome, e.g. agreement on the audience *aud′*; (2) it argues in favour of offering *aud′*, e.g. by instantiating the ArgS; 3) it considers objections (the critical questions CQs) based on alternative more desirable audiences, e.g. by considering possibly better overall scores or promoted values; and, finally, 4) it attempts to rebut these objections.

Formally, the PR has three stages: (i) the *problem formulation*, (ii) the *epistemic stage*, and (iii) the *choice of action*.

5.3.1 Problem formulation

The first step of PR consists of representing the relevant elements of the situation (i.e. conflict occurrence, involved users' preferences, possible actions and solutions, etc.), which can be performed by building an Action-Based Alternating Transition Systems with Values (AATS+V) [14]. This structure provides the underlying semantics used to describe the world and formulate arguments about *joint actions* (J_{Ag}), i.e. actions that are performed by a set of agents and that influence each other's outcome.

In the MPC context, a joint action is composed of the uploader's offer of an audience and the co-owners' response[ii]. I adapt Atkinson's definition of an AATS+V [14] to MPCs as follows:

Definition 17. *In the context of an MPC among n users, an* ***AATS+V*** *is a* $2n+8$ *tuple* $\Sigma = \langle Q, q_0, Ag, Ac_k, \rho, \tau, S, \mathcal{V}, Av_k, \delta \rangle$*, with* $k = 1, ..., n$*, where:*

[ii] As in [15], I assume the offer and the response to be "simultaneous" actions, despite their sequentiality.

- $Q = \{\textit{conflict, agreement}_{aud} \quad \forall aud \in \mathbb{A}\}$ *is a finite, non-empty set of states;*

- $q_0 = \textit{conflict}$ *is the initial state;*

- $Ag = \{up_1, co_2, ..., co_n\}$ *is the set of agents involved in the MPC, with the roles of uploader or co-owners;*

- $Ac_1 = \{\textit{offer}_{aud} \quad \forall aud \in \mathbb{A}\}$ *are the actions available to the agent* up_1*;*

- $Ac_k = \{accept_k, reject_k\}$ *are the actions available to the agent* co_k*, for* $k = 2...n$*;*

- $\rho : Ac_{Ag} \rightarrow 2^Q$ *is the action-precondition function; here, every action can be executed just from* q_0*;*

- $\tau : Q \times J_{Ag} \rightarrow Q$ *is the partial system transition function, which defines what state results from performing the joint action* j *in the state* q*, where possible; here, only the joint actions where all the co-owners accept the uploader's offer end up in an agreement state, the others stay in* q_0*;*

- $S = \{0, s_{aud} \quad \forall aud \in \mathbb{A}\}$ *is the set of collective scores characterising each state, where* $s_{q_0} = 0$*;*

- $\mathcal{V} = \{SE, ST, CO, OTC\}$ *is the set of values considered;*

- $Av_k = o_k(\mathcal{V})$ *is the preferred total order of the agent* Ag_k *over the values* $\mathcal{V}$*;*

- $\delta : Q \times Q \times Av_{Ag} \rightarrow \{+, -, =\}$ *is the valuation function, which defines the effect of a transition over each value for each agent (see Table 5.3).*

5.3.2 Epistemic stage

The epistemic stage consists of determining what the agent believes about the current situation, given the previous problem formulation. As I mentioned earlier, based on empirical evidence [175], the ELVIRA agents have a collaborative behaviour. From this underlying assumption I can further imply two epistemic assumptions:

- **EA1**: all agents share the same interpretation of the world and have the same knowledge regarding the individual evaluations of the candidate solutions: common knowledge $= \{u_{k,aud}, v_{k,aud}, s_{k,aud} \qquad \forall k \in Ag, aud \in \mathbb{A}\}$;

- **EA2**: the co-owners are believed to accept an offer in two situations, i.e. when the offered audience aud' guarantees either (i) the individual maximum score ($s_{k,aud'} = \max_{\mathbb{A}} s_{k,aud}$), or (ii) the collective maximum score ($s_{aud'} = \max_{\mathbb{A}} s_{aud}$).

These epistemic assumptions are necessary because they allow the agent to discard any CQs related to the problem formulation and its truthfulness (EA1) and to evaluate appropriately the expectations regarding the other agents' actions (EA2).

5.3.3 Choice of action

Finally, I develop a value-based argumentation framework that instantiates an appropriate argument scheme, and the agent evaluates it according to its preference over the values. Starting from ArgS, the agent discusses the CQs which contest the desirability of the audience aud':

- **CQ1** Would another audience guarantee a better overall score?

$$\exists aud \in \mathbb{A} : s_{aud} > s_{aud'}$$

- **CQ2** Would another audience with at least the same overall score promote better values?

$$\exists aud \in \mathbb{A} : s_{aud} \geq s_{aud'} \wedge v_{Ag,aud} > v_{Ag,aud'},$$

  where $v_{Ag,aud} = \sum_{k \in Ag} v_{k,aud}$

- **CQ3** Would any co-owner reject this offer? i.e.

$$\exists j \in J_{Ag}, k \in Ag : j_1 = \text{offer}_{aud'} \wedge j_k = \text{reject}$$

J_{Ag}	τ
$j_{1-8} = \langle \text{offer}_{aud}, \text{reject}_2, \text{reject}_3 \rangle$	$\tau(\text{conflict}, j_{1-8}) = \text{conflict}$
$j_{9-16} = \langle \text{offer}_{aud}, \text{accept}_2, \text{reject}_3 \rangle$	$\tau(\text{conflict}, j_{9-16}) = \text{conflict}$
$j_{17-24} = \langle \text{offer}_{aud}, \text{reject}_2, \text{accept}_3 \rangle$	$\tau(\text{conflict}, j_{17-24}) = \text{conflict}$
$j_{25-32} = \langle \text{offer}_{aud}, \text{accept}_2, \text{accept}_3 \rangle$	$\tau(\text{conflict}, j_{25-32}) = \text{agreement}_{aud_i}$

Table 5.5: Detail of the joint actions J_{Ag} and the partial transition function τ for the running example: each $aud \in \mathbb{A}$ can be offered/accepted/rejected.

If aud' collects negative answers to all of the above questions, then it is considered the most desirable offer to make. By following this process, ELVIRA uploader is granted justification for action.

5.3.4 Running example

Considering the same MPC as in the previous examples and adding that $aud_f = \langle 2, 3 \rangle$, let us discuss how the ELVIRA uploader agent, acting on behalf of Alice, performs the practical reasoning process. Figure 5.4 shows the representation of the AATS+V: for clarity and simplicity, I depict only the *accept* actions for the co-owners; the *reject* actions, which would go back to the MPC state, would neither promote nor demote any value. Table 5.5 reports all the available joint actions, and Table 5.6 shows a summary of the utilities, value promotion and scores for each pair of user and audience (recall that the overall score is given by the sum of the individual scores —see Equation (5.4)).

First, the agent considers as desirable outcome the resolution of the MPC, i.e., the agreement of all the involved users on a collective audience. For this reason, the joint actions j_{1-24} are immediately discarded. Regarding the remaining joint actions, the agent may identify agreement on Alice's preference as a desirable outcome and argues in its favour by instantiating $\text{ArgS}_{\langle 2,2 \rangle}$: *"I should offer the audience $aud' = \langle 2, 2 \rangle$, that will be accepted by the co-owners, that will generate the score $s_{aud'} = -35$ and that will promote SE"*. Then, the agent considers eventual objections to the

	Alice			Bob			Charlie			overall		
$\mathbb{A}$	u	v	s	u	v	s	u	v	s	u	v	s
$\langle 2,2\rangle$	3.5	-6	-21.0	3.5	-2	-7.0	2.3	-3	-7.0	9.3	-11	-35.0
$\langle 2,2\rangle_{mod}$	3.4	-3	-10.3	3.6	-1	-3.6	2.4	-1	-2.4	9.4	-5	-16.3
$\langle 1,3\rangle$	0.5	2	1.0	2.0	0	0.0	-1.7	-3	-5.0	0.8	-1	-4.0
$\langle 1,3\rangle_{mod}$	0.6	1	0.6	2.2	-1	-2.2	-1.4	-1	-1.4	1.4	-1	-3.0
$\langle 3,4\rangle$	2.8	0	0.0	4.2	2	8.3	3.7	-5	-18.3	10.7	-3	-10.0
$\langle 3,4\rangle_{mod}$	2.8	-1	-2.8	4.2	3	12.5	3.7	-2	-7.3	10.7	0	2.3
$\langle 2,3\rangle$	3.5	6	21.0	3.5	1	3.5	2.3	3	7.0	9.3	10	31.5
$\langle 2,3\rangle_{mod}$	3.4	4	13.7	3.6	1	3.6	2.4	3	7.2	9.4	8	24.5

Table 5.6: Utility, value promotion and score generated by each audience for each user in the example.

desirability of aud' by discussing the critical questions: (CQ1) all the other audiences would guarantee a higher score; (CQ2) all the other audiences, apart from improving the score, would also promote values that are ranked higher (see in Table 5.6 the overall value promotion); (CQ3) both co-owners are believed to reject $aud' = \langle 2,2\rangle$, because it does not guarantee the best overall nor individual score for any of them. Given the unfavourable answers to all the CQs, AS$_{\langle 2,2\rangle}$ is discarded.

The agent proceeds similarly to consider all the other possible desirable outcomes, until it eventually formulates ArgS$_{\langle 2,3\rangle}$: *"I should offer the audience $aud' = \langle 2,3\rangle$, that will be accepted by the co-owners, that will generate the score $s_{aud'} = 31.5$ and that will promote OTC, CO and SE"*. Again, the agent discusses the CQs: (CQ1) there is no other audience which would guarantee a higher score; (CQ2) there is no other audience with at least the same score and a better overall value promotion; (CQ3) the co-owners are believed to accept because of EA2 ($aud' = \langle 2,3\rangle$ guarantees the collective maximum score). Given the favourable answers to all the CQs, ArgS$_{\langle 2,3\rangle}$ is accepted and the agent identifies $aud' = \langle 2,3\rangle$ as the solution to the MPC.

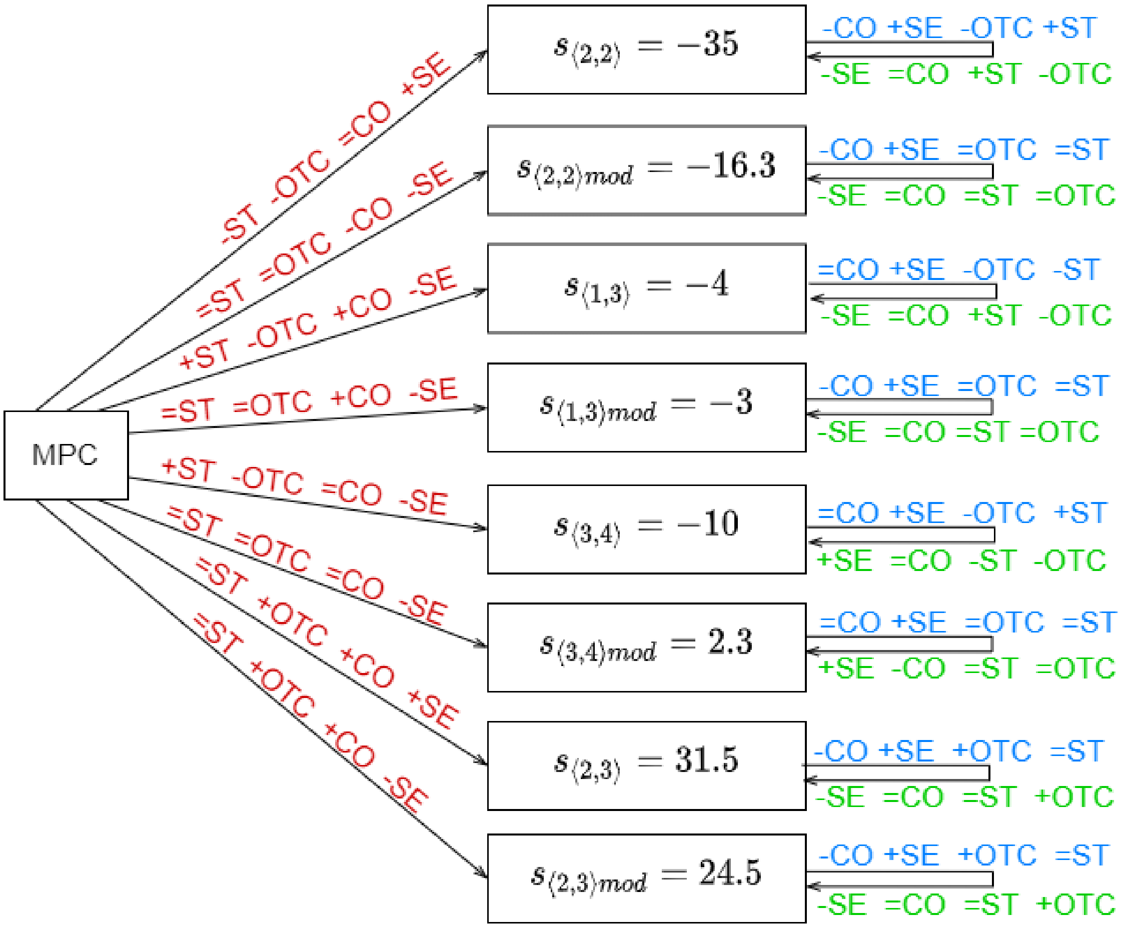


Figure 5.4: The AATS+V representing the PR performed in the example ($aud_f = \langle 2, 3\rangle$).

5.4 Formal properties

In this section, as part of the VSD technical investigation, I formally show how ELVIRA, presenting some properties such as soundness, completeness, anonimity and neutrality, fulfils the requirements of being *adaptive* and *role-agnostic.* In particular, soundness and completeness show that the model can adapt its output according to the users' preferences to always dynamically find the optimal audience, thus satisfying adaptability. Anonimity and neutrality guarantee that the preferences of uploaders and co-owners are treated equally, thus satisfying role-agnosticism.

Lemma 3 (**Soundness**)**.** *The audience recommended by ELVIRA is always optimal, i.e., it is the one which is the most coherent with everyone's utility and value preferences.*

Proof. This property can be proven by contradiction. Let us assume that ELVIRA recommends an audience aud' that is not optimal. This implies that there exists at least another audience $\widehat{aud}$ which is more desirable for the users involved in the MPC, in terms of generated utility or promoted values, both represented by the audience score. If $\widehat{aud}$ is more desirable, then it must be one of the following three cases: (i) $\widehat{aud}$ has a higher score than aud'; (ii) or $\widehat{aud}$ has the same score as aud' but a better value promotion; or (iii) aud' would be rejected by the co-owners, while $\widehat{aud}$ would be accepted. However, this contradicts the outcome of the choice-of-action stage of the practical reasoning (see Section 5.3.3), because, in order for aud' to be recommended, aud' must have collected only negative answers for the critical questions. This implies that $\widehat{aud}$ cannot exist and aud' is the optimal recommendation.

□

Lemma 4 (**Completeness**)**.** *Assuming the agents' cooperation in the computation, if an optimal audience exists, then ELVIRA finds it and recommends it to the users.*

Proof. If the optimal audience aud' exists, i.e., it has the maximum overall score and the best individual value promotion, then the argument scheme AS in favour of selecting aud' as a solution to the MPC will not be challenged by any other argument. This means that, during the choice-of-action stage in the practical reasoning process, ELVIRA collects only negative answers to the critical questions. Hence, the optimal audience aud' is identified by ELVIRA as the successful output of the practical reasoning and it will be recommended to the users. □

Lemma 5 (Anonimity). *The computation of the solution is not sensitive to permutations of the users, i.e. all the involved users are treated the same.*

Proof. Anonimity is provided by the commutative property of the sum in the Equation (5.4) and in the critical question CQ2 during the practical reasoning, where the order of aggregation of the considered elements is irrelevant. In fact, in Equation (5.4), the sum of the individual scores is independent of whose score that is; in CQ2, the promoted values $v_{Ag,aud}$ of all users are considered equally independently of their users. □

Lemma 6 (Neutrality). *The computation of the solution is not sensitive to permutations of the possible audiences, i.e., all the audiences are considered equally independently of their order.*

Proof. When performing practical reasoning, ELVIRA instantiates the argument scheme AS for every possible audience, and all the audiences are considered when discussing the critical questions. Therefore, the order of consideration of the audiences is irrelevant. □

Furthermore, adaptivity and role-agnosticism allow for the MPC solutions identified through ELVIRA to be *fair*, in the sense that anyone, independently of their role in the conflicts or of their privacy preferences, is supported in the same way.

In fact, while other models in the literature where solutions are identified in a rigid way, e.g., according to majority rule or selecting always the most private policy, may disadvantage users whose privacy preferences are out of the ordinary, ELVIRA considers whatever privacy preference is expressed by a user not on its own, but as a function of how much utility or value promotion it generates for that specific user and for the other users involved. This means that each solution is considered not in terms of how private or public it is (which may be more or less preferable according to the individuals and therefore unfair to aggregate for all the involved users in a single evaluation), but simply in terms of how happy or how satisfied it makes each user.

5.5 Explanations as social process

According to Chander and Srinivasan [40], explanations generated by AI systems should serve some cognitive-behavioural purposes, such as engender the user's trust when accounting for the user's values, or support the user's understanding of the recommendation in order to take appropriate action. However, as Miller stresses in [112] (see Section 2.6), to produce an explanation is a complex task, which involves two complementary processes: a *cognitive process*, i.e. the process of abductive inference determining the causal attribution for a given event, and a *social process*, i.e. the process of transferring knowledge between the explainer and the explainee.

In Section 5.3 I described how the practical reasoning process enables ELVIRA to gather all the necessary information to provide an explanation, i.e. ELVIRA's cognitive process, while accounting for the user's values. In this section, I now describe the steps that led to the definition of the ELVIRA's social process. First I discuss, from a theoretical point of view, the elements that should be included in the explanation for an MPC solution; then, I suggest some different explanation designs, which are evaluated through a user study as described later in Section 5.6.

5.5.1 Design of the explanations

Both Miller [112] and Langley [98] propose that social awareness is necessary for explainable agency. They suggest that a social agent must be able to transfer knowledge from itself (the explainer) to a user (the explainee) in such a way as to give the user the necessary information to understand the causes of its recommendation. This can happen when the agent is able (i) to align its knowledge base with the recipient user; (ii) to tailor the explanation according to the context, including the recipient user's needs; and (iii) to engage in counterfactual explanations, e.g. justifying the rejection of possible alternative actions. In the following, I outline how the design of ELVIRA's explanations meets these requirements.

Conflict description In order to explain the *solution* for an MPC, it is useful to provide details also about other components of the conflict, such as its *detection* and *representation* [181]. This fits the necessity for an explanation to present causal attribution [112]: it is desirable to have an explanation that not only guides the user from causes to effect, but also that describes to the user the causes and the effect. This enables the user to assess whether the agent that is providing the explanation has understood the context and has thus grounded the explanation in a realistic representation. Therefore, ELVIRA includes in the explanation a description of q_0, i.e. the initial conflictual state of the AATS+V.

Tailored explanations As part of the adaptability of the model, I argue that not only the solution but also its explanation needs to be customised and context-related. Every user may have different priorities regarding what is important to them: this influences the way the solution is identified and also the information that is worthy to be included in the explanation. Given the redundancy of reporting ELVIRA's entire PR process, I suggest that the agent could include in the explanation only the elements that regard the optimal solution, that is, the instantiation of the argumentation scheme for aud'. By doing so, the user would be made aware of the benefits of the identified solution in terms of his/her utility and value promotion.

Contrastive explanations Miller [112] clearly highlights the importance of contrastive explanations, because people may in general be not as interested in the causes of selecting the solution aud' per se, as they may be in the causes of not selecting their initial preference aud_k. Therefore, ELVIRA could include in the explanation only the elements that regard aud' in relation to aud_k, that is, the instantiation of the argumentation scheme for aud_k with the positive answers to the critical questions. By doing so, the user would be made aware of the different, and better, consequences of selecting the recommended solution rather than the initial preference.

Given these possible designs, I identified two alternative structures for the output that ELVIRA could generate and present to the users: (i) *general explanation*, and (ii) *contrastive explanation*. Both of them present first a description of the conflict, reporting the different sharing preferences of all the involved users, and then a justification for the solution, highlighting either the benefits of the solution or the positive comparison between the preferred policy and the solution. Practically speaking, for each type of explanation, I propose a rule-based template where the recommended solution, the sharing preference of the user and the value-inspired actions that would be a consequence of the solution, are variables that can be replaced with the appropriate elements when the explanation is instantiated. In Table 5.7 I report the details of the information included in each of these two types of explanation, where O is the optimal sharing policy, P is the user's preferred policy, the actions promoting/demoting the values are like in Table 5.3.

Note that the decision of what to include in the explanations is not a limitation

Conflict description	[*Example with 3 users*] A multi-user privacy conflict to share this content occurred, because the sharing preferences of the involved people do not coincide. You suggested to share {P}; {user1} opted for sharing {P1} and {user2} would like to share {P2}.
No explanation	To share {O} is the best compromise that solves the conflict.
General explanation	To share {O} is the best compromise that solves the conflict because it satisfies as much as possible everyone's initial sharing preference, and because it enables actions that are coherent with everyone's ranking of behavioural tendencies. Notably, by selecting to share {O}, the user would {*list of actions corresponding to the values promoted by selecting* {*O*}}.
Contrastive explanation	To share {O} is the best compromise that solves the conflict because it satisfies as much as possible everyone's initial sharing preferences, and because it enables actions that are coherent with everyone's ranking of behavioural tendencies. [*If* {*O*} *coincides with* {*P*}] This is also your preference! [*Else*] Notice that to share {P} (your initial sharing suggestion) would not allow the involved users to find a compromise, because other users may experience negative consequences. [*If* {*O*} *is more private than* {*P*} *and preference for undersharing*] Also, you said that it would be ok sharing with fewer people. [*If* {*O*} *is more public than* {*P*} *and preference for oversharing*] Also, you said that it would be ok sharing with more people. In addition, by selecting to share {O}, {*list of actions corresponding to the values promoted by selecting* {*O*}} that would not be the case if sharing {P}. Furthermore, by selecting to share {P}, {*list of actions corresponding to the values demoted by selecting* {*P*}}.

Table 5.7: Detailed design of the suggested structures for an explanation which includes the conflict description. {O} is the variable representing the optimal sharing policy; {P} is the variable representing the user's preferred policy; the actions promoting/demoting the values are like in Table 5.3. For the contrastive explanation, when the *if*-conditions are verified (which is optional), then the corresponding sentences are added to the explanation.

of the model: if a dialogue between the user and the agent was developed, the agent would be able to reply to any of the user's objections regarding the selection of alternative solutions based on the model described in Section 5.3.

In the user study which I describe next, I comparatively evaluate the design of the general explanation and of the contrastive explanation with a baseline, namely *no explanation*, where the recommended solution is suggested without motivation after the description of the conflict (see Table 5.7).

5.5.2 Running example

Still considering the same MPC scenario as before, I report below how the conflict description and the three explanations generated by the ELVIRA uploader agent would look like for Alice.

> *Conflict description:* A multi-user privacy conflict to share this content occurred, because the sharing preferences of the involved people do not coincide. You suggested to share with $\langle 2, 2\rangle$; Bob opted for sharing with $\langle 1, 3\rangle$ and Charlie would like to share with $\langle 3, 4\rangle$.
>
> *No explanation:* to share with $\langle 2, 3\rangle$ is the best compromise that solves the conflict.
>
> *General explanation:* to share with $\langle 2, 3\rangle$ is the best compromise that solves the conflict because it satisfies as much as possible everyone's initial sharing preference, and because it enables actions that are coherent with everyone's ranking of behavioural tendencies. Notably, by selecting to share with $\langle 2, 3\rangle$, everyone would compromise the same, everyone's privacy would be preserved and you would get your way.
>
> *Contrastive explanation:* to share with $\langle 2, 3\rangle$ is the best compromise that solves the conflict because it satisfies as much as possible everyone's initial sharing preferences, and because it enables actions that are coherent with everyone's ranking of behavioural tendencies. Notice that to share with $\langle 2, 2\rangle$ (your initial sharing suggestion) would not allow to find a compromise, because other users may experience negative consequences. Also, you said that it would be ok sharing with fewer people. In addition, by selecting to share with $\langle 2, 3\rangle$, everyone would compromise the same and everyone's privacy would be preserved, that would not be the case if sharing with $\langle 2, 2\rangle$. Furthermore, by selecting to share with $\langle 2, 2\rangle$, you would not make others happy and everyone would not compromise the same.

5.6 Evaluation of the explanations

In this section, I present the within-subjects user study that I designed and conducted in order to evaluate the design of the explanations that ELVIRA can generate: the baseline explanation (*exp0*), the general explanation (*exp1*), and the contrastive explanation (*exp2*). For the full specification of the experiment design, including the scenarios and questions presented to participants, the generated explanations and the collected data, see Appendix A.

The results of this study informed the final design of ELVIRA, which I evaluated in another user study against other models suggested in the literature, as I report in Chapter 7. Participants were recruited through Prolific[iii] and the study received ethical approval by the Ethical Board of King's College London (see Appendix B).

5.6.1 User study design

I developed a web application in Python to conduct the experiment. After eliciting the participants' moral values, the application generated some MPCs and provided for each of them, in a random order, the three alternative outputs from Table 5.7, that the participants were required to comparatively evaluate.

Values elicitation I relied on the Portrait Value Questionnaire (PVQ) designed by Schwartz [154] to elicit the value preferences of the users. As I mentioned in Section 2.3.1, among the tools suggested by Schwartz, this is the most appropriate for a broad audience and can easily be delivered online. Specifically, participants were faced with the PVQ-21, which presented 21 sentences describing behaviours of people and asked them to report how similar these people were to themselves. This version of the PVQ has been very commonly used in social studies and has been included in the European Social Survey [156] since 2002. After eliciting the

[iii] `https://www.prolific.co`

Scenario	Relationship	Sensitivity
1	colleagues	low
2	colleagues	high
3	friends	low
4	friends	high
5	family	low
6	family	high

Table 5.8: Details of the scenarios considered in the user study.

preferred order over the ten Schwartz basic values, ELVIRA computed the equivalent preferred order over the hyper-values by averaging the scores of the corresponding basic values.

Scenarios I followed an immersive scenario approach [106], which was successfully used in previous work in MPCs [173, 59], in order to elicit the participant's behaviour in MPC situations. I selected six scenarios, consisting of pairs of one picture and one description, among the ones similarly used by Fogues et al. [59][iv]. Each scenario is representative of different sensitivities (low/high) and relationship types (colleagues, friends and family), see details in Table 5.8 and in Appendix A.

MPCs Each participant was shown three randomly selected scenarios. For each scenario the participant was asked to put him/herself in the shoes of one of the depicted people and provide the following: (i) their preferred sharing policy among keeping it private, sharing with common friends, sharing with friends of friends, or sharing publicly; and (ii) their appreciation, i.e., whether they would be ok with over-sharing or under-sharing, each with a 5-point Likert scale anchored with 'very happy' and 'very unhappy'. For simplicity, the sharing policies were defined as group-based policies, which, as aforementioned and shown in [173], are equivalent to the distance-intimacy policies we used in earlier parts of this thesis, and *as-it-is* modality, because

[iv] The picture for scenario 1 is different than in [59], as I could not recover the one they used, but equivalent in terms of content and sensitivity.

they are both (policies and modalities) more familiar and intuitive for users, as that is what they currently see in mainstream online social networks [115]. Then, the application randomly generated the preferences and appreciation of two (non-participant) users involved in the scenario, making sure that an MPC was created (e.g. at least one preference would be different from the one of the participant). Note that, even if the photos and descriptions were the same, many more than just six scenarios were randomly generated, because each involved user (one participant and two simulated ones) could have one of 4 policies, one of 5 different appreciation levels, and one of 24 orders over values. Finally, the MPC was presented to the participant together with the three alternative explanation types, listed in a random order.

Satisfaction For each MPC that was presented to the participant, I asked about their satisfaction with the alternative explanation types. To measure satisfaction, I used the Satisfaction Scale proposed in [79] (see Table 5.9). This scale, based on studies in cognitive psychology, philosophy of science, and other pertinent disciplines, is meant to evaluate explanations by considering the features that make explanations good (e.g., level of detail, usefulness, accuracy, etc.). It includes 8 questions with a 5-point Likert scale anchored with 'strongly agree' (2) and 'strongly disagree' (-2). After running a pre-test, I decided to add an extra question that asked the participant to select the preferred explanation type among the three options.

Data quality measures To maximise data quality, I employed two well-known methods: attention check questions, and participants' previous performance [108, 136, 77, 133]. I recruited participants from Prolific with at least 100 submissions and an approval rate of 95% according to [136]. Also, during the experiment, the application presented participants with four attention check questions.

Satisfaction Scale
1. From the explanation, I could understand how ELVIRA works.
2. The explanation I received is satisfying.
3. The explanation provided sufficient detail about how ELVIRA works.
4. The explanation provided complete information about how ELVIRA works.
5. The explanation tells me how to use ELVIRA.
6. The explanation that ELVIRA provided is useful to my goals.
7. The explanation showed me how accurate ELVIRA is.
8. The explanation let me judge when I should trust and not trust ELVIRA.

Table 5.9: The Satisfaction Scale [79].

Age	'18-25': 35.9%, '26-35': 32.8%, '36-45': 23.5%, '46+': 7.8%
Gender	'Male': 40.6%, 'Female': 59.4%
Nationality	'UK': 26.6%, 'Portugal': 17.2%, 'Poland': 10.9%, 'Spain': 6.3%, 'Italy': 4.7%, 'USA': 4.7%, 'Mexico': 4.7%, other: 24.9%
Student status	'No': 65.6%, 'Yes': 34.4%
Social media use	'Daily': 92.2%; '2-3 times/week': 4.7%; less often: 3.1%
Privacy	'Not concerned': 4.7%; 'Concerned': 57.8%; 'Very concerned': 37.5%

Table 5.10: Demographics of participants.

5.6.2 User study results

I recruited a total of 68 participants, who were rewarded £3.00 for completing the survey, which took on average 25.3 minutes (median 20.9 minutes). I discarded 3 participants which failed at least one attention check question (4.4%) and one participant for a technical issue that led to some missing data. I conducted the analyses on the remaining 64 participants, for a total of 192 MPCs. The resulting dataset is publicly available at `https://osf.io/nrgtv/`. Table 5.10 reports the demographic distribution of the participants, including their privacy attitudes measured with the IUIPC scale [105] and social media use.

Overall satisfaction Figure 5.5 shows the evaluation through the Satisfaction Scale [79] of the three types of explanations when considering the total of 192 MPCs. I performed a Multivariate ANalysis Of Variance (MANOVA) to compare differences

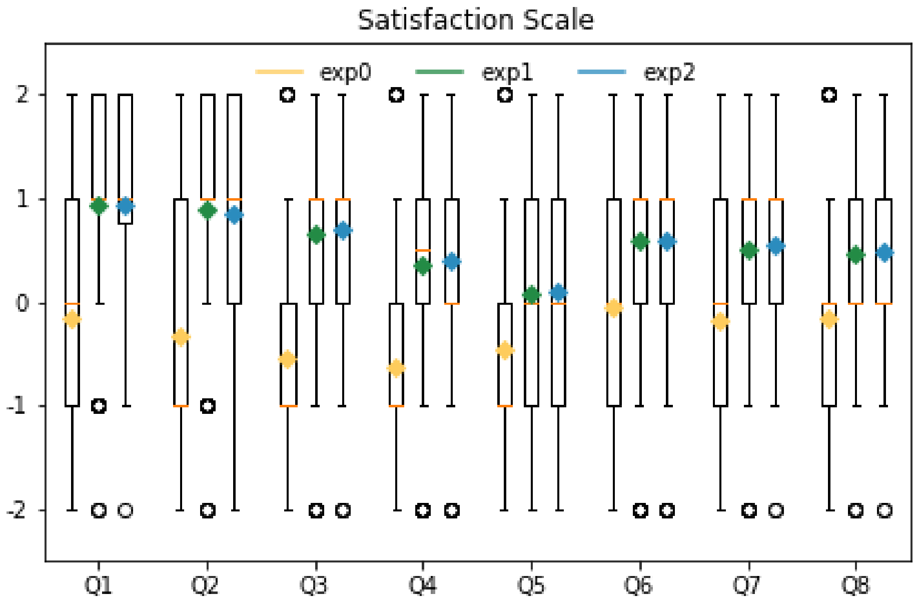


Figure 5.5: Satisfaction Scale considering all the conflicts.

in the mean scores of the Satisfaction Scale between the three types of explanations, which resulted to be significant (F = 12.81, p-value $< .05$; Wilk's $\Lambda = .717$, partial $\eta^2 = .153$). To determine how the dependent variables (i.e., the scores) differ for the independent variable (i.e., the explanation type), we need to look at the Tests of Between-Subjects Effects (see Table 5.11). More than 80% of the variance is associated with the first four questions, which I conclude being the most important main effects. Furthermore, I am interested in which specific explanations' means differ from each other. A Tukey Test, which is essentially a t-test, except that it corrects for family-wise error rate, shows that both *exp1* and *exp2* performed significantly better (p-value$< .05$) than *exp0* across all the questions, but no significant difference was detected between *exp1* and *exp2*.

General vs. contrastive In order to identify situations where one type of explanations may be preferred over another, I considered the Satisfaction Scale when splitting the dataset in complementary portions, according to whether (a) the solution of the MPC coincided with the participant's preferred policy (71 conflicts) or

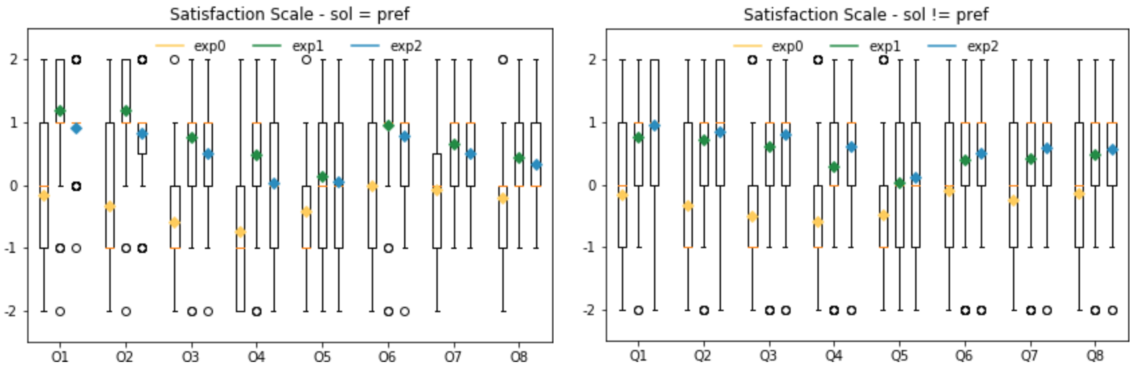
Source	Dependent Variable	F	p-value	partial η^2
expl	q1	76.044	.000	.210
	q2	76.981	.000	.212
	q3	87.488	.000	.234
	q4	56.093	.000	.164
	q5	15.612	.000	.052
	q6	24.887	.000	.080
	q7	30.537	.000	.096
	q8	26.410	.000	.084

Table 5.11: Tests of Between-Subjects Effects.

(b) the solution was different from the participant's preference (121 conflicts) (see Figure 5.6). Similarly as before, MANOVA tests showed significantly different distributions in both subsets: (a) F = 6.73, p-value < .05; Wilk's Λ = .625, partial η^2 = .21; (b) F = 7.694, p-value < .05; Wilk's Λ = .725, partial η^2 = .148. Tukey tests proved that both the general and the contrastive explanations still outperformed significantly the baseline in both subsets (p-value < .05). It is possible to notice here a general trend that made participants prefer *exp1* when (a) the solution coincided with their preference and prefer *exp2* when (b) the solution was different from their preference. This trend resulted to be a significant difference only in (a) (p-value=.034) and almost significant in (b) (p-value =.057), when considering Q4: "The explanation provided *complete* information about how the tool works.". I did not identify any other features (e.g., demographics, privacy concerns, scenarios, etc) that led to significant differences in the preference for *exp1* or *exp2*.

5.6.3 Conclusions of the user study

I summarise the above findings with three intuitions regarding the design of the explanations that ELVIRA autonomously generates. First, participants overall seem to appreciate receiving extra information that explains or justifies the recommended solution. Second, when presented with a solution that coincides with their initial



(a) When the recommended solution coincides with the user's preference.

(b) When the recommended solution is different from the user's preference.

Figure 5.6: Satisfaction Scale on subsets of the dataset.

preference, participants seem to appreciate the description of the positive consequences of selecting that policy, almost as a way of reinforcing their choice, rather than comparing or contrasting it with others. Third, when the recommended solution is different from the participant's preference, participants seem to favour contrastive explanations, i.e., they seem interested in knowing why their preference is not recommended rather than in the reasons for selecting the audience suggested.

Furthermore, following the precious comments that some colleagues provided in order to finalise the explanation design, I (i) simplified the wording of the conflict description ("The sharing preferences of the other people involved do not coincide with yours. You suggested to share P; user1 opted for sharing P1 and user2 would like to share P2." replaces what is in Table 5.7); and (ii) labelled for clarity the components of the output that ELVIRA generates ("*Conflict:*" followed by the conflict description and "*Solution:*" followed by the tailored recommendation).

In conclusion, the final design of the explanations generated by ELVIRA corresponds to a *hybrid tailored explanation* structure, where the agent typically provides a contrastive explanation whenever the solution does not coincide with the user's preference, and a general explanation otherwise.

5.7 Conclusion

In this chapter I have introduced ELVIRA, an agent-based model that can help OSNs users collaboratively manage multi-user privacy online, by recommending group sharing policies that are most aligned with the preferences of all the users involved in an MPC.

Following a Value Sensitive Design approach (VSD, see 2.4), ELVIRA is designed in order to satisfy all the requirements identified in the literature to effectively solve MPCs (see Table 1.1 in Section 1.3). In particular, the agent is both utility-driven and value-aligned (see Sections 5.2.1 and 5.2.2): regarding the utility, the agent considers the gain/loss perceived by the user when desired/undesired audience is able to access the content; regarding the moral values, the agent favours the group-deliberation behaviours that promote the values most important to the user. Then, all the agents representing users involved in the MPC collaboratively contribute to identify the optimal solution —recall that the assumption of non-adversarial behaviour is supported by empirical evidence (e.g., [175]): all the users' preferences are considered equally, thus satisfying *role-agnosticism*, and directly influence the dynamic identification of the solution, thus satisfying *adaptivity*, as I formally prove in Section 5.4.

Finally, by performing practical reasoning to identify the optimal solution, ELVIRA is able to generate explanations that justify its recommendation for the user. Such explanations, which were carefully designed and evaluated in a user study (see Section 5.6), have a tailored *hybrid* format: this means that, according to the nature of the solution, i.e., whether it coincides or not with the user's initially preferred sharing policy, the explanation focuses on highlighting either the benefits of the solution (*general explanation*) or the comparison between the solution and the user's initial preference (*contrastive explanation*).

After the preliminary definition of the JIMMY agent (see Chapter 4) and an update of the *conceptual investigation* within the VSD approach that identified new requirements for successfully solving MPCs, this chapter has reported on a second iteration of the *technical investigation*, complemented by a first empirical investigation for the explanations design. In the next chapter, Chapter 6, I will further complement the technical investigation with an analysis through software simulations of the benefits of considering both utilities and moral values when computing a solution for an MPC.

Chapter 6

Software Evaluation

6.1 Introduction

In Chapter 5 I have introduced a novel agent-based model, namely ELVIRA, to collaboratively solve multi-user privacy conflicts (MPCs) in online social networks (OSNs). ELVIRA has been designed with the aim of satisfying all the properties, introduced in Section 1.3 Table 1.1, that have emerged as crucial during the conceptual investigation within the Value Sensitive Design approach (VSD, see Section 2.4), that has driven the work presented in this thesis.

In line with the two-fold objective of the VSD *technical investigation*, I have analysed how the design of ELVIRA effectively fulfils its requirements. In particular, in Section 5.4 I have formally proven how role-agnosticism and adaptivity are guaranteed, and in Section 5.6 the effectiveness of explainability is discussed at length on an empirical basis. The advantages of considering both utility and moral values when computing the solution to an MPC, however, are still to be investigated and are the topic of this chapter.

Recall that empirical evidence [94, 175] suggests to consider both utility and values when solving an MPC, in order to better mimic the way users make privacy decisions in the real world. In fact, people may be willing to transcend their own advantage in order to accommodate someone else's preferences, but not in a blind manner, meaning that severe privacy violations or missed sharing benefits are usually taken into account. However, the approaches to identify solutions for MPCs that have been previously suggested in the literature (see Section 3.2.2 and the JIMMY agent in Chapter 4) mostly focus on the maximisation of either utility or value promotion, differently from ELVIRA, which considers both.

In the following, I present a comparative analysis through software simulations of ELVIRA (EL) and three other models inspired by approaches discussed in the literature or available in the real world:

- *Utility-based* (UB): selects the audience that maximises utility for all the involved users, similarly to works only utility-driven;
- *Value-based* (VB): selects the audience that maximises the promotion of values for all the involved users, similarly to works only value-driven;
- *Facebook* (FB): selects the uploader's preferred audience, i.e., neither utility- or value-driven, similarly to what currently happens in Facebook and other OSNs.

I compared the performance of EL, UB, VB and FB in two different types of experiments: i) experiments on synthetic data, which allow to compare the models varying all the relevant parameters and understand the influence they have on MPC solutions; ii) experiments on real data, which allow to compare the models in realistic social networks. In particular, I considered different social networks (in terms of size N and connectivity d), numbers of users involved (n) in an MPC, numbers of MPCs (T), and values of the parameters α and β; I also considered different MPCs, by varying the users' preferred audiences, their appreciation for the content to be shared, and their moral values.

In order to compare the performance of the four models, I have defined different metrics, which consider for each model M the generated utility and value promotion either in the individual conflict or cumulatively across series of conflicts:

- the individual average variation of utility (*iauc*), normalised over the size of the network, per each conflict:

$$iauc_M = \frac{1}{nTN} \sum_{k \in U_t, t<T} u_{kt,M};$$

- the individual average of value promotion (*iavc*) per each conflict:

$$iavc_M = \frac{1}{nT} \sum_{k \in U_t, t<T} v_{kt,M}$$

- the cumulative increment of social utility (*csu*):

$$csu_{t,M} = csu_{t-1} + \sum_{k \in U_t} u_{kt,M}$$

- the cumulative increment of value promotion (*csv*):

$$csv_{t,M} = csv_{t-1} + \sum_{k \in U_t} v_{kt,M}$$

where U_t are the users involved in the conflict generated at time t and $u_{kt,M}$ and $v_{kt,M}$ are the variation of utility and of value promotion, computed respectively as indicated in Equations 5.1 and 5.2, which the user k gets when selecting the solution suggested by the model M in the conflict t.

I implemented the models in Python 2.7.10 (*numpy* 1.16.2; *networkx* 2.2) and I ran all the simulations on Windows 10 64-bit, Intel(R) Core(TM) i7-7700 CPU @ 3.60GHz 16GB. In each network, intimacies were generated randomly in the range $[1, 5]$, which is the intimacy scale used by Fogues et al. [61], where 1 represents a mere acquaintance and 5 a very close relationship. Regarding the value preference, I generated randomly for each node a total order over $\mathcal{V}$, which remained static for the entire simulation; this is coherent with the individual value preference being relatively stable over the human lifetime [24]. For each simulation, an MPC among n random connected users was created, with sharing policies and appreciation functions also generated randomly. In particular, distances were in the range $[0, 5]$, which captures the vast majority of cases reported about the degrees of separation between users on Facebook[i]. Also, to generate audience aud_f, I randomly selected a tuple of distance and intimacy so that each element was contained in the range identified by the minimum and the maximum distance and intimacy of the users' preferences, but the tuple was not already contained in the set of possible solutions:

$$aud_f = \langle d_f, i_f \rangle : \quad d_f \in [\min_{\mathcal{A}} d, \max_{\mathcal{A}} d], \quad i_f \in [\min_{\mathcal{A}} i, \max_{\mathcal{A}} i],$$
$$aud_f \neq aud_k \quad \forall k \in Ag.$$

Finally, I studied different implementations of the appreciation function, by considering the random selection of just extreme values, i.e. $app = \pm 1$, or randomly selecting values from a fixed range.

6.2 Experimental settings

Here I detail the settings of each experiment. Experiments I-IV simulate MPCs over synthetic social networks, which I generated according to the *scale-free network* model by Barabasi-Albert with preferential attachment [23], where the degrees of the nodes follow a power-law distribution, in order to reproduce scenarios that would resemble as much as possible real online social networks [114]. Experiment V simulate MPCs over portions of a real social network, namely Facebook, that are available in the literature [196, 99].

Experiment I In this experiment I studied the performance of EL, UB, VB and FB after solving $T = 300$ conflicts when increasing the size of the network from

[i] https://research.fb.com/blog/2016/02/three-and-a-half-degrees-of-separation/

$N = 100$ up to $N = 2500$ while maintaining $d = 10, n = 3$, $app = \pm 1$ (random), $\alpha = 0.9$ and $\beta = 0.1$.

Experiment II In order to see the effect of other parameters in addition to the size of the network, in this experiment, I compared the models considering $N \in \{100, 200, 300, 400, 500\}$, $d \in \{10, 20, 30, 40\}$ and $n \in \{2, 3, 5, 10\}$, after $T = 1000$ conflicts, while maintaining constant the values of $\alpha = 0.9$ and $\beta = 0.1$, and $app = \pm 1$ random.

Experiment III In this experiment I evaluated how the appreciation of the content to be shared influences the average utility obtained by the user after a number of conflicts. In particular, I compared the utility generated in Experiment II when selecting randomly only the extreme values of appreciation ($app \in \{-1, +1\}$) with the utility generated when selecting randomly also intermediate values ($app \in \{-0.9, -0.45, 0, 0.45, 0.9\}$). I maintained all the other settings as in Experiment II.

Experiment IV In this experiment I studied the impact of selecting the audiences *as-it-is* or *modified*, by varying the parameters α and β. I considered $\langle \alpha = 0.9, \beta = 0.1 \rangle$, $\langle \alpha = 0.5, \beta = 0.5 \rangle$ and $\langle \alpha = 0.1, \beta = 0.9 \rangle$, in order to represent situations in which the excluded audience (both desired and extra, see sets C and D in Definition 13) has different influence on the utility. I simulated $T = 500$ conflicts with different N, d and n combinations and each conflict was solved with the three different configurations of α and β.

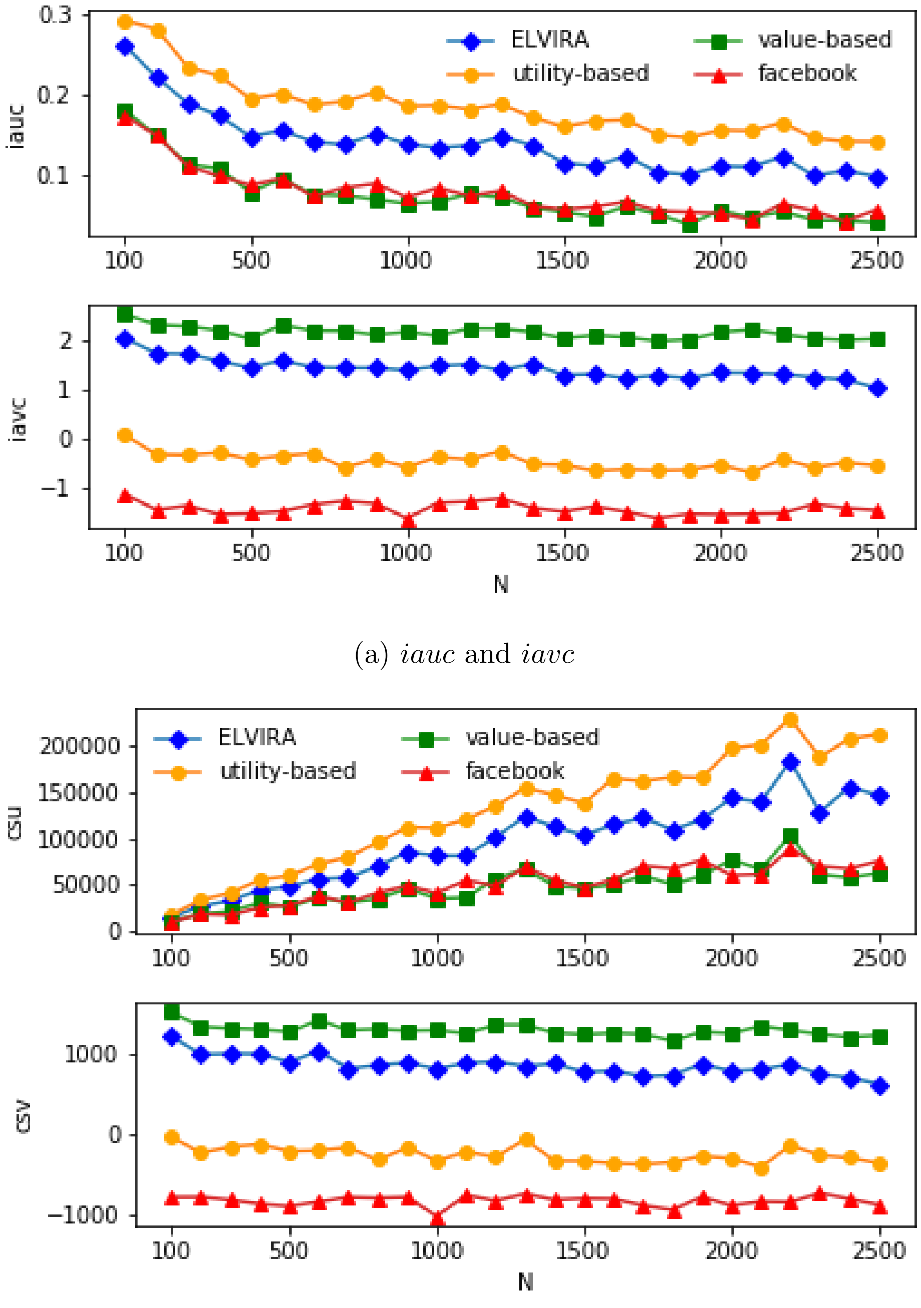
Experiment V Here I simulated conflicts on graphs corresponding to real portions of Facebook — number of nodes and edges in parenthesis: $G_1 = (769, 16656)$ and $G_2 = (1446, 59589)$ from [196], and $G_3 = (4039, 88234)$ from [99]. I maintained $\alpha = 0.9$ and $\beta = 0.1$ and $app = \pm 1$ (random), and I generated $T = 500$ MPCs among $n = 3$ random users on each graph, following the rationale that, as shown in Experiment II, the models perform similarly regardless of the number of users n involved in the MPC from 2 to 10 users, which covers the vast majority of cases regarding the number of people depicted in photos [84, 175].

6.3 Experimental results

Here I report the results of the experiments described above.

Experiment I Figure 6.1a shows the *iauc* and the *iavc* generated by each model after solving $T = 300$ MPCs. Figure 6.1b shows the *csu* and *csv* generated at the network level after $T = 300$ conflicts. Despite few peaks and drops, which may be due to the randomness of the system and therefore may smooth after generating more conflicts, a clear trend is recognisable, where ELVIRA represents the best trade-off between utility and value promotion. In particular, one can easily see how UB and FB suffer massively in terms of value promotion and VB and FB in terms of utility. The cumulative utility increases, not surprisingly, with the size of the network: therefore, in the next experiments I focus only on *iauc* and *iavc*, which I consider more significant to evaluate the performance of the models.

Experiment II Figure 6.2 shows, as an example of the performance of the four models, the results when varying the connectivity d from 10 to 40 while keeping $N = 500$ and $n = 5$. Over a more connected graph, users can in general achieve higher utilities. Table 6.1 reports an overview of the results and their statistical significance when performing pairwise t-tests (I marked with * the differences with p-value$< .05$). The results show the same similar constant trend as in Figure 6.1. Regardless of the scenarios, UB always generated the maximum *iauc*, but guaranteed a poor promotion of moral values; VB always generated the maximum *iavc*, but with



(a) *iauc* and *iavc*

(b) *csu* and *csv*

Figure 6.1: Performance of the four models in Experiment I.

Exp.	Utility (*iauc*)			Value Promotion (*iavc*)		
	ELvsUB	ELvsVB	ELvsFB	ELvsUB	ELvsVB	ELvsFB
II,III: n=2	$<$ *	$>$ *	$>$ *	$>$ *	$<$ *	$>$ *
II,III: n=3	$<$ *	$>$ *	$>$ *	$>$ *	$<$ *	$>$ *
II,III: n=5	$<$ *	$>$ *	$>$ *	$>$ *	$<$ *	$>$ *
II,III: n=10	n.s.	$>$ *	$>$ *	$>$ *	$<$ *	$>$ *

Table 6.1: ELVIRA's performance in Experiment II and III: better (>), worse (<), or not significantly different (n.s.) from the other models.

very low individual utilities; and EL represented the best utility-value trade-off. By increasing n, the distance between the utility generated by EL and UB decreased, with $iauc_{EL}$ and $iauc_{UB}$ being not significantly different when $n = 10$, while the gap with VB and FB increased. This suggests that EL might reach optimal utilities if increasing further the number of conflicting users.

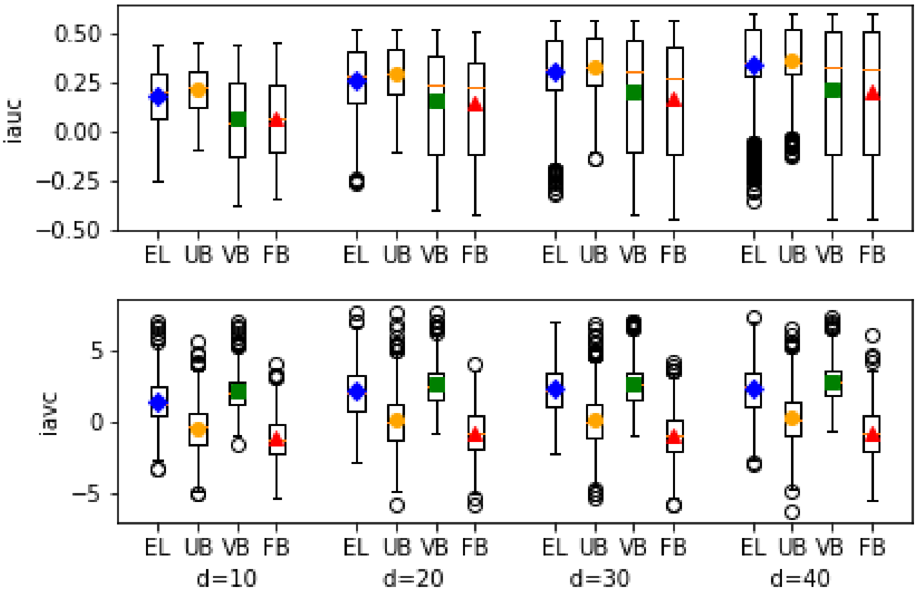


Figure 6.2: Performance of the four models in terms of *iauc* and *iavc* when varying d, with $N = 500, n = 5, T = 1000$.

Experiment III As reported in Table 6.1, when considering appreciation in a range of values ($app \in \{-0.9, -0.45, 0, 0.45, 0.9\}$), the models performed in the same way as in Experiment II, where only extreme values were allowed ($app \in \{-1, +1\}$): EL always generated a significantly worse *iauc* than UB (with the only exception of $n = 10$), but better than VB and FB, and EL always generated a significantly worse *iavc* than VB, but better than UB and FB. When comparing the *iauc* generated by ELVIRA after $T = 1000$ in the experiments II and III, I noticed that considering also intermediate values of appreciation tended to provide higher utilities, but no significant differences were observed.

Experiment IV I simulated $T = 500$ conflicts for different N, d and n combinations (n=2, N=100, d=10; n=3, N=100, d=10; n=3, N=100, d=40; n=3, N=300, d=40; n=5, N=100, d=40; N=300,n=3,d=20) and solved them with the three different pairs of α and β. In all cases, the behaviour of the models was coherent with what discussed in the previous experiments: ELVIRA produced sub-optimal *iauc* and *iavc*, and guaranteed their best trade-off. Regarding the comparison of the *iauc* generated by ELVIRA with the different α,β combinations, there were no significant differences. This suggests that there is not evident impact on the generated utility when the excluded audience does not access the content or accesses a modified version of it.

Experiment V Figure 6.3 displays the performance of the models in terms of *iauc* and *iavc*. Pairwise t-tests of EL with the other three models show significant differences between the distributions with p-value$<$ .01. The effect size of the comparison between the models is medium or large in all cases (average over the three graphs): (i) regarding *iauc*, ELvsUT: -.29, ELvsVA: .32, ELvsFB: .35; regarding *iavc*, ELvsUT: 1.09, ELvsVA: -.38, ELvsFB: 1.60. ELVIRA confirms to offer the best trade-off between maximisation of individual utility and promotion of the users' values over all the three networks. Regarding G_3, the results seem lower than the ones from G_1 and G_2, but this is due to the normalisation of *iauc* over a much bigger

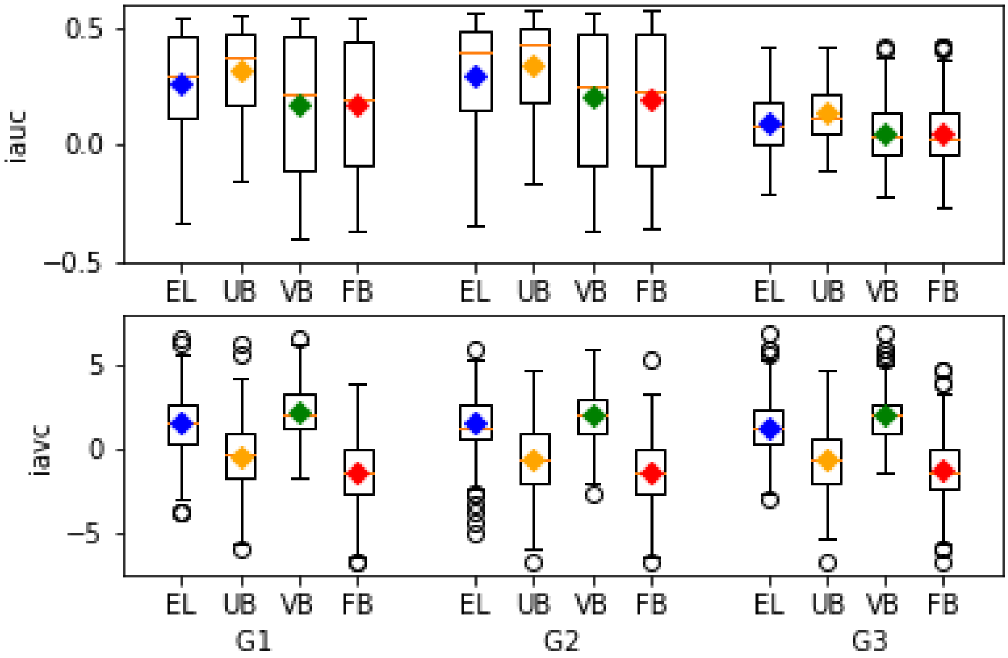


Figure 6.3: Comparison of the performance of the four models in terms of *iauc* and *iavc* generated on G_1, G_2 and G_3.

graph.

6.4 Conclusion

In this chapter I have progressed with the technical investigation of the ELVIRA model within the VSD approach. In particular, I have shown how an architecture that considers both utilities and value promotion when computing the solution for an MPC can be more beneficial for a user than other models previously suggested in the literature.

In order to do this, I have simulated the performance of ELVIRA and other three models, which resemble some approaches from the literature or the real world in terms of utility generation and value promotion. Across all the experiments, one can clearly see that these models always behaved according to a constant trend.

On the one hand, the utility-based approach outperformed the others in terms of the utility generated (both individual average, *iauc*, and social cumulative, csu_t), but with very disappointing value promotion. On the other hand, the value-based approach produced the solutions which were the most coherent with the values of the users involved in the simulated conflicts, but very poor in terms of generated utility. The Facebook approach selected the solutions with the least generation of utility and the worse value promotion. ELVIRA represented, among this selection of models, the best utility-value trade-off, by producing utilities very close to UB and value promotion close to VB. Future studies may focus on the exploration and comparison of other mechanisms to aggregate utility and value promotion, beyond their multiplication (see Equation (5.3) for the computation of the individual score).

In conclusion, the design of the ELVIRA agent fulfils all the requirements to successfully solve MPCs in OSNs that I introduced in Section 1.3. However, this is not a guarantee that the model would be actually appreciated and considered useful by OSNs users, which is the final goal of this work. Therefore, in the next chapter I describe the *empirical investigation*, still within the VSD approach, that I performed in order to evaluate whether the recommendations generated by ELVIRA are received in a more positive way than the ones generated by other models in the literature, such as UB, VB and FB.

Chapter 7

Evaluation - User Study

7.1 Introduction

In the previous chapters, I have described ELVIRA, an agent-based model that can support online social networks (OSNs) users when managing multi-user privacy online. ELVIRA's design, developed following a Value Sensitive Design (VSD) approach, is informed by some requirements suggested in the literature on privacy and autonomous systems, namely it is role-agnostic, adaptive, explainable, and both utility-driven and value-aligned. Theoretically, a model that satisfies all these properties is believed to provide satisfactory recommendations for its users. In this chapter, I describe the user study I performed in order to evaluate this hypothesis. In other words, I report on the final stage of the VSD approach, i.e., the empirical investigation, whose objective is to inform the designer about the user's perception of the developed technology.

In particular, I have designed and conducted a between-subjects user study with a two-fold goal: (i) to study the user acceptability of the recommendations identified by ELVIRA, comparing it to existing approaches; and (ii) to understand whether the cognitive and social processes introduced in Sections 5.3 and 5.5 allow ELVIRA to *convey* the recommendations in a more satisfactory way than existing approaches. Similarly to the software simulations that I presented in Chapter 6, I compared the performance of ELVIRA (EL) with three other models inspired by related work approaches: *utility-based* (UB), *value-based* (VB) and *Facebook* (FB).

Given the similarity of the models' performance across different settings, as I analysed in the previous chapter, for the evaluation through user study I decided to maintain $n = 3$, $\langle \alpha = 0.9, \beta = 0.1 \rangle$ and the appreciation in a range of possible values. For the full specification of the experiment design, including the scenarios and questions presented to participants, see Appendix A. Participants were recruited via Prolific[i], and the study received ethical approval by the Ethical Board of our university (see Appendix B).

7.2 User study design

In order to conduct this experiment, I developed another web application in Python similar to the one whose design is described in Section 5.6.1; hence, for the unchanged design details I refer the reader to that section in Chapter 5.

The application randomly assigned each participant to one treatment (between-subjects user study): ELVIRA, utility-based, value-based, and Facebook. For all treatments, the application proceeded as follows: i) first, participants were presented with multi-user privacy conflicts (MPCs) scenarios automatically generated by the tool, then, they were given the recommendations suggested by the model used in the particular treatment and asked about the acceptability of the recommendations; ii) finally, after all the scenarios, participants were asked about their satisfaction with the model of their treatment. In addition, the treatments for ELVIRA and the value-based model also included a preliminary phase to elicit the value preferences of participants through the Schwartz questionnaire PVQ-21 [156] (see Section 2.3.1 and 5.6.1). I now describe the different steps further.

[i] https://www.prolific.co/

Scenarios and MPCs I followed the same immersive scenario approach as described in Section 5.6.1. The application considered the same six scenarios (picture and description, – see [59]) and presented all of them in a random order to each participant. After eliciting the participant's sharing preference (group-based, chosen from sharing *among themselves*, *with common friends*, *with friends of friends* or *publicly*) and appreciation (i.e., happiness with over- and under-sharing, each on 5-point Likert scale), the application randomly generated the preferences and appreciation of other two (non-participant) users involved in the scenario, making sure that an MPC was created. The MPC was then presented to the participant together with the recommendation to solve it that was computed by the model of the participant's treatment (see Table 7.1). The output generated by ELVIRA corresponds to the hybrid tailored explanation described in Section 5.6.3. The utility-based and value-based models communicate the occurrence of a conflict and recommend a solution according to the works in the related literature that follow these approaches (see Section 3.2.2). The Facebook model simulates what happens in Facebook: an uploader, randomly selected among the involved users, shares the picture with the uploader's preference. Finally, the participant was asked to say how likely they would be to accept the recommendation as an individual, and how likely they thought the other involved users would accept the recommendation. Acceptabilities were given as 5-point Likert scales anchored with 'very likely' (2) and 'very unlikely' (-2).

Satisfaction After all the MPCs were presented to the participant, and as a final step, they were asked about their satisfaction with the model they had engaged with across the MPCs in terms of the output that the models generated (rather than just the acceptability of the recommendations). In order to measure satisfaction, I used again the Satisfaction Scale proposed by [79] (see Section 5.6.1 for details).

Model	Output
UB, VB	*Conflict:* The sharing preferences of the other people involved do not coincide with yours. *Solution:* The conflict would be solved by sharing {P}.
FB	{UserUploader} uploads this content online and shares it with {UploaderPolicy}.

Table 7.1: Outputs generated by the models: {P} is the sharing policy identified as a solution by UB or VB; {UserUploader} and {UploaderPolicy} are respectively the name and the preferred policy of the user defined as uploader in the FB treatment.

Data quality measures Similarly to Section 5.6.1, in order to maximise data quality, I employed both attention check questions and participants' previous performance. I recruited participants from Prolific with at least 100 submissions and an approval rate of 95%. Also, during the experiment, the application presented participants with three attention check questions.

7.3 User Study Results

I recruited 470 participants, who were rewarded £2.50 for completing the survey, which took on average 23.71 minutes (median 20.58 minutes). I discarded participants who failed at least one attention check question (28.7%), and analysed the remaining 335 participants. The resulting dataset is publicly available at `https://osf.io/v9z4s/`. Table 7.2 reports the demographic distribution of the participants, including their privacy attitudes, measured with the IUIPC scale [105], and social media use. The final split of the participants per treatment (recall this was done randomly) was: 85 ELVIRA, 82 utility-based, 85 value-based, and 83 Facebook.

Acceptability of recommendation Figure 7.1 shows the distribution of individual and collective acceptability for each model (2='Very likely', -2='Very unlikely'). The stars ($\star$) on the bottom mark the distributions that are significantly worse

Age	‘18-24’: 28.5%, ‘25-30’: 22.2%, ‘31-40’: 24.0%, ‘40+’: 25.3%
Gender	‘Male’: 55.1%, ‘Female’: 44.6%, ‘Rather not say’: 0.003%
Nationality	‘UK’: 41.6%, ‘USA’: 16.2%, ‘Poland’: 9.9%, ‘Portugal’: 6.3%, ‘Greece’: 4.8%, ‘Italy’: 2.7%, ‘Spain’: 2.1%, ‘Canada’: 2.1%, other: 14.3%
Highest education	‘Grad degree’: 27.3%, ‘Undergrad degree’: 32.9%, ‘Tech/community college’: 8.4%, ‘Secondary education’: 29.6%, other: 1.8%
Social media use	‘Daily’: 85.7%; ‘2-3 times/week’: 9.8%; ‘Once a week’: 1.8%; ‘Less than once a week’: 2.7%
Privacy	‘Not concerned’: 3.3%; ‘Concerned’: 54.0%; ‘Very concerned’: 42.7%

Table 7.2: Demographics of participants.

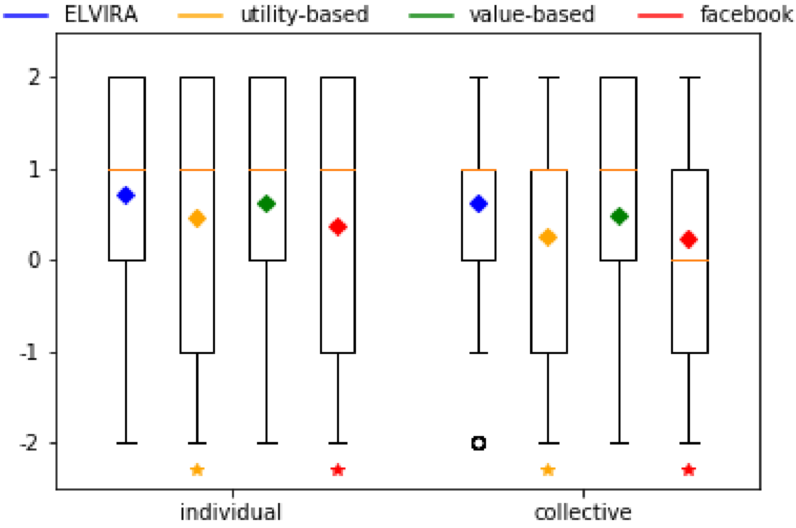


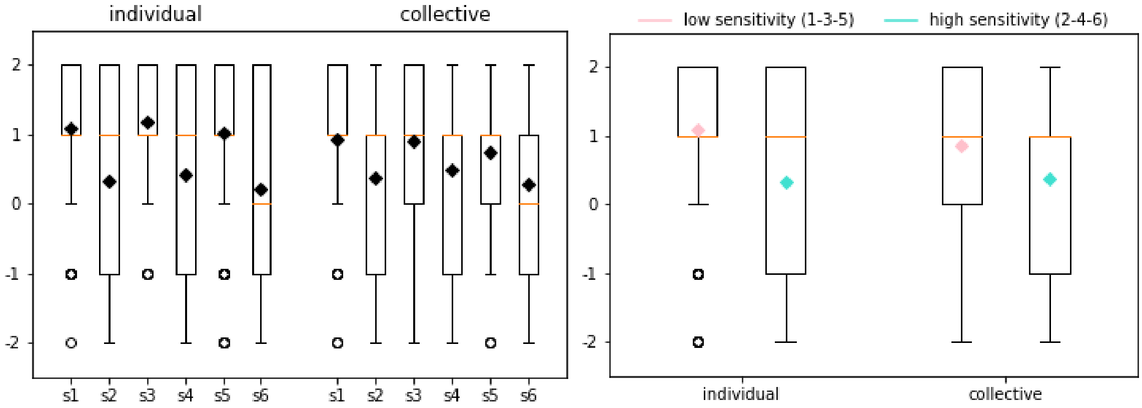
Figure 7.1: Individual and collective acceptability of the recommendations presented by each model.

than ELVIRA, when considering pairwise t-tests with p-value< .05 (effect size for individual acceptability: ELvsUT: .18, ELvsFB: .25; for collective acceptability: ELvsUT: .29, ELvsFB: .3). We can see that the recommendations generated by ELVIRA were significantly more accepted than those generated with utility-based or Facebook models.

In general, the value-based model shows a performance not significantly different from ELVIRA’s. However, there were cases where ELVIRA’s recommendations were significantly more accepted, considering both individual and collective acceptability, than the value-based ones: for participants older than 40yo (p-value$< .01$, effect size$= 0.37$); for participants who had previously experienced MPCs as co-owners (p-value$< .05$, effect size$= 0.25$); and for users accessing social media less than daily (p-value$< .06$, effect size$= 0.32$). Regarding only individual acceptability, ELVIRA performed better when the recommended solution coincided with the participant's preference (p-value$< .05$, effect size$= 0.27$). Finally, considering only the collective acceptability, we see that ELVIRA's outputs were more acceptable when the participant was mediumly privacy-aware (awareness score from IUIPC score in $[0.5, 1.5)$; p-value$< .05$, effect size$= 0.26$); for participants younger than 25yo (p-value$< .1$, effect size$= 0.25$) and for participants with at most secondary education (p-value$< .1$, effect size$= 0.21$).

Scenarios Considering only the 85 participants who interacted with ELVIRA, I investigated whether there was any difference in acceptability of the recommendations across the scenarios (see Figure 7.2a, w.r.t. the scenarios summarised in Table 5.8). Regarding the individual acceptability, in the scenarios 1-3-5 most users either agreed or strongly agreed with the recommendation received, while in the scenarios 2-4-6 the average acceptability was slightly lower, but still mostly in the positive range. A similar but less distinct trend is present also for the collective acceptability. This suggests that the sensitivity of the scenarios influenced the acceptability of the recommendation. A t-test comparing the average acceptability in scenarios with low sensitivity (s1-s3-s5) vs. scenarios with high sensitivity (s2-s4-s6) was significant with $\alpha < .05$ (effect size: individual: 0.60; collective: 0.41).

Satisfaction of the output Regarding the quality of the generated output, ELVIRA achieved by far the best performance. Figure 7.3 shows the distribution of the answers to the Satisfaction Scale (2='Strongly agree', -2='Strongly disagree';



(a) Acceptability per scenario. (b) Acceptability per sensitivity level.

Figure 7.2: Acceptability in the ELVIRA treatment.

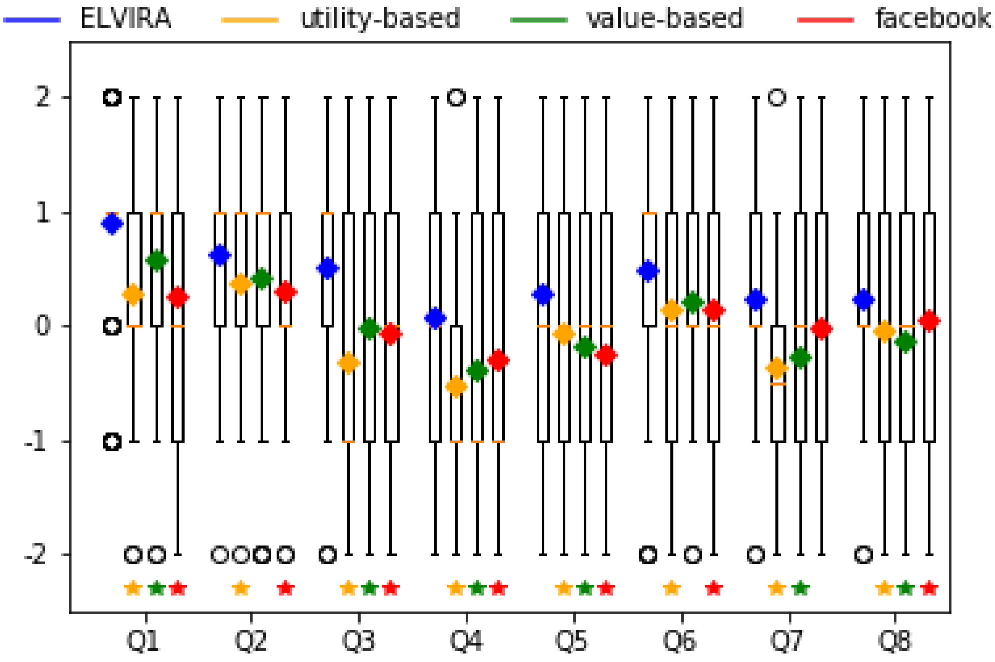


Figure 7.3: Evaluation of the outputs provided by each model, according to the Satisfaction Scale [79].

see all the questions reported in Table 5.9), with significant differences marked as above (p-value< .05, minimum effect size is .31). ELVIRA is the only model presenting a positive average score for each question, and the one with overall the most compact distribution. Particularly, ELVIRA's dominant results can be noted in Q1: "From the output, I could *understand* how the tool works"; Q3: "The output provided *sufficient* detail about how the tool works", Q4: "The explanation provided *complete* information about how the tool works.", Q5: "The explanation tells me how to use the tool.", and Q8: "The explanation let me judge when I should trust and not trust the tool".

7.4 Motivations for accepting a recommendation

When asking the users about the acceptability of each recommendation, I also investigated the motivations that supported their decisions, which were given in a single free-text box for both individual and collective acceptability. Out of the 2010 records I collected, I discarded 65 records where the participants either gave very poor answers due to low effort (e.g., id114: "Intuition", id224: "No", id0: "No motivation", etc.) or provided off-topic comments (e.g., id63: "The more I think about this, the more I wonder how FB hasn't integrated this kind of technology already... you might be on to something here :)").

I analysed the remaining *1945 responses* by applying *thematic analysis* (TA) [180], a well-known and extensively-used method for analysing qualitative data in many disciplines and fields. The purpose of TA is to identify patterns of meaning across a dataset that provide an answer to the research question being addressed. Patterns are identified through a rigorous process of data familiarisation, data coding, and theme development and revision. I followed an inductive and semantic approach to TA [180]: starting from the explicit meaning of the data, I worked bottom-up to develop codes and, ultimately, themes.

Keeping in mind the research question "*Which motivations support the acceptance or rejection of a solution to an MPC?*", I identified the following main themes. Together with the description of the theme, I report some exemplar responses with the identifier of the user (id), their treatment (t) and the scenario where they were given (s):

- **Context**: the nature of the content represented in the pictures, such as depicted people and activities, sensitivity, and sentiment, was the most commonly reported factor when evaluating a recommendation. Users very often considered also the consequences, either positive or negative, that may derive from sharing the picture online. It includes the codes: *context, context-neutral/inappropriate, consequences, consequences-bad/good/lack.* [Id276,t4,s4: "Sharing this photo with more people may lead to complications between the groom and bride". Id314,t2,s1: "Th picture is very professional and will be a nice picture if future employers want to view Felipes social media accounts before hiring him".]

- **Privacy**: the protection of someone's privacy was the second most considered factor. Users often reported concern for the privacy of their own person or of someone else (mainly children or people in a vulnerable position), associating the privacy violation with potentially very negative consequences related to their safety. It includes the codes: *privacy, safety.* [Id30,t4,s5: "Because of the children in the image, I would be keen to keep this photo private, even though it is a good photo technically, for the safety and privacy of the children involved." Id140,t2,s6: "A very personal picture that could be seen by many and used for a number of reason that might not align with me."]

- **Others**: the other people's preferences were frequently playing a role in the decision. When the others' wishes were known, the participants often respected and accommodated them. When that knowledge was not available, users sometimes were wondering what they could be and whether the picture was taken to share with the others' consent. There was often the explicit intention to identify a fair compromise: this was a highly subjective evaluation, which sometimes favoured the option that respected the wishes of the majority, and sometimes the most private preference. It includes the codes: *others, respect, consent, fairness, majority, privacy-most.* [Id21,t1,s2: "It is a very personal photo and the people asleep didnt know that they were being pictured. They did not consent prior to the photo being taken". Id22,t3,s3: "This is fair and respects all parties' privacy"].

- **Indifference**: in many cases, the participants were neutrally interested in the outcome of the MPC and were willing to accept any recommendation or compromise, sometimes just because the solution coincided with what was perceived as a common sharing behaviour. It includes the codes: *neutral attitude, compromise, compromise-accept, common behaviour.* [Id91,t1,s1: "If people do not want to share it with much people then I do not mind". Id103,t2,s5: "Whatever solves the conflict I'll be happy with".]

- **Aesthetics**: the aesthetics of the picture and its impact on the reputation of the users (more on the social network than in real life) were taken into account by many participants. It includes the codes: *flattering, unflattering, entertaining, interest, utility loss.* [Id251,t1,s3: "It would be nice for common friends to see image so they can discuss and comment and leaves comments". Id163,t2,s1: "This was a picture taken of Felipe by someone else and isnt so flattering so would be unlikely to share it further. Others may have a different opinion" Id82,t2,s5: "I don't think that friends of friends really need access to, or benefit from, what was primarily meant for family."]

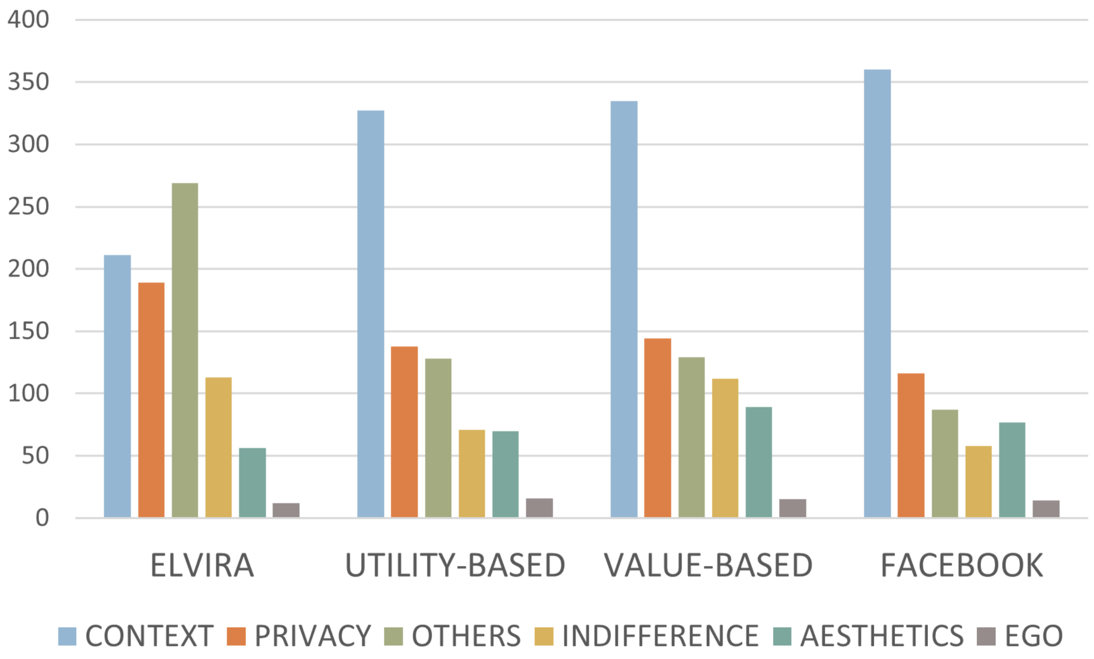


Figure 7.4: Themes distribution across the treatments.

- **Ego**: a number of participants considered the acceptability of the recommended solution just by comparison with their own preference. It includes the code: *ego*. [id258,t4,s2: "the tool has decided the same way i did". id224,t2,s5: "It was my first choice".]

Another reported factor, which is worthy of mention despite its lower frequency, was the possibility of keeping the picture private, in order to satisfy the other users' preferences, and to share more broadly an alternative one, either another picture with the same subject or a modified version of the same one. [Id198,t3,s1: "I would prefer to share this photo publicly [...]. If the other people felt uncomfortable with this then I would either crop them out of the photo or simply take a photo without them in it to post publicly.[...]."] This is a further confirmation of a common strategy considered in real situations which was already reported by previous studies [175].

Being the thematic analysis purely qualitative and exploratory in nature, I do not draw any confirmed conclusions, but I discuss some interesting trends that have

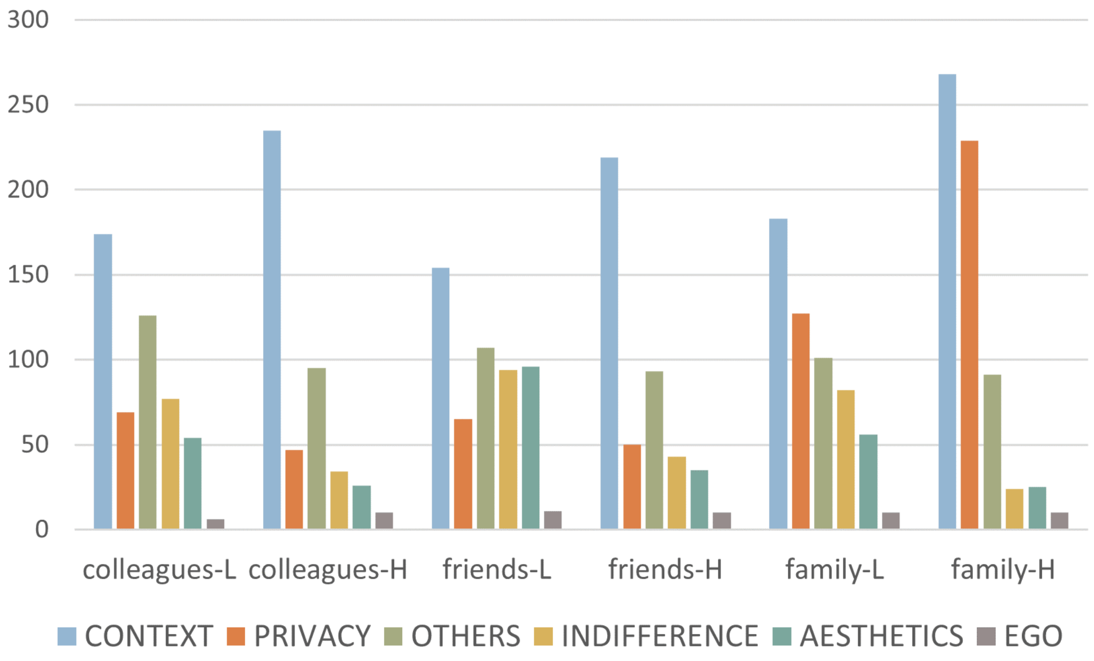


Figure 7.5: Themes distribution across the scenarios.

emerged and may be worthy of future confirmatory studies. Figures 7.4 and 7.5 report a comparative overview of the themes occurrence in the participants' answers[ii] respectively across the treatments and the scenarios. Regarding the treatments, in ELVIRA the participants were influenced most by *Others*, with less impact from *Context* and more from *Privacy* than in the other treatments. This suggests that ELVIRA successfully nudged the participants to be much more conscious of the co-owners and more *Privacy*-aware than the other models, where the participants mostly focused on *Context* when evaluating the recommendation. Regarding the scenarios, *Context* had a stronger influence in the more sensitive scenarios (s2-s4-s6), while *Indifference* and *Aesthetics* were generally more considered in the less sensitive situations (s1-s3-s5); *Privacy* concerns were more related to the familiar sphere (s5-s6) than to other types of relationships.

Figure 7.6 shows the distribution of the codes within each theme, when consider-

[ii] Note that each answer could be labelled with multiple codes and, therefore, be included in multiple themes.

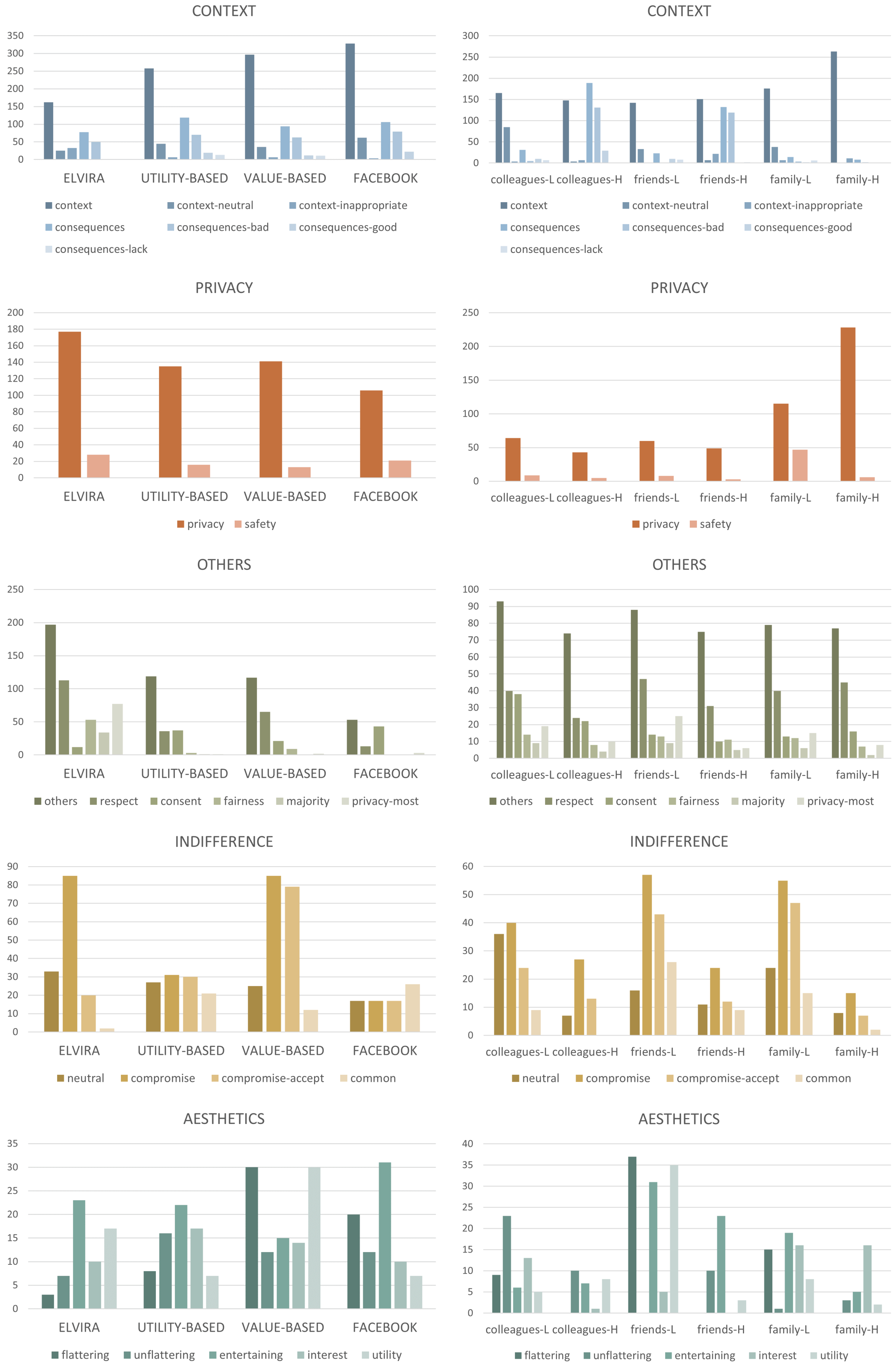


Figure 7.6: Codes distribution within each theme.

ing the treatments (on the left) or the scenarios (on the right). These are in general consistent with what seen in Figures 7.4 and 7.5. However, this more granular analysis highlights interesting new elements, especially regarding the codes within *Others*. To consider the others' preferences had different implications according to the participants: some appreciated solutions coinciding with the preference of the majority; others prioritised the protection of everyone's privacy and opted for the most private solution; some were willing to accept a solution that was not their first choice in order to accommodate the other's wishes; and, finally, some worried about the consequences that sharing the picture could have for the co-owners. The interactions with ELVIRA encouraged the participants to reflect more upon the *fairness* of the recommendation, whether it was a *compromise* (within *Indifference*), and, more generally, to be more *respectful* of the others' wishes. On the other hand, the other treatments made the participants wonder more often about the co-owners' *consent*, about *common* sharing practices (within *Indifference*), and about how *(un)flattering* the picture was (within *Aesthetics*). With ELVIRA, these factors were less relevant, because the participants were told that the received recommendation was already taking into account the others' preferences.

Regarding the scenarios, participants were particularly aware of the *(bad) consequences* (within *Context*) of sharing when considering high sensitive scenarios such as s2 and s4; attention to *safety* (within *Privacy*) was mostly relevant when thinking of children, e.g., in s5; and, finally pictures with friends, such as s3 and s4, were considered the most *entertaining* (within *Aesthetics*).

7.5 Discussion

Considering both the acceptability of the recommendations and the satisfaction with the model's output, ELVIRA outperformed all the other models.

The value-based model provides recommendations that are, generally, as accepted as ELVIRA's, but its outputs are significantly less satisfactory. Even in terms of acceptability, ELVIRA generates solutions that are more acceptable across demographics, while the value-based model seems not to cater for older, more privacy aware and less active social media users, providing recommendations that are significantly less acceptable than ELVIRA's for these groups. Significantly worse than ELVIRA, the utility-based and the Facebook models performed equivalently in terms of acceptability, with Facebook being slightly better in terms of satisfaction of the output.

Moreover, with the less sensitive scenarios, ELVIRA's recommendations were almost always accepted (neutral or positive individual acceptability in 89.8% MPCs), suggesting that the agent may be able to further reduce the user's burden to manage his/her online privacy by autonomously solving the MPCs that emerge in less threatening situations.

Regarding the participants' reasons for accepting or rejecting a recommendation, the users who interacted with ELVIRA showed a much clearer tendency to take into account and respect the co-owners' preferences, than the ones who engaged with the other models.

In conclusion, these results suggest that, in order to promote further the empirically evident collaborative behaviour in MPCs, the recommendations generated by ELVIRA may be beneficial in real-world scenarios for several reasons: (i) they would suggest solutions that are acceptable for users independently of their demographics, their privacy awareness and their OSN experience, especially in low sensitivity contexts; (ii) they would be justified by an overall satisfying explanation; (iii) they would nudge the users towards the appreciation of respectful and fair solutions for all the users involved; and, finally, (iv) they would reduce the discrepancy between very privacy-aware uploaders, who would likely worry more about the others' consent and preferences before sharing, and the less privacy-aware ones, who would more likely cause more often unintentional privacy violations.

7.6 Conclusion

In this chapter, I have described the design and the results of a between-subject user study that I performed in order to evaluate the hypothesis that ELVIRA, an agent-based model which satisfies all the requirements suggested by previous empirical and theoretical studies in privacy and autonomous systems, can offer OSN users a better support for multi-user privacy than other models previously suggested in the literature.

The participants who interacted with ELVIRA accepted significantly more often than for the other treatments the recommendations to solve some simulated MPCs; furthermore, they reported significantly higher satisfaction for the quality of the received output, i.e., the explanation that justified the privacy recommendation.

This evaluation through user study fulfils the scope of the *empirical investigation* within the VSD approach, which aims to understand the user's perception of the developed technology. In this case, the agent ELVIRA has collected broad consensus and positive feedback, for both the content and the quality of the recommendations it provides.

In the next chapter, I will discuss the dynamics that lead OSNs users to adopt and keep using ELVIRA, by presenting the conditions that guarantee its long-term dominance against competing strategies for solving MPCs.

Chapter 8

A simulation of ELVIRA's adoption in OSNs

8.1 Introduction

In Chapter 7 I have described how, in the context of solving multi-user privacy conflicts (MPCs), the participants to the user study found the recommendations generated by ELVIRA more acceptable and more satisfying than the ones generated by other models. In this chapter, I now investigate the dynamics emerging in an online social network (OSN) when ELVIRA can be adopted by users as a technology to manage multi-user privacy (MP).

In particular, I assume an online social network (OSN) to be a free market, where individuals can adopt competing technologies to manage MP. In turn, individuals, if satisfied with their technology, can influence others to adopt it, according to well-known word-of-mouth marketing dynamics [22, 25]. In the following, I show the different market settings and conditions in which ELVIRA imposes itself as the most successful, and therefore adopted, strategy. In order to do this, I draw from evolutionary game theory and study the long-term composition of the market, whose dynamics are influenced by the individual selection of competing technologies. In other words, I study how an OSN user switches strategies for managing MP over time, and how this is influenced by the strategies that are adopted by other users.

After recalling the main concepts of evolutionary game theory in Section 8.2, in Section 8.3 I define and detail the evolutionary game between ELVIRA and other competing strategies. Then, in Section 8.4, I describe and motivate the game configurations that I analyse and I report on their results in Section 8.5. Finally, I conclude this chapter in Section 8.7 with closing comments and remarks.

8.2 Evolutionary Game Theory

Classical game theory regards the study of optimal strategies in competition between adversaries. Players, who are assumed to be rational, self-interested and aware of the rules of the game being played, can select different strategies in order to maximise their own benefit or payoff, while taking into account the expected behaviour of the other players. Of particular interest in a game is the identification of *Nash equilibria*, that is the sets of strategies where no player has any incentive to switch the played strategy, i.e., everyone makes their optimal move given the others' moves. However, the assumption of rationality does not always hold, especially in case of repeated interactions with the same players. For example, when playing the Prisoner's Dilemma, each player is individually best off by defecting (when both players defect, it is a Nash equilibrium), but if the game was repeated an undetermined number of times, then pure defection would not be a strictly dominant strategy anymore [19].

Inspired from biology, where animals (or humans) sometimes act for the benefit of their species more than for their individual one, *evolutionary game theory* studies the dynamics that lead to changing strategies in a population [164]. In this context, players are not required to rationally select a strategy, they can just enter the game with any strategy and evaluate afterwards if their strategy was satisfactory. In the negative case, they can drop it and adopt a new one. The adoption of a new strategy can happen similarly to how genes evolve in the DNA of a species [184]: it can be a random *mutation*, useful to explore new opportunities, or it can be influenced by other more successful individuals that the player wants to learn from, that is traditionally referred to as *imitation.* In evolutionary game theory, we are interested in identifying *evolutionarily stable strategies* [164], which are strategies that, when adopted in a population, are impermeable to the invasion of other strategies. Evolutionarily stable strategies are a refinement of Nash equilibria and are dominant in the long-run, because random mutation is not sufficient to alter the strategy balance in the population.

In the context of this thesis, I consider the users of an OSN as a population (or a market) that can dynamically adopt, through mutation or imitation, competing strategies (or technologies) in order to manage MP; then, I investigate the emergence of evolutionary stable strategies that can provide insights regarding the optimal conditions for ELVIRA's widespread in real-world OSNs.

8.3 An evolutionary competition game

I define an evolutionary game (EG) where a population of N self-interested agents can adopt competing strategies to solve MPCs. Each iteration of the game represents the influence that the resolution of an MPC can have on users adoption of a technology in the social network: by noticing that someone else's conflict was better managed by a different strategy, a player may adopt that other strategy to solve future conflicts. Similarly to the methods implemented for studying public goods games in [163] and [153], in EG the differences between the payoffs obtained by the strategies, seen as the 'happiness' of the agent involved in the conflict about the conflict resolution with that technology, influence the probability of each strategy to be copied through social learning. This, together with mutation dynamics, defines a stochastic process describing the evolution of the frequencies of players (or users) adopting each strategy. By computing the stationary distribution, i.e. the relative frequency in the long-run, of ELVIRA, we can evaluate its evolutionarily stability given different initial settings.

In particular, consistently with the previous evaluation of the ELVIRA agent architecture, in EG there are four competing strategies $S = \{EL, UB, VB, FB\}$: ELVIRA (EL), the utility-based (UB), the value-based (VB) and the Facebook (FB) technology. These employ the same approaches as described in Chapters 5 and 6 to solve an MPC.

Algorithm 1 provides an overview of the evolutionary game. Each iteration t of the game corresponds to an MPC resolution. According to the composition of the population $k_t = [k_{t,EL}, k_{t,UB}, k_{t,VB}, k_{t,FB}]$ at the time t, i.e. how many agents play each strategy i, a payoff $P_{t,i}$ is computed and dynamics of *mutation* and *imitation* may take place.

Algorithm 1: The Evolutionary Competition Game EG

```
input  : $k_0, M, s, \mu, T$
output: $k_T$
for $t = 0$ to $T$ do
    if random(0,1) $< \mu$ then
        $k_{t+1} \leftarrow$ mutate($k_t$);
    else
        $P_t \leftarrow$ computePayoffs($k_t$);
        $k_{t+1} \leftarrow$ imitate($k_t, P_t, s$);
return $k_T$
```

Payoffs For each strategy $i \in S$, according to the proportion $k_{t,i}$ of agents that adopts it at time t, I define the payoff $P_{t,i}$ as follows:

$$P_{t,i} = (acc_{coll,i} + acc_{ind,i} + sat_i - \gamma_i c)\,\frac{k_{t,i}}{N}. \tag{8.1}$$

The collective acceptability acc_{coll}, the individual acceptability $\mathrm{acc}_{\mathrm{ind}}$ and the satisfaction with the output sat contribute positively to the payoff, while the cost c, eventually discounted by γ, contributes negatively. Table 8.1 reports for each strategy the payoff parameters used in the experiments. The values of sat, acc_{coll} and

strategy	**sat**	$\mathbf{acc_{coll}}$	$\mathbf{acc_{ind}}$
EL	0.62	0.708	0.418
UB	0.264	0.461	-0.064
VB	0.496	0.627	0.026
FB	0.245	0.363	0.02

Table 8.1: Payoff parameters obtained from user study in Chapter 7.

acc_{ind} are informed by the user study reported in Chapter 7: they represent respectively the average of the satisfaction score (across the 8 questions of the Satisfaction Scale), of the collective acceptability and of the individual acceptability, that each approach to solve MPCs obtained in the user study. In the context of EG, the cost of a technology may be interpreted and instantiated in several ways, according to the focus(es) of interest: e.g., from a user experience perspective, it could quantify the easiness of the user-technology interaction during the MPC resolution; from a software development perspective, it could represent the inter-operability, i.e., the compatibility of a technology in generating a MPC solution when interacting with different technologies; from an economical perspective, it could represent the price the user needs to pay, in terms of money, personal data, etc., in order to use the technology to solve MPCs; and so on. Given the lack of real data that can inform the selection of appropriate costs, in the first tranche of experiments both the costs and the discount factors will be unitarian for all the strategies, in order to evaluate the influence of the other elements of the model on the emerging dynamics. Later, I will vary the discount factors for the costs of the technologies competing against ELVIRA, in order to identify the *relative* maximum cost of ELVIRA, in comparison with the other strategies, that would allow ELVIRA to be evolutionary stable.

Mutation In each iteration t of the evolutionary game EG, an agent is randomly selected to randomly change its strategy, i.e. to adopt a new technology to manage MP, according to the *mutation rate* μ. This parameter influences the component of noise in the evolutionary game.

Imitation In each iteration t of the evolutionary game EG when mutation does not occur, *imitation dynamics*, also referred to as *social learning*, happen. In this case, two agents are randomly selected and one adopts the strategy of the other one according to the imitation strength parameter s and the payoffs of the two agents' strategies. In particular, given two sampled agents a and b and the difference between their payoffs $\delta = P_{t,a} - P_{t,b}$, if a number randomly selected between 0 and 1 is lower than $f(s, \delta)$, then a adopts b's strategy, otherwise vice-versa. According to the tradition in evolutionary games (e.g., see [163, 153]), I define $f(s, \delta)$ as follows:

$$f(s, \delta) = \frac{1}{1 + e^{-s\delta}}.$$

If s is relatively high, then the learning capabilities of the agents are stronger and it is more likely that the agent that receives the lowest payoff adopts the strategy of the better off agent; if s is low, then the learning is weaker and an agent may adopt a worse strategy.

8.4 Experimental settings

In order to explore different situations in which ELVIRA may result to be the evolutionarily stable strategy, i.e., the dominant technology in the long-run, I evaluate the evolutionary game EG with different settings, which I describe below.

EG1 - Uniform population composition In this setup, I assume the four strategies to be initially equally represented in the population, i.e., $k_{0,i} = N/4$ for each strategy i. This is helpful to study the effect on the strategy adoption when considering different values for (i) the size of the population N, (ii) the number of iterations played T, (iii) the mutation rate μ and (iv) the imitation rate s, while maintaining the same discount factor $\gamma = 1$ for the cost of all the strategies.

EG2 - Varying population composition In this setup, I consider fixed values for the size of the population $N = 1000$, the number of iterations $T = 10^4$, the mutation rate $\mu = 0.001$, the imitation rate $s = 10$ and the discount factor $\gamma = 1$, while varying the percentage of population that initially adopts each strategy k_0. This is helpful to identify the types of population composition that guarantee the evolutionary stability of ELVIRA, i.e., for understanding the ability of the ELVIRA technology to be adopted by the majority of users in as OSN even when starting from disadvantaged distributions.

EG3 - Varying discount factors In this setup, I first define sub-games where ELVIRA's competition is represented by a single technology at time and the OSN is initially uniformly distributed: $k_{0,EL} = N/2$, and $k_{0,i} = N/2$ for one $i \in \{UB, VB, FB\}$ at time. The discount factor for the cost of ELVIRA remains constant, $\gamma_{EL} = 1$, while I study the effect of decreasing the ones of each other strategy, making them "cheaper" to adopt than ELVIRA. Then, I consider again the main game EG and I look at the effect that discounting the cost of the other strategies when they all compete against each other has on the adoption of ELVIRA. This is helpful to understand how much more expensive than each other strategy can ELVIRA afford to be in order to still guarantee its wide adoption in the OSN (i) in each direct competition (one-to-one) game and (ii) in the main game with an initially uniformly distributed population.

I implemented the evolutionary competition game EG in Python 3.6 and I ran all the simulations on Windows 10 64-bit, Intel(R) Core(TM) i7-7Y75 CPU @ 1.30GHz 1.60 GHz 8GB.

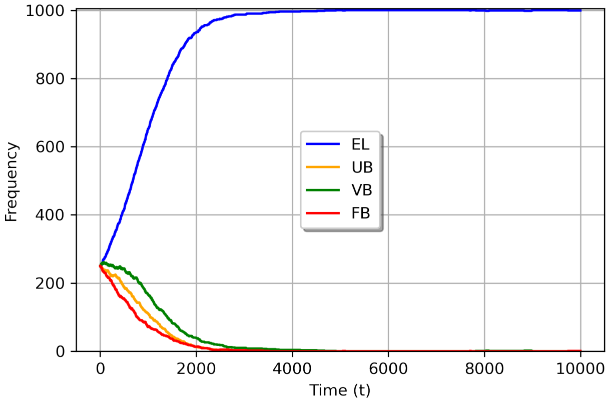


Figure 8.1: Evolution of a uniformly distributed initial population: $N = 1000$ and $k_{0,i} = N/4$ for each strategy $i \in \{EL, UB, VB, FB\}$.

8.5 Experimental results

In this section I report on the results of the simulated evolutionary game EG with the settings previously introduced.

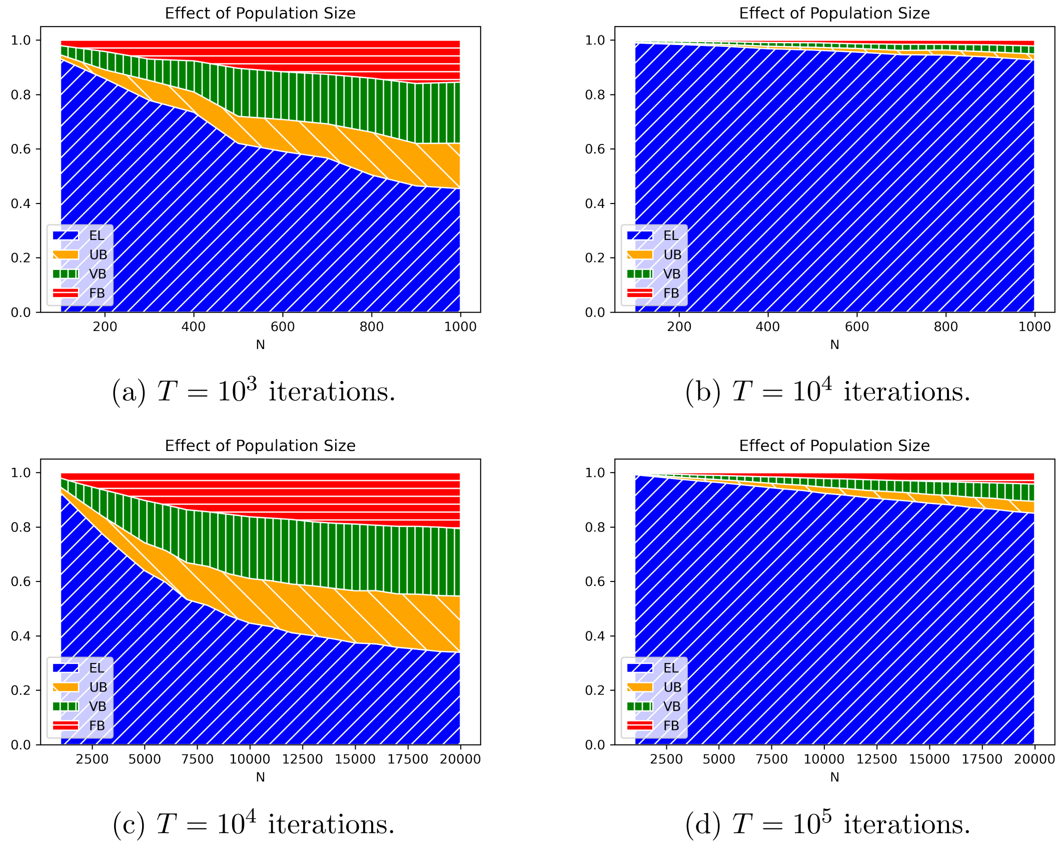
8.5.1 EG1 - Uniform population composition

First, I explored the variation in population composition during $T = 10^4$ iterations of the competition game between the strategies ELVIRA, utility-based, value-based and Facebook when fixing the size of the population $N = 1000$, the mutation rate $\mu = 0.001$, the imitation rate $s = 10$ and the cost $c = 1$. Figure 8.1 shows that, when starting from a uniformly distributed population, the agents quickly drop the other strategies and adopt ELVIRA in an evolutionarily stable manner, i.e. ELVIRA invades the OSN and is the dominant technology in the long run. The stability of this emerging dynamics is confirmed by the average and the standard deviation, computed over 100 runs of EG1, of the long-run average frequency of each strategy (see Table 8.2).

Second, I explored the population composition when varying the population size N from 100 to 20k agents. Figure 8.2 shows that ELVIRA is the evolutionarily stable

strategy	mean	st. dev.
EL	0.9249	0.0027
UB	0.0220	0.0013
VB	0.0316	0.0023
FB	0.0215	0.0012

Table 8.2: Long-run average frequency of a uniformly distributed initial population.



(a) $T = 10^3$ iterations.

(b) $T = 10^4$ iterations.

(c) $T = 10^4$ iterations.

(d) $T = 10^5$ iterations.

Figure 8.2: Population distribution after T iterations, when varying the population size N.

strategy in any size of population, as long as enough game iterations are allowed for. Hence, the bigger the population, the longer it takes for the word-of-mouth marketing to impose ELVIRA as the dominant technology in the OSN, which is anyways the guaranteed outcome.

Next, I explored the impact of the mutation rate μ on the population composition (see Figure 8.3), when maintaining $N = 1000$, $T = 10^4$ and $s = 10$. Intuitively, the

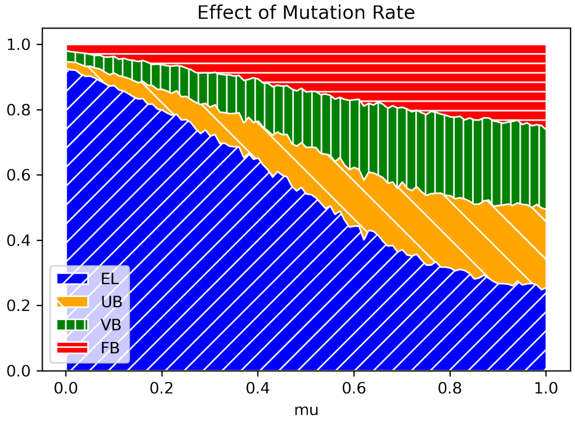
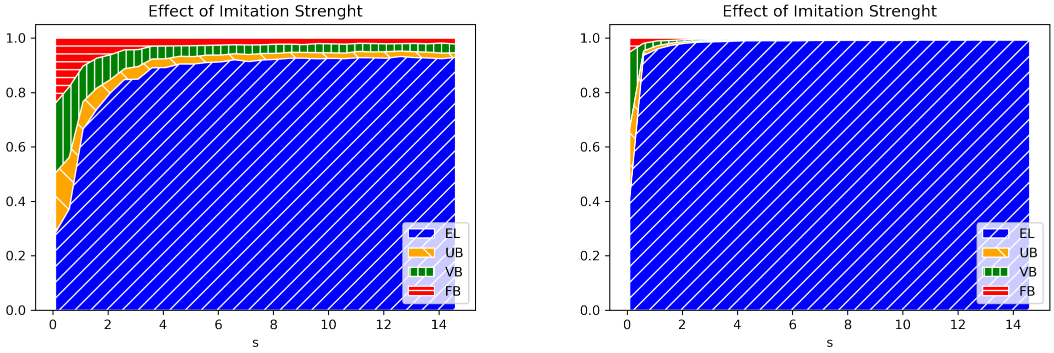


Figure 8.3: Population distribution after $T = 10^4$ iterations, when varying the mutation rate μ.



(a) $T = 10^4$ iterations.

(b) $T = 10^5$ iterations.

Figure 8.4: Population distribution after T iterations, when varying the imitation strength s.

larger the mutation rate, the more random is the behaviour of the population: in particular, when $\mu = 1$, the strategy distribution remains unchanged after $T = 10^4$ iterations, with a quarter of agents adopting each strategy. On the other hand, for smaller values of μ, i.e., for less noisy configurations of game, ELVIRA tends to be the most preferred strategy by the agents.

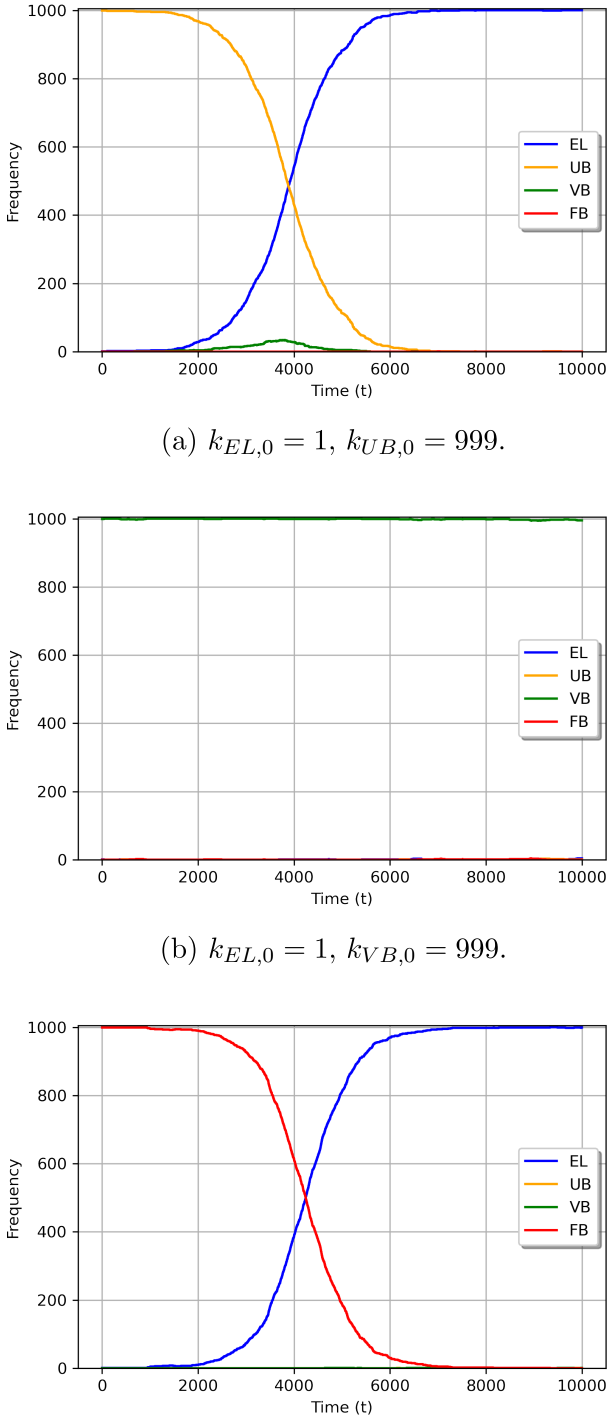
Finally, I explored the impact of the imitation strength s on the population composition (see Figure 8.4), when maintaining $N = 1000$ and $\mu = 0.001$. Relatively small values of s already resemble "strong" imitation ($s \to +\infty$, see [163]), where more successful agents (such as the ones playing the ELVIRA strategy) are always imitated, and less successful ones never. Strong imitation guarantees the efficacy of the word-of-mouth marketing for imposing ELVIRA on the OSN.

8.5.2 EG2 - Varying population composition

Here I explore the ability of the strategy ELVIRA to invade the population, i.e., to be adopted in a stable manner by the majority of the agents, when varying the initial composition of the population. For these simulations, I maintained $T = 10^4$, $N = 1000$, $\mu = 0.001$, $s = 10$ and ran 100 evolutionary games for each configuration.

Figures 8.5a, 8.5b and 8.5c show the time series of the strategy frequencies during one evolutionary game where at $t = 0$ there is only one agent playing ELVIRA and all the others playing utility-based, value-based and Facebook, respectively. Note that in each scenario there are two strategies that are initially not represented in the population, even though they may be adopted by mutation. The utility-based and Facebook strategies seem to be easily dropped in favour of ELVIRA, but a homogeneous population of value-based players does not get influenced by a single ELVIRA player.

Then, I look at the long-run frequency of ELVIRA, when consecutively increasing the initial number of agents adopting ELVIRA $k_{EL,0}$ from 0 to 400 with steps of 5. Figures 8.6, 8.7 and 8.8 show its average and standard deviation computed over 100 runs of each EG2. Again, utility-based and Facebook react similarly to the invasion of the ELVIRA strategy. In both cases, the more ELVIRA agents are in the population at the beginning, the faster ELVIRA becomes, and remains, evolutionarily stable. On the other hand, when engaging with a population of VB players, there need to be at least about 150 ELVIRA players in order for ELVIRA to have a chance of invading the population: with a number of ELVIRA players between 150 and 200, the system is very volatile and ELVIRA may or may not

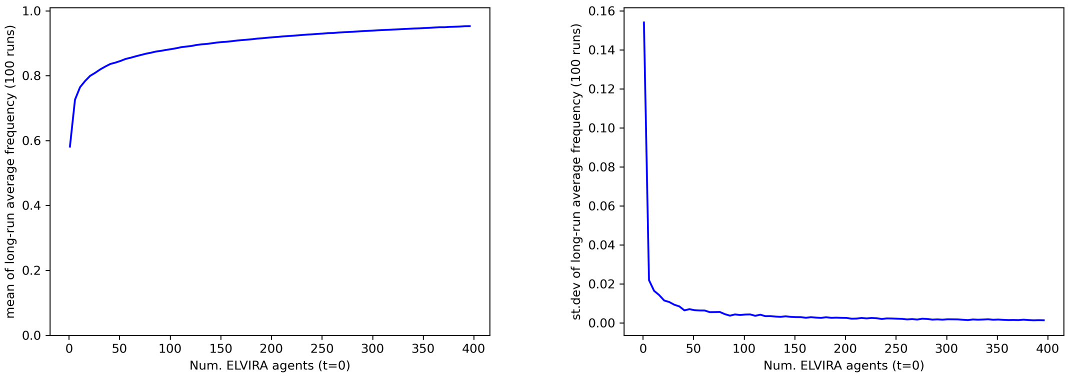


(a) $k_{EL,0} = 1$, $k_{UB,0} = 999$.

(b) $k_{EL,0} = 1$, $k_{VB,0} = 999$.

(c) $k_{EL,0} = 1$, $k_{FB,0} = 999$.

Figure 8.5: Evolution of a population with different initial distribution k_0.



(a) Avg. of long-run average frequency.

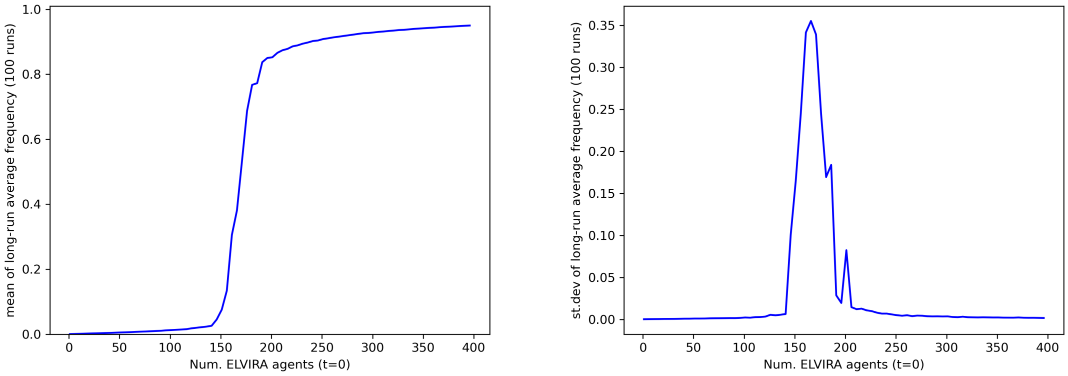
(b) Std. dev. of long-run average frequency.

Figure 8.6: EL vs UB: Long-run average frequency of ELVIRA agents when varying the initial population composition: $k_{EL,0} \in [0, 400]$, $k_{UB,0} = N - k_{EL,0}$.

invade (see Figure 8.9). Only with more than 205 initial agents, the evolutionary dominance of ELVIRA is guaranteed. This means that, while an OSN saturated by the UB or FB technologies would be easy to invade for ELVIRA, the conquest of an OSN dominated by VB would require a more substantial effort in terms of initial resources: for example, more users should be initially paid to try the new ELVIRA technology, before the word-of-mouth marketing could be effective. Further research could investigate whether the number of initial ELVIRA players could be reduced if these were selected strategically in the network, i.e. well-connected users or "influencers", and the social connections were influencing the imitation dynamics, which is not the case in the current EG model.

8.5.3 EG3 - Varying discount factors

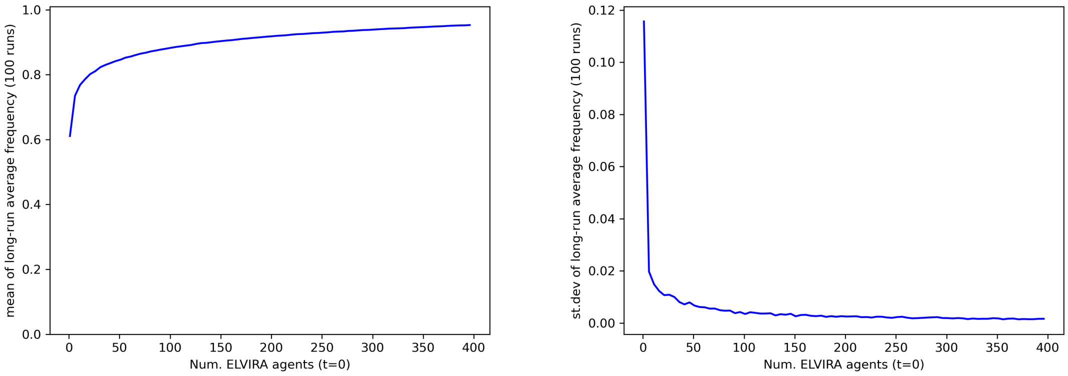
Here I explore the ability of the technology ELVIRA to invade the OSN, i.e., to be adopted in a stable manner by the majority of the users, when discounting the cost of adopting the other competing technologies. For these simulations, I maintained $T = 10^4$, $N = 1000$, $\mu = 0.001$, $s = 10$ and ran 100 evolutionary games for each configuration.



(a) Avg. of long-run average frequency. (b) Std. dev. of long-run average frequency.

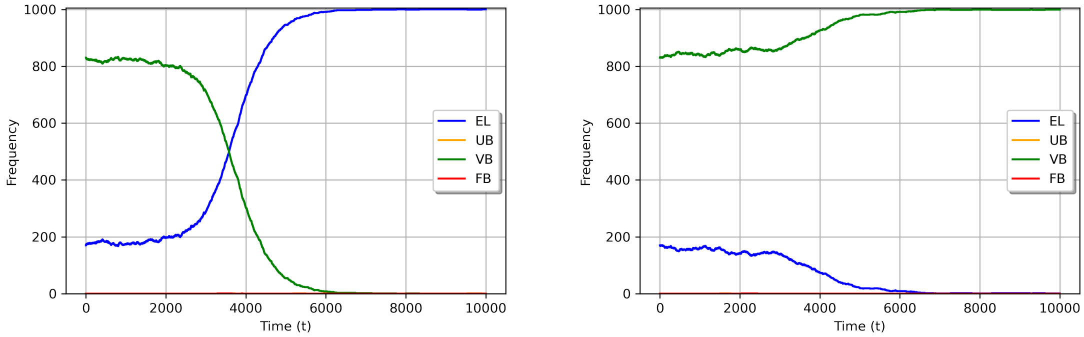
Figure 8.7: EL vs VB: Long-run average frequency of ELVIRA agents when varying the initial population composition: $k_{EL,0} \in [0, 400]$, $k_{VB,0} = N - k_{EL,0}$.

First, in order to study the effects of a direct competition between ELVIRA and each other strategy, I define three sub-games where only two strategies are involved. At the beginning of each sub-game, the population is uniformly distributed: in all the three cases $k_{EL,0} = N/2$ and (EG3a) $k_{UB,0} = N/2$, (EG3b) $k_{VB,0} = N/2$ and (EG3c) $k_{UB,0} = N/2$. Note that, differently from EG2, each sub-game involves only two strategies: the other two cannot be adopted even by mutation. These more limiting assumptions about the model allow the derivation of stronger and more reliable results, because less influenced by randomness, which could better inform the future deployment of the ELVIRA technology in real-world OSN. Figures 8.10, 8.11 and 8.12 show the average and the standard deviation of the long-run average frequency of agents adopting ELVIRA (computed over 100 runs of each evolutionary sub-game), when increasing the discount factor γ (steps by 0.05) for the cost of utility-based, value-based and Facebook, respectively. In EG3a and EG3c, even by discounting the cost of the utility-based and the Facebook strategies at the point of making them "free", ELVIRA still gets quickly adopted by the entire population. However, in EG3b there emerges a more interesting dynamics: if the adoption of value-based is expensive at most 30% of the cost of adopting ELVIRA, then VB



(a) Avg. of long-run average frequency. (b) Std. dev. of long-run average frequency.

Figure 8.8: EL vs FB: Long-run average frequency of ELVIRA agents when varying the initial population composition: $k_{EL,0} \in [0, 400]$, $k_{FB,0} = N - k_{EL,0}$.

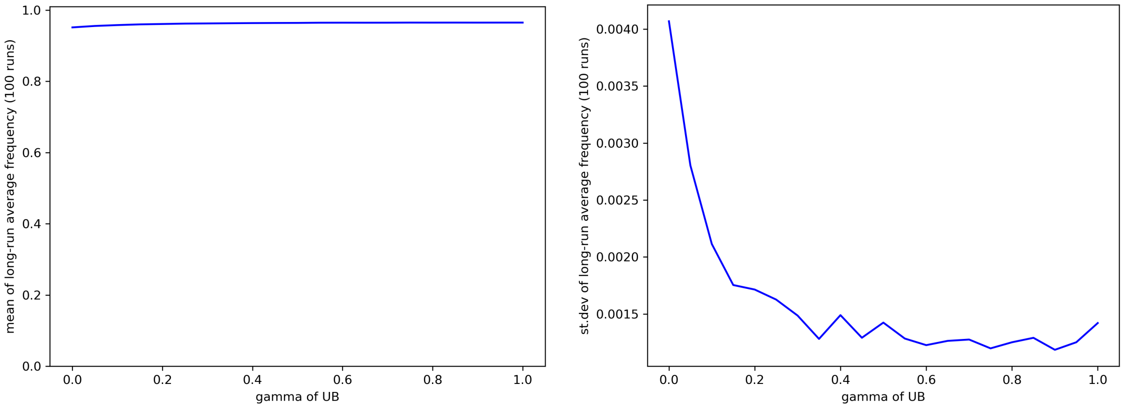


(a) Dominance of ELVIRA strategy. (b) Dominance of VALUE-BASED strategy.

Figure 8.9: Alternative evolutions of a population with the same initial distribution $k_{EL,0} = 170$, $k_{VB,0} = 830$.

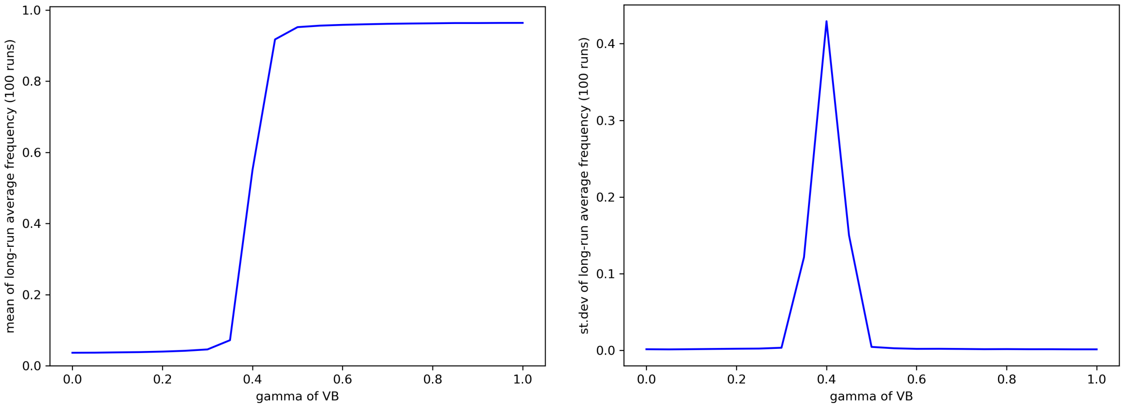
invades the population; if the cost of VB is equal or more than 50% of the cost of EL, then EL invades the population; finally, for costs of VB between 30% and 50% of the cost of EL, the outcome of the competition among the two strategies is more uncertain, as the pick in standard deviation of the long-run average frequency of ELVIRA agents shows in Figure 8.11b.

Then, I considered the main game EG3, where the four strategies compete with each other starting from a uniformly distributed population. Without discounting the cost of ELVIRA, $\gamma_{EL} c_{EL} = 1$, I explored all the combinations of discount factors



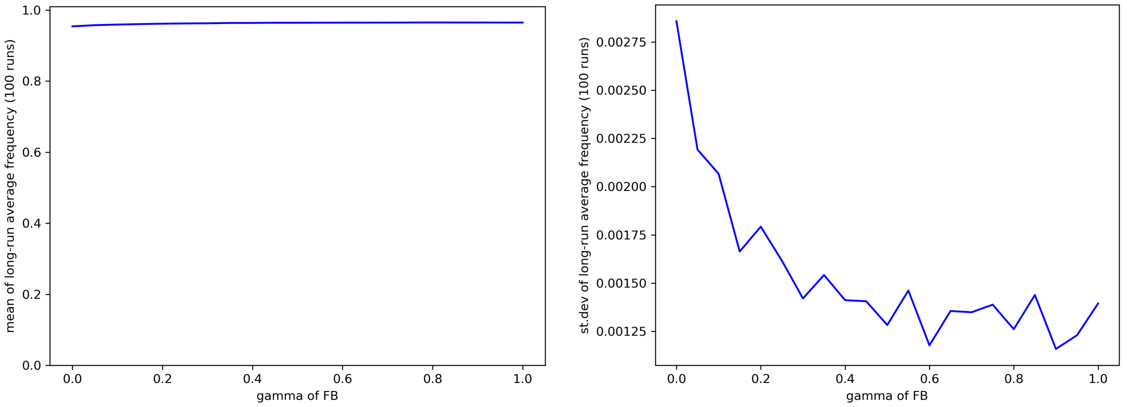
(a) Avg. of long-run average frequency. (b) Std. dev. of long-run average frequency.

Figure 8.10: ELVIRA vs UTILITY-BASED: Long-run average frequency of ELVIRA agents when varying the cost of UT: $c_{EL} = 1$, $c_{UB} = \gamma c_{EL}$, where $\gamma \in [0, 1]$. Average and standard deviation computed over 100 runs.



(a) Avg. of long-run average frequency. (b) Std. dev. of long-run average frequency.

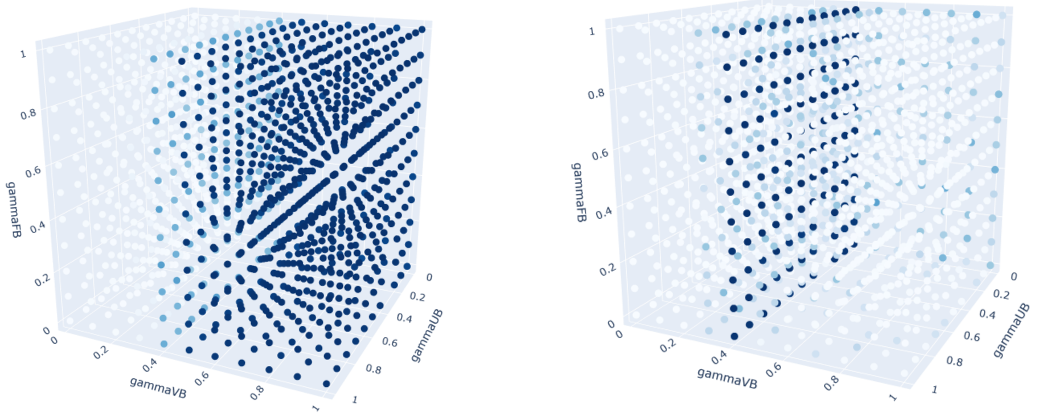
Figure 8.11: ELVIRA vs VALUE-BASED: Long-run average frequency of ELVIRA agents when varying the cost of VB: $c_{EL} = 1$, $c_{VB} = \gamma c_{EL}$, where $\gamma \in [0, 1]$. Average and standard deviation computed over 100 runs.



(a) Avg. of long-run average frequency. (b) Std. dev. of long-run average frequency.

Figure 8.12: ELVIRA vs FACEBOOK: Long-run average frequency of ELVIRA agents when varying the cost of FB: $c_{EL} = 1$, $c_{FB} = \gamma c_{EL}$, where $\gamma \in [0, 1]$. Average and standard deviation computed over 100 runs.

γ_{UB}, γ_{VB} and γ_{FB} (steps by 0.1). Figure 8.13 shows the average and the standard deviation of the long-run frequency of ELVIRA agents in the population with these settings, where each axis reports the variation of one discount factor. It is evident that different values of γ_{UB} and γ_{FB} did not have any visible impact on the long-run frequency of ELVIRA, that is instead clearly influenced by γ_{VB}. In particular, values of γ_{VB} lower than 0.4 led to a close-to-zero long-run frequency of ELVIRA (white dots in Figure 8.13a), while values higher than 0.4 resulted in ELVIRA being the evolutionary stable strategy (dark dots in Figure 8.13a). Coherently, $\gamma_{VB} = 0.4$ generated the noisiest outcomes, as shown by the peak of standard deviation in Figure 8.13b. For the interested reader, since the figure does not allow to infer the details, I report in Table 8.3 the configurations of the discounts γ_{UB}, γ_{VB} and γ_{FB} that generated values of interest (minimum and maximum average and standard deviation) for the long-run average frequencies of the agents playing ELVIRA and value-based strategies. Both strategies were the least stable, i.e., their long-run average frequencies show the highest standard deviation, in the same configuration, where UB and FB were little or not discounted and the cost of VB was 40% of



(a) Avg. of long-run average frequency. (b) Std. dev. of long-run average frequency.

Figure 8.13: Long-run average frequency of ELVIRA in an initially uniformly distributed population ($k = N/4$ for each strategy) when varying the cost of the other strategies. The darker the dot, the higher the value represented; note, however, that each colour corresponds to different value ranges in the two plots – see Table 8.3 for details.

ELVIRA's one. ELVIRA shows the best performance, i.e., the highest average of long-run average frequencies, in the same scenario as in CG1, where all the strategies have the same cost. The value-based strategy shows the strongest dominance when $\gamma_{VB} = 0$, intuitively, and UB and FB are little or not discounted ($\gamma_{UB} = 0.7, \gamma_{FB} = 1$).

In conclusion, the dynamics that emerged in the sub-games are confirmed also in the main game: whether the ELVIRA technology manages to be widely adopted in the OSN depends mostly on its relative price compared with the value-based technology, while the other two technologies seem to have a negligible influence on ELVIRA's chances of success.

8.6 Discussion

In order to understand the most successful conditions that allow the ELVIRA technology to be adopted by the majority of the users when deployed in an OSN, I have simulated a number of evolutionary games where ELVIRA competes against

			γ_{UB}	γ_{VB}	γ_{FB}	avg freq (avg)	avg freq (std)
EL	min	avg	0.0	0.0	0.0	**0.028431**	0.002625
		std	0.9	1.0	0.9	0.924263	**0.002287**
	max	avg	1.0	1.0	1.0	**0.92523**	0.002609
		std	1.0	0.4	0.9	0.473429	**0.414981**
VB	min	avg	0.0	1.0	0.0	**0.02044**	0.001505
		std	0.1	1.0	0.0	0.020896	**0.001333**
	max	avg	0.7	0.0	1.0	**0.92567**	0.003334
		std	1.0	0.4	0.9	0.488075	**0.415105**

Table 8.3: Details of the discount configurations that generated minimum and maximum values for average and standard deviation of long-run average frequency for the strategies ELVIRA and value-based.

other technologies and the imitation dynamics reproduce word-of-mouth marketing approaches.

In EG1, where the population was initially uniformly distributed and all the strategies had the same cost, independently of the size of the population, with an imitation strength $s \geq 1$ and a mutation rate $\mu \leq 0.2$, ELVIRA was always the evolutionary stable strategy. This suggests that ELVIRA has a stronger capability to spread than the other technologies: i.e., if all the technologies were relying on word-of-mouth marketing, ELVIRA would be the most successful and would be widely adopted in the OSN.

In EG2, where the strategies still had the same cost, but the simulations started with strongly unbalanced initial distributions of agents, one ELVIRA agent alone was sufficient to quickly and stably invade the populations composed by only utility-based or Facebook users, but about a quarter of the population size was necessary to win the competition against a majority of value-based users. This suggests that the nature of the competing technologies available in the OSN directly influences the amount of resources to be invested in the deployment of ELVIRA. If the OSN is saturated by either the utility-based or the Facebook technologies, then ELVIRA can invade it with minimal effort, e.g. by involving in a trial a small number of users. On the other hand, if most of the users in the OSN have already adopted the value-based technology, then a more significant effort must be made in order to guarantee the widespread of ELVIRA.

In EG3, the exploration of the discount factors for the cost of each competing technologies showed that ELVIRA dominated the utility-based and the Facebook technologies even when the latter were free, and the value-based technology when its cost was no cheaper than 50% of the cost of ELVIRA, independently of the initial population distribution. This means that a higher cost of ELVIRA, compared to the other technologies, does not hinder its widespread in the OSN. For instance, despite its cost being higher due to the more sophisticated usability of the technology (being fully explainable) or the demands of more personal data (individual utilities and values), the dominance of ELVIRA on the OSN is guaranteed.

Considering these results, I conclude that, when assuming competing technologies to manage MP in OSNs such as the utility-based, the value-based and the Facebook technologies, ELVIRA would generally be widely adopted and preferred to the alternatives in the long term.

8.7 Conclusion

In this chapter I have finally concluded the empirical and technical investigation of the ELVIRA model within the Value Sensitive Design approach.

In particular, I have studied the minimal conditions for the widespread adoption of ELVIRA as a technology to manage multi-user privacy in OSNs. In order to do this, I have designed an evolutionary game where ELVIRA competes against other technologies (utility-based, value-based and Facebook – consistently with the rest of this thesis) to be adopted by the users of an OSN. At each iteration of the game, a user can change their strategy, i.e. adopt another technology, either by randomly mutating or by imitating another user. The imitation dynamics, also referred to as *social learning* [163, 153], are particularly significant in the OSN as a free market setting, because they can represent the effects of a word-of-mouth advertisement of a technology: a user who is happy with the adopted technology would recommend its adoption to other users.

Experimental results have shown that ELVIRA is the evolutionary stable strategy, that is it seems that ELVIRA could be adopted by the majority of users, in a variety of settings. Similarly to what noticed in the user study described in Chapter 7, the value-based approach to manage MP is the strongest competitor of ELVIRA. In fact, the influence of utility-based and Facebook on the overall technology-adoption dynamics are negligible, independently of the specific technologies initially adopted by the users and of their costs. On the other hand, in order to dominate the value-based technology, and therefore be adopted by most of the OSN, ELVIRA can still be successful if more resources are invested in its deployment and its adoption cost is no higher than twice the cost of adopting the value-based technology.

In conclusion, by considering the users' appreciation for the ELVIRA's solutions, as reported in Chapter 7, I have shown here that, under certain conditions and assumptions, ELVIRA would be the preferred technology to manage multi-user privacy in OSNs in the long term.

Chapter 9

Conclusion

9.1 Summary

The main focus of this thesis revolves around the answer to the research question stated in Chapter 1: *RQ: How do we design autonomous systems (ASs) that can effectively help users of online social networks (OSNs) manage multi-user privacy?*. In order to be effectively helpful and safely deployed in society, ASs need to be *value-aligned*, i.e., act coherently with their users moral values, and *explainable*, i.e., able to provide justification for their actions. In addition to these, informed by the literature in Privacy and Social Sciences, I have identified other requirements that ASs should satisfy in the context of solving multi-user privacy conflicts (MPCs), that is to be adaptive, role-agnostic and utility-driven.

After introducing the topic of this thesis, in Chapter 2 I have summarised the interdisciplinary previous work on which I have built my research. First, I have described different theories of *privacy* and of *values* and outlined the *Value Sensitive Design* (VSD) approach, a design methodology to embed values in technology which I have followed throughout this thesis. Then, I have illustrated the *practical reasoning* process, as defined by Atkinson and Bench-Capon [14], as a way to identify the optimal behaviour of an autonomous agent in given contexts; this process is particularly useful in order to *explain* autonomous behaviour, explanation that needs to be both cognitive and *social*, as remarked by Miller [112].

Then, in Chapter 3 I have given an overview of the ASs, with a particular focus on the agent-based models, that have been previously presented by other scholars in order to support users manage multi-user privacy. When considering the requirements described earlier, none of the works available in the related literature can be considered successful. Keeping in mind the research sub-questions related to the safe deployment of ASs, I have also summarised the main attempts of producing value-aligned and explainable autonomous agents in the AI literature.

Chapter 4 is dedicated to the introduction of JIMMY, an agent-based model that, similarly to the Jiminy Cricket character in Pinocchio, helps users manage multi-user privacy while supporting a morally aligned interaction between users. In particular, in order to identify a commonly acceptable sharing policy, JIMMY drives its user through a negotiation process with the other individuals involved in the MPC, where each negotiating action (e.g., make/accept/reject an offer) is identified according to the user's preference over some moral values, modelled according to the Schwartz Theory of Basic Values. Despite being *fair* and *value-aligned* by design, and proving to be sound and terminating in a finite time, JIMMY presents some drawbacks, such as a limited support for explainability and potential user dissatisfaction due to only value-driven solutions.

By following the iterative process of VSD, I have revised the JIMMY model and, in Chapter 5, I have defined ELVIRA, an agent-based model that satisfies all the identified requirements for successfully supporting the resolution of MPCs. After aggregating the individuals' appreciation for any suggested solution in terms of value promotion and utility generation, the ELVIRA agent performs practical reasoning in order to identify the optimal solution, that is the solution most coherent with all the involved users' preferences. This makes the agent adaptive, role-agnostic, and both value- and utility-driven. Explainability is also provided as the combination of a cognitive process, represented by the practical reasoning, and a social process, which defines how to best convey to the user the necessary information that justifies the selection of the optimal solution. I have accurately studied the design of the explanations generated by ELVIRA through a user study, which identified the hybrid (contextually general or contrastive) tailored explanation to be the most satisfactory explanation structure for users.

In Chapter 6 I have presented and discussed the results of the ELVIRA evaluation through software simulations, consistent with the technical investigation within VSD. In particular, I have compared the performance of ELVIRA and other three models, namely utility-based (UB), value-based (VB) and Facebook (FB), which are representative of the solutions previously proposed by scholars or available in real-world OSNs. The performance of each model was measured in terms of generation of utility, i.e. content availability to approved or disapproved audience, and promotion of values, i.e. coherency of the solution with the individual's moral value preferences. Across all the experiments, that I have performed on both synthetic and real networks, UB achieved the highest levels of generated-utility but with poor value promotion; VB was the most coherent with the users' values but generated poor utility; FB had the worse performance for both utility and values; and ELVIRA consistently showed the best utility-values trade-off, by producing utilities very close to UB and promoting values almost as much as VB.

According to the next step of VSD, the empirical investigation, I have described in Chapter 7 the design and the results of a between-subjects user study where the MPCs solutions recommended by ELVIRA, UB, VB and FB were compared in terms of degrees of acceptability and satisfaction for the users across different scenarios. Even though VB generated solutions that were generally as acceptable as ELVIRA's, ELVIRA outperformed all the other models when considering the users' satisfaction of their outputs, while securing the highest acceptability independently of the users' demographics and privacy awareness. Furthermore, by performing a thematic analysis on the users' reasons for accepting or rejecting the recommendations, I have noticed that ELVIRA nudged, more than the other models, the users to consider and respect the preferences of their co-owners, supporting a more pro-social behaviour.

Finally, after showing the benefits and the users' predilection for ELVIRA's MPCs solutions, respectively through software simulations and a user study, in Chapter 8 I have studied the conditions for the widespread of ELVIRA as a technology to manage multi-user privacy in an OSN. Inspired by evolutionary game theory, where individuals can randomly mutate their strategy or imitate the strategy of other more successful players, I argue that word-of-mouth advertisement, where users satisfied with a technology recommend its adoption to others, would be the best approach to guarantee the large-scale adoption of ELVIRA over time. Therefore, I have modelled the competition between the technologies defined by ELVIRA, UB, VB and FB as an evolutionary game. In OSNs where the initial strategy distribution is uniform, i.e. the same number of users have adopted each strategy, ELVIRA quickly emerges as the evolutionary stable strategy, that is the technology that is constantly adopted by the majority of the OSN. This emerging dynamics is confirmed also when the OSN is initially dominated by either UB or FB. The competition is harder for ELVIRA when the majority of users have adopted VB: in this case, the success of ELVIRA depends on a more significant deployment effort. In any case, the higher cost of the ELVIRA technology, in terms of production and/or adoption, would not hinder its widespread adoption in the OSN and ELVIRA would be the preferred technology to manage MP in OSNs in the long term.

After having gathered positive conclusions to both the technical and the empirical investigation, I considered complete the VSD approach that has driven the design and evaluation of ELVIRA, which represents my answer to the research question

RQ.

9.2 Main Contributions

In this thesis, by answering the research questions presented in the Introduction (see Section 1.2), I have contributed to the fields of privacy-preserving autonomous agents, value-aligned autonomous agents and explainable autonomous agents.

Privacy With respect to privacy, and more specifically multi-user privacy, I have identified the crucial features that enable autonomous systems to *effectively* help users of online social networks manage multi-user privacy (cf. **RQ-A**). Specifically, whenever multi-user privacy conflicts occur, an autonomous agent should provide and justify (i.e., be explainable) solutions that are impartial to the role of the user (i.e., be role-agnostic), that depend on the contextual preferences of all the involved users (i.e., be adaptive), that preserve the user's interests while respecting their moral values (i.e., be both utility- and value-driven). This combination of features has been implemented in an agent-based model, ELVIRA, which provides MPC solutions of better quality and more satisfactory for users than other state-of-the-art models, according to the simulations and the user study that I have performed. Hence, ELVIRA is an autonomous agent that would effectively help users of online social networks manage multi-user privacy (cf. **RQ**), and my simulations suggest that it would be adopted by users, as opposed to current or alternative mechanisms, when deployed under different conditions.

Value-alignment I designed two agent-based models, namely JIMMY and ELVIRA, that are aligned with their users' values when recommending decisions w.r.t. multi-party privacy in OSNs (cf. **RQ-B**). By relying on the Schwartz Theory of Basic Values, which asserts that values drive human behaviour, the two agent architectures aim to recommend, during the interactions with other users, decisions and behaviour that are consistent with the moral and attitude-related preferences of their users. However, the architectures are independent of the specific theory of values and could be easily adapted to other values and behavioural attitudes. Furthermore, the architectures are not strictly coupled to the privacy scenario and could be easily tailored to identify value-aligned actions or solutions in other contexts.

Explainability While considering explainable AI as a design requirement that improves the usability of AI tools rather than a solution concept to 'black-box' models, I have designed an explainable agent-based model, ELVIRA, that generates tailored justifications for the MPC solutions that it recommends (cf. **RQ-C**). In order to provide explanations, the agent first engages in practical reasoning to identify the best action to solve the privacy conflict, and then it conveys the outcome of its abductive reasoning process to the user. I have designed the format of the explanation being informed by a user study, which has provided useful insights on what users prefer to see in an explanation w.r.t. privacy decisions. The satisfaction for these explanations has been positively evaluated in another, subsequent, between-subjects user study, where the agent's recommendations and explanations were more frequently and better accepted than the ones generated by other state-of-the-art models. Again, the explainable agent architecture is independent of the privacy context and could be easily adapted to generate solutions and explanations in other scenarios.

9.3 Limitations

Despite the positive theoretical and empirical results that follow from the design and evaluation of the ELVIRA agent, this model to support the collaborative management of online multi-user privacy is far from being perfect. In the following, I report and discuss its main limitations.

Values and behaviour Even though the Schwartz values have been recognised across different cultures and it is widely accepted that moral values influence people's behaviours [24], there is no empirical knowledge available regarding the mapping of values onto actions in specific contexts, such as the privacy one. Therefore, I designed the value layer of ELVIRA in the most objective way I could, but the relationship between values and behaviour in privacy contexts that I have identified may not be universally valid and may reflect a Western cultural bias (mine) [32, 212]. Anyways, if a better value theory or more universal insights on mappings of values and behaviours were made available, ELVIRA's architecture could be effortlessly adapted without losing any of their properties.

Collaborative behaviour My entire research has built on the assumption that online social networks users are generally collaborative w.r.t. the resolution of MPCs, as suggested in previous empirical studies [175]. However, in reality there exist some other circumstances, e.g. (i) when some user has malicious intentions and (ii) when users are more self-interested than interested in the social good. In circumstance (i), which fortunately has a low incidence, ELVIRA may not be able to support the resolution of the MPC, but it would still hinder the malicious user's intention by not allowing to directly upload and share the critical content on the OSN without everyone's approval (similarly to the cases when ELVIRA's optimal recommendation is not accepted by all the users and they may proceed to an offline negotiation). On the other hand, in circumstance (ii), users (or agents) may be tempted to share untruthful preferences in order to game the system and receive a better outcome (notice that untruthful evaluations of the solutions are not permitted by the ELVIRA model). While this behaviour may at least partially be modelled by the self-enhancement value ("getting your way") and prioritised for anyone who prefers this value over the others, ELVIRA is not able to detect and/or oppose self-interested users. To conclude on a positive note, let me remind the reader that the user study described in Section 7 suggests that, by interacting with the ELVIRA agent, users are nudged towards a more pro-social behaviour than with the other tested models, hinting that perceived fairness in the decision making process may generally be more important than self-interest.

9.4 Future Directions

The body of work that I have presented in this thesis offers several further directions of research, which I briefly outline in the following.

Dialogical explanations According to the feedback received by users (see Chapter 7), the explanations generated by ELVIRA are satisfactory and helpful for solving MPCs. However, it may be possible to improve their quality further by making them even more tailored and dynamic. For instance, the definition of a dialogical human-agent interaction protocol would enable the agent to provide just the information that is explicitly requested by the user, avoiding redundancy. Furthermore, the agent may learn the user's preferences not only about how to manage MPCs, which is already considered in the literature, but also about the human-agent interaction itself, where explanation features such as technicality, length, content, and so on, may be tailored further over time.

Evolution on a network When considering the adoption of ELVIRA for solving MPCs by a population, ELVIRA is shown to be, most of the times, the dominant strategy in an evolutionary competition game, as I described in Chapter 8. In the analysis I performed, I had considered social learning, i.e., the dynamics of imitating other more successful players, among any member of the population. Yet, online social networks, given their structure, would allow for a more specific investigation of social learning, not only at the entire population level, but also in the more restricted context of cliques, groups and communities. In fact, some influential users may significantly speed up or slow down the invasion of a strategy, or may modify the overall emerging dynamics.

Malicious behaviour In this thesis I have tackled the challenge of *collaboratively* managing multi-user privacy. This is important, as it has been shown that the large majority of MPCs occurs in non adversarial settings, where the incapability of uploaders to make appropriate privacy decisions when managing co-owned content is one of the main causes for MPCs [175]. However, even though much more rarely, MPCs may also be generated on purpose by users with ill-intentions, such as in the cases of cyber-bulling and revenge-porn [197]. Through evolutionary games, similarly to what presented in Chapter 8, it would be possible to study the evolutionary stability of ELVIRA when malicious players are part of the population. Different malicious types of strategies may be defined for this purpose, for instance representing (i) casual disruptors, who are just rarely adversarial, (ii) constant disruptors, who are always adversarial, and (iii) vindictive disruptors, who are reactively adversarial.

Beyond OSNs The management of multi-user privacy is not a problem peculiar to OSNs, but it is common to several other contexts such as smartphones [107, 132], collaborative platforms in the cloud [143, 128], smart homes [69, 1] and so on. Further research may shed light on the possibility of adapting the ELVIRA architecture, as defined in Chapter 5, so that it could support users also beyond ONSs. In fact, by eventually revising the behavioural interpretation of the Schwartz values (or other values) and defining new appropriate actions, one could adjust the current practical reasoning process in order to identify optimal multi-user privacy-related actions in new collaborative scenarios. This would still allow to generate explanations to justify value-aligned decisions, which are crucial features for the safe deployment of autonomous systems in society.

Appendix A

Design of the User Studies

Here I report the full specifications of the design of the user study for evaluating ELVIRA's explanations (**US1**) (see Chapter 5, Section 5.6.1) and for the user study for evaluating the quality of ELVIRA's recommendations against other models (**US2**) (see Chapter 7, Section 7.2). A relevant portion of the questionnaires for the two studies coincides, hence I describe the two designs together. If not specified otherwise, each of the following parts appears in both questionnaires.

The text parts in bold were not shown to the participants. The pictures were all taken from [59], but I needed the original version of them, as they had been published blurred. Showing them unblurred to participants was important so they could feel more immersed in the scenarios. I contacted the authors of [59], and I was able to get all pictures unblurred but one, which is about a lunch with colleagues, but we simply replaced it with a very similar one (Figure A.1a). Also, for details about participants' perceptions on sensitivities and relationships of the photos, please refer to [59], in which this information is reported. Importantly, they are representative of different sensitivities and relationship types.

A.1 Part 1: Value Elicitation through PVQ-21

In US2, only for ELVIRA and value-based treatment [156]

Please read the following descriptions and select how much the person in the description is like you. *['Very much like me', 'Like me', 'Somewhat like me', 'A little like me', 'Not like me', 'Not like me at all']*

1. Thinking up new ideas and being creative is important to him/her. He/she likes to do things in his/her own original way.

2. It is important to him/her to be rich. He/she wants to have a lot of money and expensive things.

3. He/she thinks it is important that every person in the world should be treated equally. He/she believes everyone should have equal opportunities in life.

4. It is important to him/her to show his/her abilities. He/she wants people to admire what he/she does.

5. It is important to him/her to live in secure surroundings. He/she avoids anything that might endanger his/her safety.

6. He/she likes surprises and is always looking for new things to do. He/she thinks it is important to do lots of different things in life.

7. He/she believes that people should do what they're told. He/she thinks people should follow rules at all times, even when no-one is watching.

8. It is important to him/her to listen to people who are different from him/her. Even when he/she disagrees with them, he/she still wants to understand them.

9. It is important to him/her to be humble and modest. He/she tries not to draw attention to him/herself.

10. Having a good time is important to him/her. He/she likes to "spoil" him/herself.

11. It is important to him/her to make his/her own decisions about what he/she does. He/she likes to be free and not depend on others.

12. It is very important to him/her to help the people around him/her. He/she wants to care for their well-being.

13. Being very successful is important to him/her. He/she hopes people will recognise his/her achievements.

14. It is important to him/her that the government ensures his/her safety against all threats. He/she wants the state to be strong so it can defend its citizens.

15. He/she looks for adventures and likes to take risks. He/she wants to have an exciting life.

16. It is important to him/her always to behave properly. He/she wants to avoid doing anything people would say is wrong.

17. It is important to him/her to get respect from others. He/she wants people to do what he/she says.

18. It is important to him/her to be loyal to his/her friends. He/she wants to devote him/herself to people close to him/her.

19. He/she strongly believes that people should care for nature. Looking after the environment is important to him/her.

20. Tradition is important to him/her. He/she tries to follow the customs handed down by his/her religion or his/her family.

21. He/she seeks every chance he/she can to have fun. It is important to him/her to do things that give him/her pleasure.

22. **Attention Check 1:** He/she looks like a yellow zebra. Click on not like me at all.

Feedback on Schwartz values elicitation

Given your previous answers, we think that your ranking of behavioural tendencies is as follows. Each item represents an action that may be executed in the context of a multi-party deliberation, for instance when a group of people needs to decide how to share online a picture that involves them all (e.g., they are all depicted in the photo). [(1) represents what you care most about, (4) represents what you care least about.]

(1) action1

(2) action2

(3) action3

(4) action4

Actions to be presented in the rank above according to the answers to the PVQ-21:

[SE] ‘getting your way’

[ST] ‘making others happy’

[OTC] ‘everyone compromising the same’

[CO] ‘preserving everyone’s privacy’

1. Do you agree with this ranking of your behavioural tendencies? *[5 points Likert scale anchored with ‘Strongly agree’ and ‘Strongly disagree’]*

2. If you disagree with this ranking, how would you change it? Please provide your own ranking. *[open text box]*

3. Do you have any other comments about the ranking? *[open text box]*

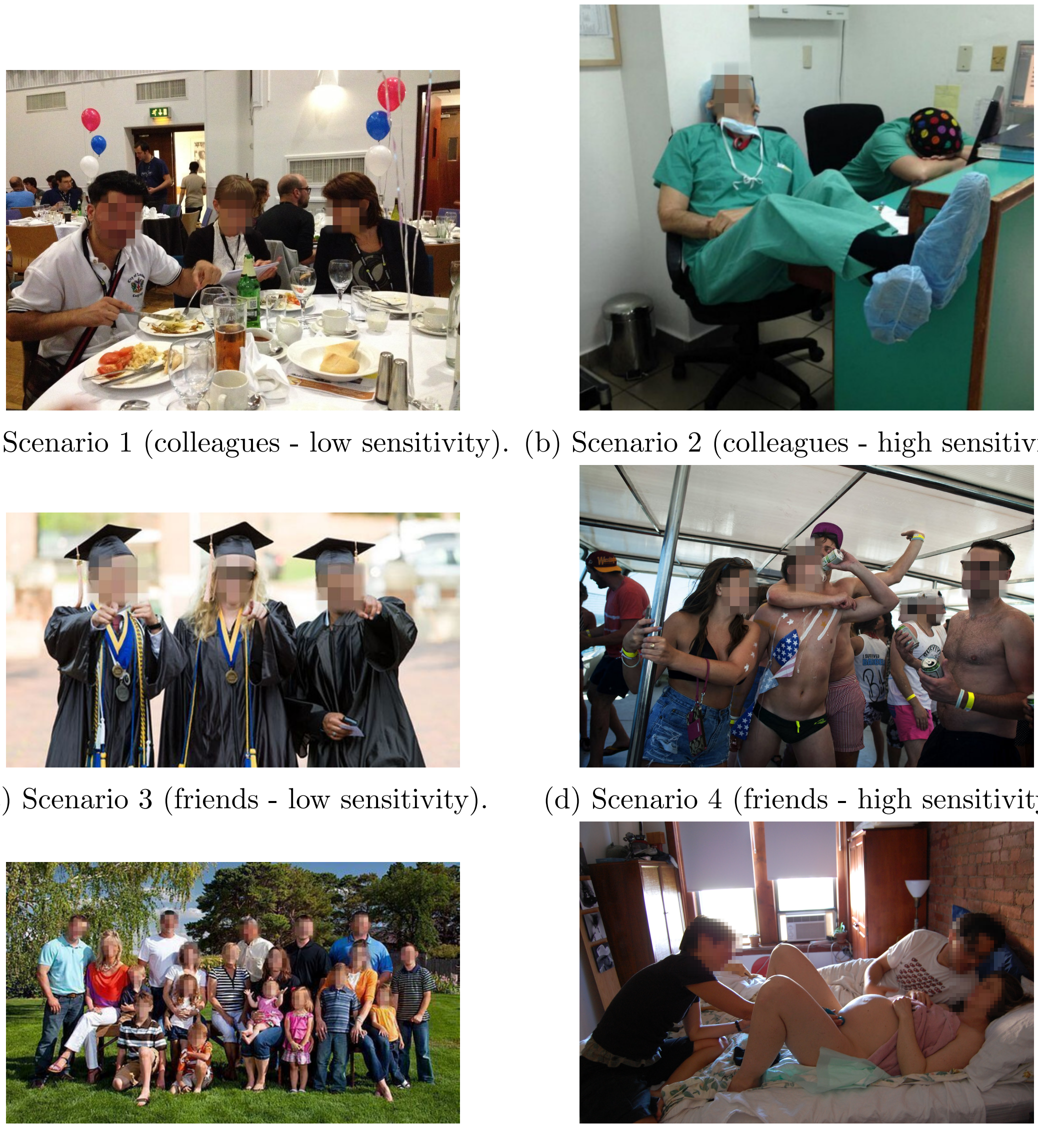
A.2 Part 2: MPCs Scenarios

The same scenarios were used in US1 (a random sample of 3) and US2 (all 6)

US1: In the next part we will ask you to consider, one at a time, three photos. You will be asked to suggest the level of publicity/privacy you would assign to them. Then, ELVIRA will recommend a solution to a simulated multiuser privacy conflict that may emerge from people having different preferences about how to share the same photo online. ELVIRA will show three different explanations for the same recommended solution and you will be asked to evaluate them.

US2: In the next part we will ask you to consider, one at a time, six photos. For each photo, you will be asked to suggest the level of publicity/privacy you would assign to it. Then, we will describe a multiparty privacy conflict that may emerge from people having different preferences about how to share the same photo online.

For each photo, you will be asked about the acceptability of the solution that the tool will recommend. As part of this research, we implemented different versions of a tool that aims to recommend a solution for each simulated conflict. In this study you will be exposed to only one of these versions, so please be very honest with your answers: a sugar-coated feedback will not be helpful to comparatively evaluate the different versions of the tool.



(a) Scenario 1 (colleagues - low sensitivity). (b) Scenario 2 (colleagues - high sensitivity).

(c) Scenario 3 (friends - low sensitivity). (d) Scenario 4 (friends - high sensitivity).

(e) Scenario 5 (family - low sensitivity). (f) Scenario 6 (family - high sensitivity).

Figure A.1: Pictures presented as part of the scenarios in the user study.

Presentation of the scenarios, in a random order; for each scenario,

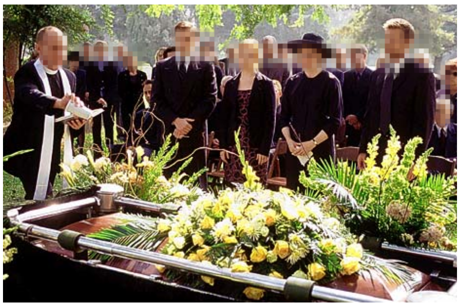


Figure A.2: Attention Check.

questions a-b-c are shown.

Consider this situation: [Figures A.1a,A.1b,A.1c,A.1d,A.1e,A.1f]

1. Felipe, Maria and Carla, three junior employees in a company, attended a business lunch in which they meet their seniors. One of the other employees took the following picture and sent it to Felipe. Felipe wants to upload the picture to his social media account.

2. The hospital where Bryan, Martin, and Sophia work has recently changed its shift policy making shifts much longer. Doctors complain that these shifts leave them exhausted. During one such long shift, at 4am, Bryan takes a picture of his two colleagues Martin and Sophia sleeping while they wait for another patient to come to emergencies. Bryan wants to upload the picture to his social media account, a few days after the picture was taken.

3. Tim, Ashley, and Jerry just graduated. Tim's father took the picture above after the graduation ceremony. Tim wants to upload the picture to his social media account.

4. Three friends, Mark (the groom), Alex, and John, go on a boat in Ibiza during a bachelor party. They get drunk and meet some girls. This is one of the pictures Alex took during that party. Alex wants to upload the picture to his social media account, the day after the party.

5. The Moore brothers (Frank, James and Nick) and their parents, wives, and

children took part in a photoshoot. The following is the best picture from the photo shoot. Frank wants to upload the picture to his social media account.
6. Dolores and Philip decide to have their baby, Rose, at home with the help of Ann, who is Dolores' sister and a midwife. They took the picture below during the labour. Philip wants to upload the picture to his social media account, a few days after Rose was born.

a. If you were [Felipe/Bryan/Tim/Alex/Frank/Philip], with whom would you share this picture on a social network? *['Just among themselves', 'With common friends', 'With friends of friends', 'Publicly']*
b. If fewer people than the ones included in your preferred choice had access to the picture, how would you feel? *['I would be very unhappy about it.', 'I would be unhappy about it.', 'I would be neutral about it.', 'I would be happy about it.', 'I would be very happy about it.']*
c. If more people than the ones included in your preferred choice had access to the picture, how would you feel? *['I would be very unhappy about it.', 'I would be unhappy about it.', 'I would be neutral about it.', 'I would be happy about it.', 'I would be very happy about it.']*

Questions b. and c. are shown just when appropriate (e.g., if the answer to a. is 'Publicly', only b. is shown).

### Just in US1 ###
The following explanations are presented for each scenario in a random order (identified by A, B, C): {user}, {user1} and {user2} are the actors of the scenario; {P} is the participant's preference (answer to question a. of each scenario), {P1} and {P2} are the preferences of other two simulated users involved in the conflict; {O} is the optimal solution identified by ELVIRA. The actions corresponding to promoted or demoted

behaviours are the following:

[SE] ‘The user would [not] get his/her way.’

[ST] ‘The user would [not] make others happy.’

[CO] ‘Everyone’s privacy would [not] be preserved.’

[OTC] ‘Everyone would [not] compromise the same.’

A multi-user privacy conflict to share this content occurred, because the sharing preferences of the involved people do not coincide. You suggested for {user} to share {P}; {user1} opted for sharing {P1} and {user2} would like to share {P2}. Given the occurrence of this conflict, ELVIRA computes an optimal solution and can present it in different ways:

[**No explanation**] to share {O} is the best compromise that solves the conflict.

[**General explanation**] to share {O is the best compromise that solves the conflict because it satisfies as much as possible everyone’s initial sharing preference, and because it enables actions that are coherent with everyone’s ranking of behavioural tendencies. Notably, by selecting to share O, the user would actions corresponding to the promoted values.

[**Contrastive explanation**] to share {O} is the best compromise that solves the conflict because it satisfies as much as possible everyone’s initial sharing preferences, and because it enables actions that are coherent with everyone’s ranking of behavioural tendencies. [*If O coincides with P*] This is also your preference! [*Else*] Notice that to share {P} (your initial sharing suggestion) would not allow the involved users to find a compromise, because other users may experience negative consequences. [*If O is more private than P and preference for undersharing (answer to question b. for each scenario)*] Also, you said that it would be ok sharing with fewer people. [*If O is more public than P and preference for oversharing (answer to question b. for each scenario)*] Also, you said that it would be ok sharing with more people. In addition, by selecting to share {O}, {actions promoting values} that would not be the case if sharing {P}. Furthermore, by selecting to share {P}, {actions demoting values}.

For each scenario, the participant evaluates the three types of explanation, A, B and C, through the Satisfaction Scale (Q1-8). Q9 is an attention check. *Question text*

A. [5-point scale anchored with 'strongly disagree' and 'strongly agree']

B. [5-point scale anchored with 'strongly disagree' and 'strongly agree']

C. [5-point scale anchored with 'strongly disagree' and 'strongly agree']

1. From the explanation, I could understand how ELVIRA works.
2. The explanation I received is satisfying.
3. The explanation provided sufficient detail about how ELVIRA works.
4. The explanation provided complete information about how ELVIRA works.
5. The explanation tells me how to use ELVIRA.
6. The explanation that ELVIRA provided are useful to my goals.
7. The explanation showed me how accurate ELVIRA is.
8. The explanation let me judge when I should trust and not trust ELVIRA.
9. Giraffes are blue and orange. Select 'Strongly disagree' for all A, B and C. **(Attention Check: Different but equivalent for each scenario)**
10. If you were Felipe in this situation, which explanation would you prefer to receive from ELVIRA? [A,B,C]

### Just in US2 ###

After 3 scenarios, Attention Check 2

Consider this situation and follow the instructions. [Figure A.2]

Albert, Sarah, Violet and John (in this order in the foreground, from left to right) attended their mother's funeral. After some days, they discuss about sharing this picture on a social network.

In this question, you don't need to express your preference, please click on 'Publicly'. *['Just among themselves', 'With common friends', 'With friends of friends', 'Publicly']*

Presentation of the MPC and the Recommended Solution

For ELVIRA:

Conflict: The sharing preferences of the other people involved do not coincide with yours. You suggested for {user} to share {P}; {user1} opted for sharing {P1} and {user2} would like to share {P2}.

Solution: {ELVIRA recommendation}

For utility-based and value-based:

Conflict: The sharing preferences of the other people involved do not coincide with yours.

Solution: The conflict would be solved by sharing solutionUtility or solutionValues

For Facebook model:

{UserUploader} uploads this content online and shares it with {UploaderPolicy}.

Where {user}, {user1} and {user2} are the actors of the scenario; {userUploader} is the uploader, randomly selected among the users; {P} is the participant's preference (answer to question a. of each scenario), {P1} and {P2} are the preferences of other two simulated users involved in the conflict.

d. If you were {user1}, how likely would you be to accept to share {solution}? *[5-points scale anchored with 'Very unlikely' and 'Very likely']*

e. In your opinion, how likely would it be for everyone involved to accept to share this content {solution}? *[5-points scale anchored with 'Very unlikely' and 'Very likely']*

f. Please provide some details on the motivations that supported your answers in the questions 1 and 2. *[open text box]*

A.3 Part 3: Satisfaction

Only in US2

You interacted a number of times with a tool that aims to support users to resolve multi-user privacy conflicts on social media. In each scenario, the tool recommended to you what it thought to be the optimal solution given the simulated scenarios we presented to you. Based on this, we would like to ask you for feedback on the tool itself.

Please be very honest with your answers. In this study, you have interacted with only one out of four alternative designs that have been implemented as part of this research, which may lead to different solutions or ways of explaining them. It is very important to us that you evaluate your experience truthfully, to help us understand which of the alternatives might be better and adapt our future research accordingly.

Satisfaction Scale [79] *[5-point scale anchored with 'strongly disagree' and 'strongly agree']*

1. From the output, I could understand how the tool works.
2. The output I received is satisfying.
3. The output provided sufficient detail about how the tool works.
4. The output provided complete information about how the tool works.
5. The output tells me how to use the tool.
6. The output that the tool provided are useful to my goals.
7. The output showed me how accurate the tool is.
8. The output let me judge when I should trust and not trust the tool.
9. **Attention Check 3:** Giraffes are blue and orange. Select 'Strongly disagree'.

A.4 Part 4: Privacy

Questions 1-10 are the IUIPC scale [105]. *[5-point scale anchored with 'Strongly disagree' and 'Strongly agree']*

1. Consumer online privacy is really a matter of consumers' right to exercise control and autonomy over decisions about how their information is collected, used, and shared.

2. Consumer control of personal information lies at the heart of consumer privacy.

3. I believe that online privacy is invaded when control is lost or unwillingly reduced as a result of a marketing transaction.

4. Companies seeking information online should disclose the way the data are collected, processed, and used.

5. A good consumer online privacy policy should have a clear and conspicuous disclosure.

6. It is very important to me that I am aware and knowledgeable about how my personal information will be used.

7. It usually bothers me when online companies ask me for personal information.

8. When online companies ask me for personal information, I sometimes think twice before providing it.

9. It bothers me to give personal information to so many online companies.

10. I am concerned that online companies are collecting too much personal information about me.

11. Which of the following social networking and social media platforms do you use? *[Facebook, Instagram, WhatsApp, Flickr, Pinterest, Twitter, YouTube, Snapchat, other]*

12. How often do you access any of the above platforms? *[Daily, 2-3 times per week, once a week, 2-3 times per month, once a month, less than once a month]*

13. How often do you share photographic content depicting other people on any of the above-mentioned platforms? *[Daily, 2-3 times per week, once a week, 2-3 times per month, once a month, less than once a month]*

14. In the last year, how often was someone unhappy about a picture that you had shared online? [never, once, twice, more than 2 times]

15. How often other people share photographic content depicting you on any of the above-mentioned platforms? *[Daily, 2-3 times per week, once a week, 2-3 times per month, once a month, less than once a month]*

16. In the last year, how often were you unhappy about a picture that some other person had shared online? *[never, once, twice, more than 2 times]*

Appendix B

Ethical Approval for the User Studies

B.1 Information Sheet for Participants

[version 1.0, 11/08/2020]

Ethical Clearance Reference Number: LRS-19/20-19405

Title of project Evaluation of ELVIRA, an explainable agent for value- and utility-driven multiuser privacy

Invitation Paragraph I would like to invite you to participate in this research project which forms part of my PhD research. Before you decide whether you want to take part, it is important for you to understand why the research is being done and what your participation will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask me if there is anything that is not clear or if you would like more information.

What is the purpose of the project? The purpose of the project is to evaluate the appropriateness of the support that ELVIRA, a software agent that I developed as part of my PhD, offers to users of online social networks when managing multiuser privacy. ELVIRA provides recommendations regarding the level of publicity or privacy to assign to some item that involves multiple people (e.g., a picture) when eventually sharing it online. The recommendation is informed by the individual privacy and behavioural preferences of all the users involved. The results of this project may contribute to find future and more effective solutions to privacy conflicts in social media.

Why have I been invited to take part? You are being invited to participate in this project as part of the pool of participants that the platform Prolific offers to researchers.

What will happen if I take part? If you choose to take part in the project you will be asked to complete three parts of a questionnaire. In the first part, you will answer some questions regarding your general behavioural attitudes. The second part, which will be partly informed by your answers to the first part, will ask you to consider, one at a time, three photos. You will be asked to suggest the level of publicity/privacy you would assign to them. Then, ELVIRA will recommend a solution to a simulated multiuser privacy conflict that may emerge from people having different preferences about how to share the same photos online. The recommended solution will be accompanied by alternative explanations, which you will be asked to evaluate. In the last part, you will be asked some questions regarding general privacy preferences. Participation will take place online and the questionnaire is expected to take about 30 minutes.

Do I have to take part? Participation is completely voluntary. You should only take part if you want to and choosing not to take part will not disadvantage you in anyway. Once you have read the information sheet, please contact me if you have any questions that will help you make a decision about taking part.

Incentives Every participant who completes successfully the study will receive a reward equal to £3.75 through the Prolific platform.

What are the possible risks of taking part? There are no foreseeable risks in taking part in this study.

What are the possible benefits of taking part? The information that will be collected during the study may help in developing novel solutions to multi-privacy conflicts in social media. Also, the study could increase the awareness of the participant regarding the management of multiuser privacy.

Data handling and confidentiality Your data will be processed in accordance with the General Data Protection Regulation 2016 (GDPR). It will not be possible to identify you through your Prolific ID and I will only access anonymised demographic data from Prolific. After 30 days from the completion of the study, I will permanently and irreversibly remove the information on the Prolific IDs of the participants too, therefore completely anonymising the resulting dataset. I may also publish this resulting anonymised version of the dataset for enabling further research.

Data Protection Statement Your data will be processed in accordance with the General Data Protection Regulation 2016 (GDPR). If you would like more information about how your data will be processed in accordance with GDPR please visit the link below: `https://www.kcl.ac.uk/research/support/research-ethics/kings-college-london-statement-on-use-of-personal-data-in-research`

What if I change my mind about taking part? You are free to withdraw at any point of the project, without having to give a reason. Withdrawing from the project will not affect you in any way. You are able to withdraw your data from the project up until 30 days after the completion of the study, after which withdrawal of your data will no longer be possible due to anonymisation of the dataset that will precede its analysis. If you choose to withdraw from the project in this timeframe, you need to contact me by email (francesca.mosca@kcl.ac.uk) including your Prolific ID. I will not retain the information you have given thus far.

How is the project being funded? This project is being funded by King's College London.

What will happen to the results of the project? The results of the project will be summarised in academic publications and as part of my PhD dissertation, where I will refer to only anonymised information. The anonymised dataset will be made publicly available.

Who should I contact for further information? If you have any questions or require more information about this project, please contact me using the following contact details:

Francesca Mosca
PhD Candidate in Computer Science at King's College London
francesca.mosca@kcl.ac.uk

What if I have further questions, or if something goes wrong? If this project has harmed you in any way or if you wish to make a complaint about the conduct of the project you can contact King's College London using the details below for further advice and information:

Dr Jose M. Such
Reader (Associate Professor) in Security and Privacy
Director, KCL Cybersecurity Centre
Department of Informatics

King's College London
jose.such@kcl.ac.uk

Thank you for reading this information and for considering taking part in this research.

CONSENT FORM FOR PARTICIPANTS IN RESEARCH PROJECTS

1. I consent voluntarily to be a participant in this project and understand that I can refuse to take part and can withdraw from the project at any time, without having to give a reason, in the next 30 days.

2. I consent to the processing of my personal information for the purposes explained to me in the Information Sheet. I understand that such information will be handled in accordance with the terms of the General Data Protection Regulation (GDPR) and the UK Data Protection Act 2018.

3. I understand that my information may be subject to review by responsible individuals from the College for monitoring and audit purposes.

4. I understand that confidentiality and anonymity will be maintained, and it will not be possible to identify me in any research outputs.

5. I consent to my data being shared publicly in the form of an anonymised dataset.

I confirm that I have read and understood the description of this study and the consequences of my participation. [**to be ticked to proceed**]

B.2 Ethical Approval

Research Ethics
Office

Franklin Wilkins Building
5.9 Waterloo Bridge Wing
Waterloo Road
London SE1 9NH
Telephone 020 7848 4020/4070/4077
rec@kcl.ac.uk

Francesca Mosca

11 August 2020

Dear Francesca

LRS-19/20-19405 - Evaluation of ELVIRA, an explainable agent for value- and utility-driven multiuser privacy

Thank you for submitting your application for the above project. I am pleased to inform you that your application has now be approved with the provisos indicated at the end of this letter. All changes must be made before data collection commences. The Committee does not need to see evidence of these changes, however supervisors are responsible for ensuring that students implement any requested changes before data collection commences.

IMPORTANT CORONAVIRUS UPDATE: In light of the COVID-19 pandemic, the College Research Ethics Committee has temporarily suspended all primary data collection involving face to face participant interactions until further notice. **Ethical clearance for this project is granted. However, the clearance outlined in the attached letter is contingent on your adherence to the latest College measures when conducting your research.** Please do not commence data collection until you have carefully reviewed the update and made any necessary project changes:

Ethical approval has been granted for a period of **three years** from11 August 2020 You will not be sent a reminder when your approval has lapsed and if you require an extension you should complete a modification request, details of which can be found here:

https://internal.kcl.ac.uk/innovation/research/ethics/applications/modifications.aspx

Please ensure that you follow the guidelines for good research practice as laid out in UKRIO's Code of Practice for research: http://ukrio.org/publications/code-of-practice-for-research/

Any unforeseen ethical problems arising during the course of the project should be reported to the panel Chair, via the Research Ethics Office.

Please note that we may, for the purposes of audit, contact you to ascertain the status of your research.

We wish you every success with your research.

Yours sincerely,
Miss Elizabeth Chuck

Senior Research Ethics Officer

For and on behalf of:
BDM Research Ethics Panel

Approved with Provisos

Review Reference: LRS-19/20-19405

Major Issues (will require substantial consideration by the applicant before approval can be granted)

N/A

Minor Issues related to application (the reviewer should identify the relevant section number before each comment)

1. B2: Please ensure that data collection does not commence until full approval is granted.

2. E5: It is recommended that research data should be password protected and stored with KCL using OneDrive or SharePoint https://www.kcl.ac.uk/researchsupport/managing/store

3. E7: Please retain research data in accordance with the King's Data Retention Schedule: https://www.kcl.ac.uk/aboutkings/orgstructure/ps/audit/records/retention

Minor Issues related to recruitment documents

Information Sheet

4. Insert a date and version number.

5. Data handling and confidentiality: ‘The only personally identifiable information that we will collect is your Prolific ID.’ It is understood that participants cannot be identified by the researcher using their Prolific ID, please clarify. Note that the Prolific website states ‘We will not disclose personal data between Participants and Researchers, although Researchers will see anonymized demographic data relating to Participants for screening purposes.’

6. What if I change my mind: Please outline the process for withdrawing data, for example, will participants be required to email you with their Prolific ID?

Advice and Comments (do not have to be adhered to, but may help to improve the research)

N/A

Bibliography

[1] Noura Abdi, Xiao Zhan, Kopo M Ramokapane, and Jose Such. Privacy norms for smart home personal assistants. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–14, 2021.

[2] A. Acquisti, L. Brandimarte, and G. Loewenstein. Privacy and human behavior in the age of information. *Science*, 347(6221):509–514, 2015.

[3] Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, et al. Nudges for privacy and security: Understanding and assisting users' choices online. *ACM Computing Surveys (CSUR)*, 50(3):44, 2017.

[4] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, 6:52138–52160, 2018.

[5] Nirav Ajmeri, Hui Guo, Pradeep K Murukannaiah, and Munindar P Singh. Elessar: Ethics in norm-aware agents. In *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pages 16–24, 2020.

[6] 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.

[7] Davide Alberto Albertini, Barbara Carminati, and Elena Ferrari. Privacy settings recommender for online social network. In *2016 IEEE 2nd International Conference on Collaboration and Internet Computing (CIC)*, pages 514–521. IEEE, 2016.

[8] Gordon Willard Allport, Philip Ewart Vernon, and Gardner Lindzey. *Study of values: Manual.* Riverside Publishing Company, 1970.

[9] Irwin Altman. The environment and social behavior: privacy, personal space, territory, and crowding. 1975.

[10] Saleema Amershi, James Fogarty, and Daniel Weld. Regroup: Interactive machine learning for on-demand group creation in social networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 21–30, 2012.

[11] Michael Anderson, Susan Leigh Anderson, and Vincent Berenz. A value driven agent: Instantiation of a case-supported principle-based behavior paradigm. In *AAAI Workshops*, 2017.

[12] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

[13] K. Atkinson and T. Bench-Capon. Action-based alternating transition systems for arguments about action. In *AAAI*, volume 7, pages 24–29, 2007.

[14] K. Atkinson and T. Bench-Capon. Practical reasoning as presumptive argumentation using action based alternating transition systems. *Artificial Intelligence*, 171(10-15):855–874, 2007.

[15] K. Atkinson and T. Bench-Capon. Taking account of the actions of others in value-based reasoning. *Artificial Intelligence*, 254:1–20, 2018.

[16] Katie Atkinson and Trevor Bench-Capon. Addressing moral problems through practical reasoning. *Journal of Applied Logic*, 6(2):135–151, 2008.

[17] Katie Atkinson, Trevor Bench-Capon, and Peter McBurney. Computational representation of practical argument. *Synthese*, 152(2):157–206, 2006.

[18] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The moral machine experiment. *Nature*, 563(7729):59–64, 2018.

[19] Robert Axelrod and William D Hamilton. The evolution of cooperation. *science*, 211(4489):1390–1396, 1981.

[20] Tim Baarslag, Alper T Alan, Richard C Gomer, Ilaria Liccardi, Helia Marreiros, Enrico H Gerding, and MC Schraefel. Negotiation as an interaction mechanism for deciding app permissions. In *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*, pages 2012–2019, 2016.

[21] Tim Baarslag, Alan Alper, Richard Gomer, Muddasser Alam, Perera Charith, Enrico Gerding, et al. An automated negotiation agent for permission management. 2017.

[22] Ana Babić Rosario, Kristine de Valck, and Francesca Sotgiu. Conceptualizing the electronic word-of-mouth process: What we know and need to know about ewom creation, exposure, and evaluation. *Journal of the Academy of Marketing Science*, 48(3):422–448, 2020.

[23] A.L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[24] A. Bardi and S.H Schwartz. Values and behavior: Strength and structure of relations. *Personality and social psychology bulletin*, 29(10):1207–1220, 2003.

[25] Daniela Baum, Martin Spann, Johann Füller, and Carina Thürridl. The impact of social media campaigns on the success of new product introductions. *Journal of Retailing and Consumer Services*, 50:289–297, 2019.

[26] Filipe Beato and Roel Peeters. Collaborative joint content sharing for online social networks. In *2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS)*, pages 616–621. IEEE, 2014.

[27] Trevor JM Bench-Capon. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation*, 13(3):429–448, 2003.

[28] A. Besmer and H.R. Lipford. Moving beyond untagging: photo privacy in a tagged world. In *CHI*, pages 1563–1572. ACM, 2010.

[29] Igor Bilogrevic, Kévin Huguenin, Berker Agir, Murtuza Jadliwala, and Jean-Pierre Hubaux. Adaptive information-sharing for privacy-aware mobile social networks. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 657–666, 2013.

[30] Edvard P Bjørgen, Simen Madsen, Therese S Bjørknes, Fredrik V Heimsæter, Robin Håvik, Morten Linderud, Per-Niklas Longberg, Louise A Dennis, and Marija Slavkovik. Cake, death, and trolleys: dilemmas as benchmarks of ethical decision-making. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 23–29, 2018.

[31] Jack Block. The five-factor framing of personality and beyond: Some ruminations. *Psychological Inquiry*, 21(1):2–25, 2010.

[32] Alan Borning and Michael Muller. Next steps for value sensitive design. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1125–1134, 2012.

[33] Christine Boshuijzen-van Burken, Ross Gore, Frank Dignum, Lamber Royakkers, Phillip Wozny, and F LeRon Shults. Agent-based modelling of values: The case of value sensitive design for refugee logistics. *JASSS: Journal of Artificial Societies and Social Simulation*, 23(4), 2020.

[34] Nick Bostrom. *Superintelligence.* Dunod, 2017.

[35] Brigitte Burgemeestre, Joris Hulstijn, and Yao-Hua Tan. Value-based argumentation for justifying compliance. *Artificial Intelligence and Law*, 19(2):149–186, 2011.

[36] Gul Calikli, Mark Law, Arosha K Bandara, Alessandra Russo, Luke Dickens, Blaine A Price, Avelie Stuart, Mark Levine, and Bashar Nuseibeh. Privacy dynamics: Learning privacy norms for social software. In *2016 IEEE/ACM 11th International Symposium on Software Engineering for Adaptive and Self-Managing Systems (SEAMS)*, pages 47–56. IEEE, 2016.

[37] B. Carminati and E. Ferrari. Collaborative access control in on-line social networks. In *CollaborateCom*, pages 231–240. IEEE, 2011.

[38] Tathagata Chakraborti, Sarath Sreedharan, Sachin Grover, and Subbarao Kambhampati. Plan explanations as model reconciliation–an empirical study. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 258–266. IEEE, 2019.

[39] Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. *arXiv preprint arXiv:1701.08317*, 2017.

[40] Ajay Chander and Ramya Srinivasan. Evaluating explanations by cognitive value. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 314–328. Springer, 2018.

[41] Benjamin P Chapman and Lewis R Goldberg. Act-frequency signatures of the big five. *Personality and Individual Differences*, 116:201–205, 2017.

[42] Joe Collenette, Katie Atkinson, and Trevor JM Bench-Capon. An explainable approach to deducing outcomes in european court of human rights cases using adfs. 2020.

[43] Jeffrey D Cone Jr, C Stephen Byrum, Wyatt G Payne, and David J Smith Jr. A novel adjuvant to the resident selection process: the hartman value profile. *Eplasty*, 12, 2012.

[44] Paul T Costa Jr and Robert R McCrae. *The Revised NEO Personality Inventory (NEO-PI-R).* Sage Publications, Inc, 2008.

[45] Stephen Cranefield, Nir Oren, and Wamberto W Vasconcelos. Accountability for practical reasoning agents. In *International Conference on Agreement Technologies*, pages 33–48. Springer, 2018.

[46] Stephen Cranefield, Michael Winikoff, Virginia Dignum, and Frank Dignum. No pizza for you: Value-based plan selection in bdi agents. In *IJCAI*, pages 178–184, 2017.

[47] Natalia Criado, Estefania Argente, and V Botti. Open issues for normative multi-agent systems. *AI communications*, 24(3):233–264, 2011.

[48] Natalia Criado and Jose Such. Implicit contextual integrity in online social networks. *Information Sciences*, 325:48–69, 2015.

[49] Mary L Cummings. Integrating ethics in design through the value-sensitive design approach. *Science and Engineering Ethics*, 12(4):701–715, 2006.

[50] Kristijonas Čyras, Myles Lee, and Dimitrios Letsios. Schedule explainer: An argumentation-supported tool for interactive explanations in makespan scheduling. In *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, pages 243–259. Springer, 2021.

[51] Kristijonas Čyras, Dimitrios Letsios, Ruth Misener, and Francesca Toni. Argumentation for explainable scheduling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2752–2759, 2019.

[52] Kristijonas Čyras, Antonio Rago, Emanuele Albini, Pietro Baroni, and Francesca Toni. Argumentative XAI: A survey. *arXiv preprint arXiv:2105.11266*, 2021.

[53] Judith DeCew. Privacy. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2018 edition, 2018.

[54] Morteza Dehghani, Emmett Tomai, Kenneth D Forbus, and Matthew Klenk. An integrated reasoning approach to moral decision-making. In *AAAI*, pages 1280–1286, 2008.

[55] Louise A Dennis and Nir Oren. Explaining BDI agent behaviour through dialogue. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021)*. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 2021.

[56] Virginia Dignum. *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Springer International Publishing, 2019.

[57] Lujun Fang and Kristen LeFevre. Privacy wizards for social networking sites. In *Proceedings of the 19th international conference on World wide web*, pages 351–360, 2010.

[58] Shaheen Fatima, Sarit Kraus, and Michael Wooldridge. *Principles of automated negotiation*. Cambridge University Press, 2014.

[59] R. Fogues, P. Murukannaiah, J. Such, and M. Singh. Sharing policies in multiuser privacy scenarios: Incorporating context, preferences, and arguments in decision making. *ACM TOCHI*, 24(1):5:1–5:29, 2017.

[60] R. Fogues, P. Murukannaiah, J. Such, and M. Singh. Sosharp: Recommending sharing policies in multiuser privacy scenarios. *IEEE Internet Computing*, 21(6):28–36, 2017.

[61] R. Fogues, J. Such, A. Espinosa, and A. Garcia-Fornes. Bff: A tool for eliciting tie strength and user communities in social networking services. *Information Systems Frontiers*, 16(2):225–237, 2014.

[62] Ricard Fogues, Jose Such, Agustin Espinosa, and Ana Garcia-Fornes. Open challenges in relationship-based privacy mechanisms for social network services. *International Journal of Human-Computer Interaction*, 31(5):350–370, 2015.

[63] Ricard Fogues, Jose Such, Agustin Espinosa, and Ana Garcia-Fornes. Tie and tag: A study of tie strength and tags for photo sharing. *PLOS ONE*, 13(8):1–22, 08 2018.

[64] Philippa Foot. The problem of abortion and the doctrine of the double effect. *Oxford review*, 5, 1967.

[65] Maria Fox, Derek Long, and Daniele Magazzeni. Explainable planning. *arXiv preprint arXiv:1709.10256*, 2017.

[66] B. Friedman, P.H. Kahn, and A. Borning. Value sensitive design and information systems. *The handbook of information and computer ethics*, pages 69–101, 2008.

[67] Batya Friedman and Peter H Kahn Jr. Human values, ethics, and design. In *The human-computer interaction handbook*, pages 1267–1292. CRC press, 2007.

[68] Carrie Gates. Access control requirements for web 2.0 security and privacy. *IEEE Web*, 2(0):12–15, 2007.

[69] Christine Geeng and Franziska Roesner. Who's in control? interactions in multi-user smart homes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019.

[70] Azin Ghazimatin, Oana Balalau, Rishiraj Saha Roy, and Gerhard Weikum. Prince: Provider-side interpretability with counterfactual explanations in recommender systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 196–204, 2020.

[71] Keith Gibbins and Iain Walker. Multiple interpretations of the rokeach value survey. *The Journal of Social Psychology*, 133(6):797–805, 1993.

[72] Lewis R Goldberg. An alternative" description of personality": the big-five factor structure. *Journal of personality and social psychology*, 59(6):1216, 1990.

[73] Lewis R Goldberg. The development of markers for the big-five factor structure. *Psychological assessment*, 4(1):26, 1992.

[74] Seda Gürses and Jose M Del Alamo. Privacy engineering: Shaping an emerging field of research and practice. *IEEE Security & Privacy*, 14(2):40–46, 2016.

[75] Werner Güth, Rolf Schmittberger, and Bernd Schwarze. An experimental analysis of ultimatum bargaining. *Journal of economic behavior & organization*, 3(4):367–388, 1982.

[76] Robert S Hartman. Formal axiology and the measurement of values. *The Journal of Value Inquiry*, 1(1):38–46, 1967.

[77] David J Hauser and Norbert Schwarz. Attentive turkers: Mturk participants perform better on online attention checks than do subject pool participants. *Behavior research methods*, 48(1):400–407, 2016.

[78] Samaneh Heidari, Maarten Jensen, and Frank Dignum. Simulations with values. In *Advances in Social Simulation*, pages 201–215. Springer, 2020.

[79] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608*, 2018.

[80] Geert Hofstede. Dimensionalizing cultures: The hofstede model in context. *Online readings in psychology and culture*, 2(1):2307–0919, 2011.

[81] H. Hu, G.J. Ahn, and J. Jorgensen. Detecting and resolving privacy conflicts for collaborative data sharing in online social networks. In *ACSAC*, pages 103–112. ACM, 2011.

[82] Hongxin Hu, Gail-Joon Ahn, and Jan Jorgensen. Multiparty access control for online social networks: model and mechanisms. *IEEE Transactions on Knowledge and Data Engineering*, 25(7):1614–1627, 2012.

[83] M. Humbert, B. Trubert, and K. Huguenin. A survey on interdependent privacy. *ACM Computing Surveys*, page 35, 2019.

[84] P. Ilia, I. Polakis, E. Athanasopoulos, F. Maggi, and S. Ioannidis. Face/off: preventing privacy leakage from photos in social networks. In *CCS*, pages 781–792, New York, New York, USA, 2015. ACM Press.

[85] Ronald Inglehart and Christian Welzel. *Modernization, cultural change, and democracy: The human development sequence*. Cambridge university press, 2005.

[86] Ronald Inglehart and Christian Welzel. Changing mass priorities: The link between modernization and democracy. *Perspectives on politics*, pages 551–567, 2010.

[87] Maritza Johnson, Serge Egelman, and Steven M Bellovin. Facebook and privacy: it's complicated. In *Proceedings of the eighth symposium on usable privacy and security*, pages 1–15, 2012.

[88] D. Kekulluoglu, N. Kökciyan, and P. Yolum. Preserving privacy as social responsibility in online social networks. *ACM TOIT*, 18(4):42, 2018.

[89] Berkant Kepez and Pιnar Yolum. Learning privacy rules cooperatively in online social networks. In *Proceedings of the 1st International Workshop on AI for Privacy and Security*, pages 1–4, 2016.

[90] Joseph Kim, Christian Muise, Ankit Shah, Shubham Agarwal, and Julie Shah.

Bayesian inference of linear temporal logic specifications for contrastive explanations. In *IJCAI*, pages 5591–5598, 2019.

[91] N. Kökciyan, N. Yaglikci, and P. Yolum. An argumentation approach for resolving privacy disputes in online social networks. *ACM TOIT*, 17(3):27, 2017.

[92] Nadin Kökciyan, Isabel Sassoon, Elizabeth Sklar, Sanjay Modgil, and Simon Parsons. Applying metalevel argumentation frameworks to support medical decision making. *IEEE Intelligent Systems*, 36(2):64–71, 2021.

[93] Richard E Kopelman, Janet L Rovenpor, and Mingwei Guan. The study of values: Construction of the fourth edition. *Journal of Vocational Behavior*, 62(2):203–220, 2003.

[94] H. Krasnova, S. Spiekermann, K. Koroleva, and T. Hildebrand. Online social networks: Why we disclose. *JIT*, 25(2):109–125, 2010.

[95] Sarit Kraus, Amos Azaria, Jelena Fiosina, Maike Greve, Noam Hazon, Lutz Kolbe, Tim-Benjamin Lembcke, Jorg P Muller, Soren Schleibaum, and Mark Vollrath. Ai for explaining decisions in multi-agent environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13534–13538, 2020.

[96] A Can Kurtan and Pınar Yolum. Assisting humans in privacy management: an agent-based approach. *Autonomous Agents and Multi-Agent Systems*, 35(1):1–33, 2021.

[97] A. Lampinen, V. Lehtinen, A. Lehmuskallio, and S. Tamminen. We're in it together: interpersonal management of disclosure in social network services. In *CHI*, pages 3217–3226. ACM, 2011.

[98] Pat Langley. Explainable, normative, and justified agency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9775–9779, 2019.

[99] J. Leskovec and J.J. Mcauley. Learning to discover social circles in ego networks. In *NIPS*, pages 539–547, 2012.

[100] Qingrui Li, Juan Li, Hui Wang, and Ashok Ginjala. Semantics-enhanced privacy recommendation for social networking sites. In *2011IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications*, pages 226–233. IEEE, 2011.

[101] Xiao-Hui Li, Caleb Chen Cao, Yuhan Shi, Wei Bai, Han Gao, Luyu Qiu, Cong Wang, Yuanyuan Gao, Shenjia Zhang, Xun Xue, et al. A survey of data-driven and knowledge-aware explainable AI. *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[102] Beishui Liao, Marija Slavkovik, and Leendert van der Torre. Building jiminy cricket: An architecture for moral agreements among stakeholders. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 147–153, 2019.

[103] Enrico Liscio, Michiel van der Meer, Luciano C Siebert, Catholijn M Jonker, Niek Mouter, and Pradeep K Murukannaiah. Axies: Identifying and evaluating context-specific values. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, pages 799–808, 2021.

[104] Andrea Loreggia, Nicholas Mattei, Francesca Rossi, and K Brent Venable. Preferences and ethical principles in decision making. In *2018 AAAI Spring Symposium Series*, 2018.

[105] Naresh K Malhotra, Sung S Kim, and James Agarwal. Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information systems research*, 15(4):336–355, 2004.

[106] Clara Mancini, Yvonne Rogers, Arosha K Bandara, Tony Coe, Lukasz Jedrzejczyk, Adam N Joinson, Blaine A Price, Keerthi Thomas, and Bashar Nuseibeh. Contravision: exploring users' reactions to futuristic technology. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 153–162, 2010.

[107] Maximilian Marsch, Jens Grossklags, and Sameer Patil. Won't you think of others?: Interdependent privacy in smartphone app permissions. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–35, 2021.

[108] Winter Mason and Siddharth Suri. Conducting behavioral research on amazon's mechanical turk. *Behavior research methods*, 44(1):1–23, 2012.

[109] Peter McBurney and Simon Parsons. Dialogue games for agent argumentation. In *Argumentation in artificial intelligence*, pages 261–280. Springer, 2009.

[110] David C McClelland. How motives, skills, and values determine what people do. *American psychologist*, 40(7):812, 1985.

[111] Yavuz Mester, Nadin Kökciyan, and Pınar Yolum. Negotiating privacy constraints in online social networks. In *International Workshop on Multiagent Foundations of Social Computing*, pages 112–129. Springer, 2015.

[112] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 2018.

[113] Tim Miller, Piers Howe, and Liz Sonenberg. Explainable ai: Beware of inmates

running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint arXiv:1712.00547*, 2017.

[114] A. Mislove, M. Marcon, K.P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *ICM*, pages 29–42. ACM, 2007.

[115] G. Misra and J. Such. How socially aware are social media privacy controls? *IEEE Computer*, 49(3):96–99, 2016.

[116] G. Misra and J. Such. Pacman: Personal agent for access control in social media. *IEEE Internet Computing*, 21(6):18–26, 2017.

[117] G. Misra, J. Such, and H. Balogun. Improve - identifying minimal profile vectors for similarity based access control. In *IEEE Trustcom*, pages 868–875, 2016.

[118] G. Misra, J. Such, and H. Balogun. Non-sharing communities? an empirical study of community detection for access control decisions. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 49–56, 2016.

[119] Gaurav Misra and Jose Such. React: Recommending access control decisions to social media users. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, page 421–426, 2017.

[120] James H Moor. The nature, importance, and difficulty of machine ethics. *IEEE intelligent systems*, 21(4):18–21, 2006.

[121] F. Mosca, J. Such, and P. McBurney. Value-driven collaborative privacy decision making. In *AAAI PAL Symposium*, 2019.

[122] Francesca Mosca. Value-aligned and explainable agents for collective decision making: Privacy application. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pages 2199–2200, 2020.

[123] Francesca Mosca, Ştefan Sarkadi, Jose Such, and Peter McBurney. Agent expri: Licence to explain. In *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems (EXTRAAMAS)*, pages 21–38. Springer, 2020.

[124] Francesca Mosca and Jose 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)*, 2021.

[125] Francesca Mosca and Jose Such. An explainable assistant for multiuser privacy. *Autonomous Agents and Multi-Agent Systems*, 36(1):10, 2022.

[126] Francesca Mosca, Jose Such, and Peter McBurney. Towards a value-driven explainable agent for collective privacy. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1937–1939, 2020.

[127] Md Sultan Al Nahian, Spencer Frazier, Mark Riedl, and Brent Harrison. Learning norms from stories: A prior for value aligned agents. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 124–130, 2020.

[128] Michael Nebeling, Matthias Geel, Oleksiy Syrotkin, and Moira C Norrie. Mubox: Multi-user aware personal cloud storage. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1855–1864, 2015.

[129] Helen Nissenbaum. A contextual approach to privacy online. *Daedalus*, 140(4):32–48, 2011.

[130] Piotr K Oles and Hubert JM Hermans. Allport-vernon study of values. *The Corsini encyclopedia of psychology*, pages 1–2, 2010.

[131] Alexandra-Mihaela Olteanu, Kévin Huguenin, Italo Dacosta, and J-P Hubaux. Consensual and privacy-preserving sharing of multi-subject and interdependent data. In *Proceedings of the 25th Network and Distributed System Security Symposium (NDSS)*, pages 1–16. Internet Society, 2018.

[132] Alexandra-Mihaela Olteanu, Kévin Huguenin, Reza Shokri, Mathias Humbert, and Jean-Pierre Hubaux. Quantifying interdependent privacy risks with location data. *IEEE Transactions on Mobile Computing*, 16(3):829–842, 2016.

[133] Leonard J Paas and Meike Morren. Please do not answer if you are reading this: Respondent attention in online panels. *Marketing Letters*, 29(1):13–21, 2018.

[134] F. Paci, A. Squicciarini, and N. Zannone. Survey on access control for community-centered collaborative systems. *ACM Computing Surveys*, 51(1), 2018.

[135] Sampo V Paunonen and Douglas N Jackson. What is beyond the big five? plenty! *Journal of personality*, 68(5):821–835, 2000.

[136] Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti. Reputation as a sufficient condition for data quality on amazon mechanical turk. *Behavior research methods*, 46(4):1023–1031, 2014.

[137] Sandra Petronio. Brief status report on communication privacy management theory. *Journal of Family Communication*, 13(1):6–14, 2013.

[138] Leon Pomeroy. *The new science of axiological psychology*, volume 169. Rodopi, 2005.

[139] Antonio Rago, Oana Cocarascu, Christos Bechlivanidis, David Lagnado, and Francesca Toni. Argumentative explanations for interactive recommendations. *Artificial Intelligence*, 296:103506, 2021.

[140] Iyad Rahwan and Guillermo R Simari. *Argumentation in artificial intelligence*, volume 47. Springer, 2009.

[141] S. Rajtmajer, A. Squicciarini, C. Griffin, S. Karumanchi, and A. Tyagi. Constrained social-energy minimization for multi-party sharing in online social networks. In *Proceedings of the International Conference on Autonomous Agents & Multiagent Systems (AAMAS)*, pages 680–688, 2016.

[142] S. Rajtmajer, A. Squicciarini, J. Such, J. Semonsen, and A. Belmonte. An ultimatum game model for the evolution of privacy in jointly managed content. In *GAMESEC*, pages 112–130. Springer, 2017.

[143] Kopo Marvin Ramokapane, Awais Rashid, and Jose Miguel Such. "i feel stupid i can't delete": A study of users' cloud deletion practices and coping strategies. In *Thirteenth symposium on usable privacy and security (SOUPS 2017)*, pages 241–256, 2017.

[144] Kai Rannenberg, Denis Royer, and André Deuker. *The future of identity in the information society: Challenges and opportunities.* Springer Science & Business Media, 2009.

[145] Arunee Ratikan and Mikifumi Shikida. Privacy protection based privacy conflict detection and solution in online social networks. In *International Conference on Human Aspects of Information Security, Privacy, and Trust*, pages 433–445. Springer, 2014.

[146] Delphine Reinhardt, Franziska Engelmann, and Matthias Hollick. Can i help you setting your privacy? a survey-based exploration of users' attitudes towards privacy suggestions. In *Proceedings of the 13th International Conference on Advances in Mobile Computing and Multimedia*, pages 347–356, 2015.

[147] Sonia Roccas, Lilach Sagiv, Shalom H Schwartz, and Ariel Knafo. The big five personality factors and personal values. *Personality and social psychology bulletin*, 28(6):789–801, 2002.

[148] M. Rokeach. *The nature of human values.* Free press, 1973.

[149] F. Rossi, K.B. Venable, and T. Walsh. A Short Introduction to Preferences: Between Artificial Intelligence and Social Choice. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 5(4):1–102, 2011.

[150] Ramon Ruiz-Dolz, José Alemany, Stella Heras, and Ana García-Fornes. Automatic generation of explanations to prevent privacy violations. 2019.

[151] Ramon Ruiz-Dolz, Stella Heras, José Alemany, and Ana García-Fornes. Towards an argumentation system for assisting users with privacy management in online social networks. In *CMNA@ PERSUASIVE*, pages 17–28, 2019.

[152] Ricardo Santos, Goreti Marreiros, Carlos Ramos, José Neves, and José Bulas-Cruz. Personality, emotion, and mood in agent-based group decision making. *IEEE Annals of the History of Computing*, 26(06):58–66, 2011.

[153] Ştefan Sarkadi, Alex Rutherford, Peter McBurney, Simon Parsons, and Iyad Rahwan. The evolution of deception. *Royal Society open science*, 8(9):201032.

[154] S. H. Schwartz. An overview of the Schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1):11, 2012.

[155] Shalom H Schwartz. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*, volume 25, pages 1–65. Elsevier, 1992.

[156] Shalom H Schwartz. A proposal for measuring value orientations across nations. *Questionnaire package of the european social survey*, 259(290):261, 2003.

[157] Shalom H Schwartz. Basic human values: Theory, measurement, and applications. *Revue française de sociologie*, 47(4):929, 2007.

[158] Marc Serramia, Maite López-Sánchez, Juan A Rodríguez-Aguilar, Javier Morales, Michael Wooldridge, and Carlos Ansotegui. Exploiting moral values to choose the right norms. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 264–270, 2018.

[159] Qurat-ul-ain Shaheen, Alice Toniolo, and Juliana K. F. Bowles. Dialogue games for explaining medication choices. In Víctor Gutiérrez-Basulto, Tomáš Kliegr, Ahmet Soylu, Martin Giese, and Dumitru Roman, editors, *Rules and Reasoning*, pages 97–111, Cham, 2020. Springer International Publishing.

[160] Mohamed Shehab and Hakim Touati. Semi-supervised policy recommendation for online social networks. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 360–367. IEEE, 2012.

[161] Nisha P Shetty, Balachandra Muniyal, and Saleh Mowla. Policy resolution of shared data in online social networks. *International Journal of Electrical & Computer Engineering (2088-8708)*, 10, 2020.

[162] Carles Sierra, Nardine Osman, Pablo Noriega, Jordi Sabater-Mir, and Antoni Perelló. Value alignment: a formal approach. *arXiv preprint arXiv:2110.09240*, 2021.

[163] Karl Sigmund, Hannelore De Silva, Arne Traulsen, and Christoph Hauert. Social learning promotes institutions for governing the commons. *Nature*, 466(7308):861–863, 2010.

[164] John Maynard Smith. *Evolution and the Theory of Games.* Cambridge university press, 1982.

[165] Nate Soares. The value learning problem.

[166] A. Squicciarini, M. Shehab, and F. Paci. Collective privacy management in social networks. In *WWW*, pages 521–530. ACM, 2009.

[167] A. Squicciarini, S. Sundareswaran, D. Lin, and J. Wede. A3p: adaptive policy prediction for shared images over popular content sharing sites. In *HT*, pages 261–270. ACM, 2011.

[168] Anna Squicciarini, Cornelia Caragea, and Rahul Balakavi. Toward automated online photo privacy. *ACM Transactions on the Web (TWEB)*, 11(1):1–29, 2017.

[169] Anna Cinzia Squicciarini, Dan Lin, Smitha Sundareswaran, and Joshua Wede. Privacy policy inference of user-uploaded images on content sharing sites. *IEEE transactions on knowledge and data engineering*, 27(1):193–206, 2014.

[170] Ramya Srinivasan and Ajay Chander. Explanation perspectives from the cognitive sciences — A survey. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 4812–4818, 2021.

[171] Mark Stamp. *Information security: principles and practice.* John Wiley & Sons, 2011.

[172] Ilia Stepin, Jose M Alonso, Alejandro Catala, and Martín Pereira-Fariña. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, 9:11974–12001, 2021.

[173] J. Such and N. Criado. Resolving Multi-Party Privacy Conflicts in Social Media. *IEEE TKDE*, 28(7):1851–1863, 2016.

[174] J. Such and N. Criado. Multiparty privacy in social media. *Communications of the ACM*, 61(8):74–81, 2018.

[175] J. Such, J. Porter, S. Preibusch, and A. Joinson. Photo privacy conflicts in social media: a large-scale empirical study. In *CHI*, pages 3821–3832. ACM, 2017.

[176] J. Such and M. Rovatsos. Privacy policy negotiation in social media. *ACM TAAS*, 11(1):1–29, 2016.

[177] Jose Such. Privacy and autonomous systems. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4761–4767, 2017.

[178] Roykrong Sukkerd, Reid Simmons, and David Garlan. Toward explainable multi-objective probabilistic planning. In *2018 IEEE/ACM 4th International Workshop on Software Engineering for Smart Cyber-Physical Systems (SEsCPS)*, pages 19–25. IEEE, 2018.

[179] Roykrong Sukkerd, Reid Simmons, and David Garlan. Tradeoff-focused contrastive explanation for mdp planning. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1041–1048. IEEE, 2020.

[180] Gareth Terry, Nikki Hayfield, Victoria Clarke, and Virginia Braun. Thematic analysis. *The Sage handbook of qualitative research in psychology*, pages 17–37, 2017.

[181] C Tessier, L Chaudron, and H-J Müller. *Conflicting agents: conflict management in multi-agent systems*, volume 1. Springer Science & Business Media, 2006.

[182] K. Thomas, C. Grier, and D. Nicol. Unfriendly: Multi-party privacy risks in social networks. In *PET*, pages 236–252. Springer, 2010.

[183] Suzanne Tolmeijer, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein. Implementations in machine ethics: A survey. *ACM Computing Surveys (CSUR)*, 53(6):1–38, 2020.

[184] Arne Traulsen, Christoph Hauert, Hannelore De Silva, Martin A. Nowak, and Karl Sigmund. Exploration dynamics in evolutionary games. *Proceedings of the National Academy of Sciences*, 106(3):709–712, 2009.

[185] Onuralp Ulusoy and Pınar Yolum. Agents for preserving privacy: Learning and decision making collaboratively. In *Multi-Agent Systems and Agreement Technologies*, pages 116–131. Springer, 2020.

[186] Onuralp Ulusoy and Pinar Yolum. Norm-based access control. In *Proceedings of the 25th ACM Symposium on Access Control Models and Technologies*, pages 35–46, 2020.

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

[188] Steven Umbrello and Angelo Frank De Bellis. A value-sensitive design approach to intelligent agents. *Artificial Intelligence Safety and Security (2018) CRC Press (. ed) Roman Yampolskiy*, 2018.

[189] Jeroen Van den Hoven. Ict and value sensitive design. In *The information society: Innovation, legitimacy, ethics and democracy in honor of Professor Jacques Berleur SJ*, pages 67–72. Springer, 2007.

[190] Thomas L van der Weide. *Arguing to motivate decisions*. PhD thesis, Utrecht University, 2011.

[191] Thomas L van der Weide, Frank Dignum, J-J Ch Meyer, Henry Prakken, and Gerard AW Vreeswijk. Practical reasoning using values. In *International Workshop on Argumentation in Multi-Agent Systems*, pages 79–93. Springer, 2009.

[192] Alexandros Vassiliades, Nick Bassiliades, and Theodore Patkos. Argumentation and explainable artificial intelligence: a survey. *The Knowledge Engineering Review*, 36, 2021.

[193] Bart Verheij. Formalizing value-guided argumentation for ethical systems design. *Artificial Intelligence and Law*, 24(4):387–407, 2016.

[194] Sami Vihavainen, Airi Lampinen, Antti Oulasvirta, Suvi Silfverberg, and Asko Lehmuskallio. The clash between privacy and automation in social media. *Pervasive Computing, IEEE*, 13(1):56–63, 2014.

[195] N. Vishwamitra, Y. Li, K. Wang, H. Hu, K. Caine, and G.J. Ahn. Towards pii-based multiparty access control for photo sharing in online social networks. In *SACMAT*, pages 155–166. ACM, 2017.

[196] B. Viswanath, A. Mislove, M. Cha, and K.P. Gummadi. On the evolution of user interaction in facebook. In *Proceedings of the 2nd ACM workshop on Online social networks*, pages 37–42. ACM, 2009.

[197] K. Walker and E. Sleath. A systematic review of the current knowledge regarding revenge pornography and non-consensual sharing of sexually explicit media. *Aggression and violent behavior*, 36:9–24, 2017.

[198] Douglas N Walton. *Argumentation schemes for presumptive reasoning.* Psychology Press, 1996.

[199] Yang Wang, Gregory Norcie, Saranga Komanduri, Alessandro Acquisti, Pedro Giovanni Leon, and Lorrie Faith Cranor. "I regretted the minute I pressed share" A qualitative study of regrets on Facebook. In *Proceedings of the seventh symposium on usable privacy and security*, pages 1–16, 2011.

[200] Maya Wardeh, Adam Wyner, Katie Atkinson, and Trevor Bench-Capon. Argumentation based tools for policy-making. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*, pages 249–250, 2013.

[201] J. Watson, H.R. Lipford, and A. Besmer. Mapping user preference to privacy default settings. *ACM TOCHI*, 22(6):32, 2015.

[202] Alan F Westin. Privacy and freedom. *Washington and Lee Law Review*, 25(1):166, 1968.

[203] Norbert Wiener. Some moral and technical consequences of automation. *Science*, 131(3410):1355–1358, 1960.

[204] Michael Winikoff. Towards trusting autonomous systems. In *International Workshop on Engineering Multi-Agent Systems*, pages 3–20. Springer, 2017.

[205] Michael Winikoff, Virginia Dignum, and Frank Dignum. Why bad coffee? Explaining agent plans with valuings. In *International Conference on Computer Safety, Reliability, and Security*, pages 521–534. Springer, 2018.

[206] Till Winkler and Sarah Spiekermann. Twenty years of value sensitive design: a review of methodological practices in vsd projects. *Ethics and Information Technology*, pages 1–5, 2018.

[207] P. Wisniewski, H. Lipford, and D. Wilson. Fighting for my space: Coping mechanisms for sns boundary regulation. In *CHI*, pages 609–618. ACM, 2012.

[208] Michael Wooldridge and Wiebe Van Der Hoek. On obligations and normative ability: Towards a logical analysis of the social contract. *Journal of Applied Logic*, 3(3-4):396–420, 2005.

[209] Lei Xu, Chunxiao Jiang, Nengqiang He, Zhu Han, and Abderrahim Benslimane. Trust-based collaborative privacy management in online social networks. *IEEE Transactions on Information Forensics and Security*, 14(1):48–60, 2018.

[210] Sergej Zerr, Stefan Siersdorfer, Jonathon Hare, and Elena Demidova. Privacy-aware image classification and search. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 35–44, 2012.

[211] Haoti Zhong, Anna Squicciarini, and David Miller. Toward automated multiparty privacy conflict detection. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1811–1814, 2018.

[212] John Zoshak and Kristin Dew. Beyond Kant and Bentham: How ethical

theories are being used in artificial moral agents. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.