Autonomous Agents and Multi-Agent Systems (2022) 36:10
https://doi.org/10.1007/s10458-021-09543-5

An explainable assistant for multiuser privacy

Francesca Mosca[1] · **Jose Such**[1]

Accepted: 24 December 2021

Abstract
Multiuser Privacy (MP) concerns the protection of personal information in situations where such information is co-owned by multiple users. MP is particularly problematic in collaborative platforms such as online social networks (OSN). In fact, too often OSN users experience privacy violations due to conflicts generated by other users sharing content that involves them without their permission. Previous studies show that in most cases MP conflicts could be avoided, and are mainly due to the difficulty for the uploader to select appropriate sharing policies. For this reason, we present ELVIRA, the first fully explainable personal assistant that collaborates with other ELVIRA agents to identify the optimal sharing policy for a collectively owned content. An extensive evaluation of this agent through software simulations and two user studies suggests that ELVIRA, thanks to its properties of being role-agnostic, adaptive, explainable and both utility- and value-driven, would be more successful at supporting MP than other approaches presented in the literature in terms of (i) trade-off between generated utility and promotion of moral values, and (ii) users' satisfaction of the explained recommended output.

Keywords Explainable agent · Multiuser privacy · Agent-based simulations · User study

1 Introduction

Privacy in Online Social Networks (OSNs) depends on not just what one user reveals about herself but also on what others reveal about her [1]. OSN platforms have proved to be particularly unsuitable to manage multi-user privacy in a satisfying way for the users [1–3]. One specific privacy problem is that, whenever the content to be shared involves more than a person, the privacy policies should be understood and approved by all the users involved. If this does not happen, a *multi-user privacy conflict* (MPC) is likely to occur. MPCs are frequent, and have been suffered by the majority of OSN users [4, 5]. A common example in the literature is the case of a picture representing a group of friends, where each one of them would assign different degrees of publicity/privacy to the picture on the OSN.

✉ Francesca Mosca
francesca.mosca@kcl.ac.uk

Jose Such
jose.such@kcl.ac.uk

[1] Department of Informatics, King's College London, London, England

Published online: 15 January 2022

Currently, most OSN platforms lack built-in mechanisms that allow the users to discuss and agree on a policy in advance [4, 6], and the responsibility of selecting one is generally left solely to the uploader. The other involved users, if unhappy with the uploader's choice, can only resort to reparative solutions, such as untagging or asking to remove the content. These solutions are not considered to be satisfying [5, 7]: the damage may be immediately perceived and the content is not guaranteed to disappear.

Recently, models for better supporting users to deal with MPCs have been proposed in the related literature. However, all these models lack one or more of crucial properties to successfully supporting MPC resolution, such as being able to explain the solution achieved and considering users' values. On the one hand, one of the requirements for collaborative access control models to successfully address MPC is that they can explain users why the particular solution was reach so that users would ultimately be able to understand the solutions suggested by the models [8]. Hence, the capability of models to solve MPC to provide an explanation of their processes in a human comprehensible way [9] is highly desirable. On the other hand, users consider different values when they share information in OSN. For instance, empirical evidence suggests that some users go beyond their perceived personal gain (or utility) and consider the consequences of their actions on others in terms of MPCs [5].

Although some MPCs occur in adversarial settings (e.g., revenge porn [10]), the vast majority of MPCs happen in *non-adversarial settings* [5], where it is simply too difficult for uploaders to identify the optimal sharing policy for an item that involves other co-owners [2, 4, 11]. For this reason, we present in this paper ELVIRA, an explainable agent-based model that aims to support OSNs users to manage multiuser privacy and collaboratively identify a sharing policy that would solve the MPC with everyone's satisfaction.

1.1 Requirements for MPC solutions

Informed by previous literature on online privacy and autonomous systems [12–14] and, more specifically, by theoretical studies and empirical evidence on multiuser privacy on OSNs [4, 5, 8, 11], a model should match the following requirements in order to support the collaborative resolution of MPCs [15, 16].

First, *role-agnosticism* (RA), i.e., models 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 only considering the perspective of one user, who tends to be the uploader [4].

Second, *adaptability* (AD), i.e., a model should behave differently depending on the users' subjective preferences, because different individuals manage privacy in different ways depending on the context [12, 17].

Third, *utility-driven* (UD), i.e. models 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 [14]; and negatively experiencing privacy violations [4, 11].

Fourth, *value-driven* (VD), i.e., models should consider moral values, because empirical evidence suggests that users do so in order to collaboratively solve MPCs [5]. For instance, some users go beyond their perceived personal gain (or utility) to consider the consequences of their actions on others, or self-transcend to accommodate others' preferences.

Last but not least, *explainability* (EX), i.e., the capability of a model to provide an explanation of its processes [9], is desirable in any autonomous system for reasons of trustworthiness [18], accountability [19], and responsibility [20]; this is particularly crucial in the MPC context for allowing users to know why a solution is suggested and its effects [8], and to align the differences between uploaders and co-owners [5].

While this list may be not exhaustive, we show in this paper that the combination of all these properties by design is crucial to adequately support multiuser privacy management in OSN.

1.2 Contributions

In this paper[1] we report three main contributions to the resolution of MPCs in OSNs:

1. we define ELVIRA, an explainable agent architecture that is both utility and value-driven and can support OSNs users when collaboratively managing multiparty privacy;
2. we formally define the explainable layer of ELVIRA and we evaluate it through a user study, gathering relevant insights on the most proper design of explanations in the MPCs context;
3. we show through software simulations and a user study that ELVIRA generates the best recommendations to solve MPCs in OSNs, in terms of utility-value trade-off and user's satisfaction, when compared with other state-of-the-art approaches.

1.3 Organisation

The rest of this paper is organised as follows. Section 7 provides an overview of state-of-the-art approaches to manage individual and collective privacy in OSNs. Section 2 introduces the concepts of utility and values in the MPC scenario, with the necessary definitions and notation. Section 3 details how the ELVIRA agents collaboratively identify the optimal solution to MPC by performing practical reasoning. Section 4 describes the design and the evaluation through a user study of the explanations that ELVIRA autonomously generates. Sections 5 and 6 present a comparative evaluation of ELVIRA with other state-of-the-art approaches, respectively through software simulations and a user study. Section 8 summarises and discusses the contributions, and Sect. 9 concludes the paper with final remarks. Finally, in "Appendix A" we report a summary of the main symbols and notations that we have used across the paper.

[1] The work reported in this paper substantially and significantly extends the model and the results published in [21]. In particular, we: (i) extended our analysis of related work, (ii) included a new user study to evaluate different types of explanations, (iii) expanded the evaluation of the model by performing new simulations studying the effect of all model parameters on the model behaviour, and (iv) included a completely novel, qualitative (thematic) analysis on the previously presented user study data, which provides new, valuable insights that allowed us to substantially and complementarily extend the discussion of the results.

2 Preliminaries

We represent a 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 [22].

Among other activities, users can engage with the network by sharing online content[2] 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 [5]: 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.

Definition 1 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 [23], but noting that this is equivalent and can be translated to and back from the group-based access control models used in OSN platforms [24], we define a *sharing policy* as follows:

Definition 2 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.

We assume that every user has a *preferred sharing policy* for each content they are involved in (i.e., they *own*), and that it can be elicited automatically (e.g. see Sect. 7.1). We 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, i.e., the *co-owners* Ag of the item, have contradictory preferred sharing policies which lead to different preferred audiences.

Definition 3 An **MPC** regarding an item $x \in X$ co-owned by users $k, j \in Ag$, 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 4 When considered from all the involved users' point 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:

[2] In this paper we mostly focus on photographic content, but similar solutions can be applied also to other types of content.

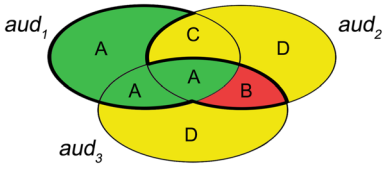


Fig. 1 MPC between 3 users, a possible solution aud' (represented with bold borders), and the $\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}$ sets for user 1

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

In the remaining part of the paper, we will refer to *candidate solutions* for an MPC as collective audiences.

Furthermore, we 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 [25]. In fact, empirical evidence [5] suggests sharing modified content, even if not completely true to the original [26], 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.

Users are known to benefit from sharing in social media [14], 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 [5].

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, we 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, we first define the following sets with respect to the user k and her preferred audience $aud_{sp_k,k}$, considering the collective audience aud' as a potential solution to a MPC where k is involved (see Fig. 1 for a graphical representation), then the appreciation function capturing the tendencies to under/over-share, and finally the utility function.

Definition 5 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., $\mathcal{D} = \bigcup_{l \neq k} aud_{sp_l,l} \setminus aud'$.

Table 1 Variation of the individual utility for item x, considering audience sets, appreciation and mode of sharing

$\Delta utility$		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

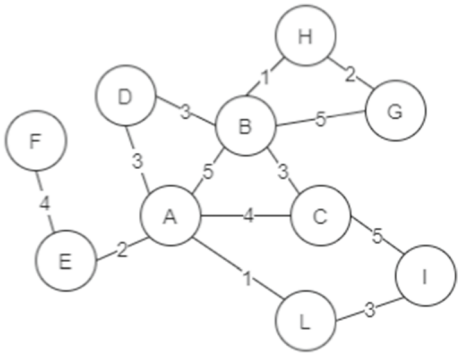


Fig. 2 The simplified online social network discussed in the example

Definition 6 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 undershare.

Definition 7 Given an audience aud, its **utility** for user k is:

$$u_{k,aud} = \sum_{j \in \mathcal{A}} \frac{i_j}{d_j} - \alpha \sum_{j \in \mathcal{C}} \frac{i_j}{d_j} + app(x)\left(\sum_{j \in \mathcal{B}} \frac{i_j}{d_j} + \beta \sum_{j \in \mathcal{D}} \frac{i_j}{d_j} \right). \quad (1)$$

For the sake of clarity, Table 1 shows the individual contributions of each audience set to the variation in utility. Note that the components for the sets C and 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 Sect. 5) that the optimal choice of these two parameters does not seem critical, because we did not find any significant impact on the differences between individual utilities achieved under different values for the parameters.

Example Let us consider the simplified OSN in Fig. 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 preferred audiences do not coincide. Furthermore, Alice and Charlie prefer to eventually undershare the content x ($app_A(x) = app_C(x) = -1$), while Bob prefers to overshare it ($app_B(x) = +1$).

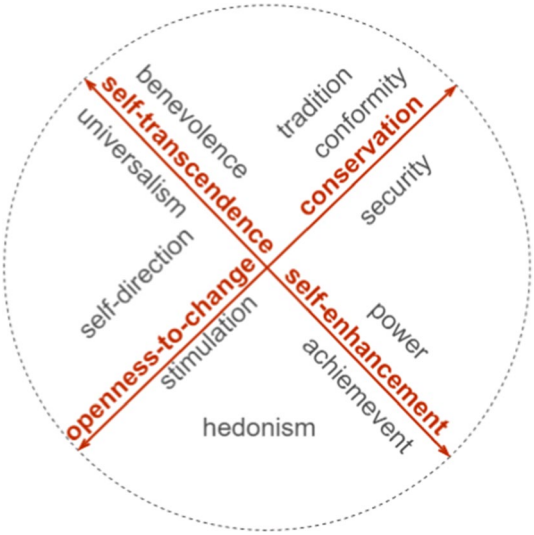


Fig. 3 The Schwartz values and hypervalues arranged in a circular structure

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.

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

Then, Alice, Bob and Charlie would perceive the following variation in utility:

$$u_{A,aud_{sp'}} = \sum_{j \in \{B,C,D,G,I\}} \frac{i_j}{d_j} - \alpha \sum_{j \in \{E,F\}} \frac{i_j}{d_j}$$

$$u_{B,aud_{sp'}} = \sum_{j \in \{A,C,D,G\}} \frac{i_j}{d_j} + 1 \cdot \left(\sum_{j \in \{I\}} \frac{i_j}{d_j} + \beta \sum_{j \in \{E,F\}} \frac{i_j}{d_j} \right)$$

$$u_{C,aud_{sp'}} = \sum_{j \in \{A,B,G,I\}} \frac{i_j}{d_j} - 1 \cdot \left(\sum_{j \in \{D\}} \frac{i_j}{d_j} + \beta \sum_{j \in \{E,F\}} \frac{i_j}{d_j} \right)$$

2.1 Schwartz basic values

The *theory of basic values* by Schwartz [27] is one of the most well-known and established socio-cultural theories of human values backed by strong empirical evidence. In

Table 2 Details of promotion and demotion of the values for a user, comparing different sharing options with own preference, and corresponding behaviours

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	

this theory, values are socially desirable concepts that represent the mental goals which drive human behaviour and influence any people's decision.

As depicted in Fig. 3, Schwartz identifies ten main values and orders them in a circular way, considering reciprocal similarities and influences. Two dimensions emerge overall and define four directions that represent higher order values, or *hypervalues*, which pull apart while influencing the human behaviours. On one axis, *openness to change* (OTC) is opposed of *conservation* (CO), representing dynamic and independent ways of acting versus conservative and self-restraining attitudes. On the other axis, *self-transcendence* (ST) reflects tolerant and altruistic behaviours in opposition to *self-enhancement* (SE), that characterises authoritarian and image-conscious conducts.

Schwartz designed and broadly validated some tools that allow to elicit the individual preferred order over the values (see [27, 28] and some more details in Sect. 4.2.1). Such tools are more reliable than the ones offered by other value theories (see for instance [29]), which do not provide an overall value architecture or direct insights on the behavioural impact of the values. Furthermore, the individual preferred order over the values is proven to be relatively stable over the lifetime [30]. This suggests that it should not be necessary to elicit it from the users for every MPC, allowing the model we present in this paper to be scalable and applicable in the real world. In fact, if this model was to be applied in the real world, e.g., as a service provided by the OSN itself or by a third-party application, it may be sufficient to elicit the user's preferred values just once at the time of signing up and then to confirm or update them at regular intervals (e.g., after a few years of use). A real-world application based on our proposal may then guarantee that data would be used exclusively for the purpose of recommending appropriate actions when managing multi-user privacy and would not be shared with any other entity.

We model behaviours in a MPC along the four main directions: OTC, meant as appreciating compromises which differ from anyone's initial preference; CO, meant as the effort of preserving individual and social security; ST, meant as making the others happy; and SE, meant as getting one's own way, e.g., by maintaining or increasing one's own utility. The selection of any audience as a solution promotes or demotes these values for each involved user as shown in Table 2. In the remaining part of the paper, we refer to these value-directions as $\mathcal{V}$.

Table 3 Users' preferences in the MPC discussed in the running example

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

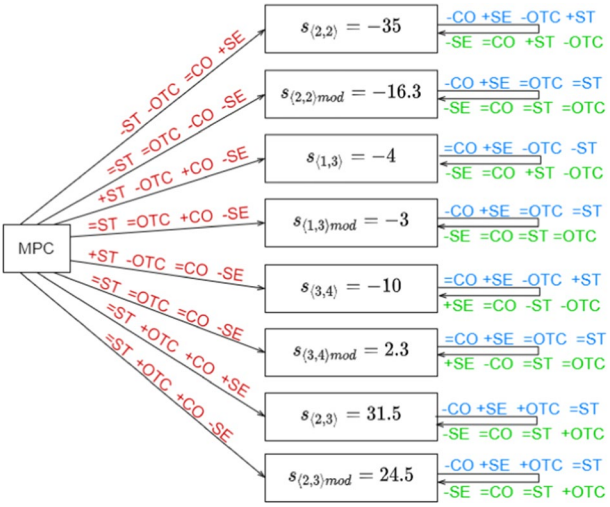


Fig. 4 The AATS+V representing the PR performed in the example ($aud_f = \langle 2,3 \rangle$)

Running example Considering the same MPC as in the previous example (see a summary of the user's preferences in Table 3), let us discuss how Alice may 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 simplicity, we 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. For a complete view of the value promotion for all the involved users, see later Fig. 4.

3 ELVIRA

We now describe in detail ELVIRA, an agent that supports the collaborative resolution of MPCs. The design of ELVIRA is such that it complies with all the desired requirements described in Sect. 1: *explainability* is given by the practical reasoning approach (Sect. 3.1) and the process to describe MPCs and their recommended solution (Sect. 4), which are evaluated in Sect. 4.2; *role-agnosticism* and *adaptability* are guaranteed by its formal properties (Sect. 3.2); and, finally, both individual utility and moral values are explicitly

considered to compute the solution to the MPC as described below (and evaluated in Sects. 5 and 6).

We assume that there is one ELVIRA agent representing each user involved in an MPC, and that they will all be working together collaboratively to resolve the MPC, as we focus in this paper on the majority of MPCs which happen in non-adversarial settings [2, 4, 11]. 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. For clarity and in the interest of space, we present ELVIRA from the perspective of the uploader agent, which considers everyone's individual preferences in collaboration with the co-owner agents, and identifies a solution for the MPC.

In order to solve an MPC over one item,[3] the uploader can offer to the co-owners an audience *aud*, chosen *as-it-is* or *modified*, from a finite set of options $\mathbb{A}$ which includes the n collective audiences $aud_1, \ldots, aud_n$ 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 $\text{card}(\mathbb{A}) \leq 2(n+1)$ possible solutions to the MPC ($\leq$ because two or more co-owners may have the same preferred audience). In the remaining part of this section, 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 4.

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

$$s_{k,aud} = \begin{cases} -u_{k,aud} \cdot v_{k,aud} & \text{if } u_{k,aud} < 0 \text{ and } v_{k,aud} < 0 \\ u_{k,aud} \cdot v_{k,aud} & \text{otherwise} \end{cases} \tag{2}$$

The utility $u_{k,aud}$ is computed as in Eq. (1); the value promotion $v_{k,aud}$ takes as input an order o over $\mathcal{V}$, so that:

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

where $I = \text{card}(\mathcal{V}) + 1$, and $prom(o_i) = 1$ if the i-th preferred value is promoted by selecting *aud*, $prom(o_i) = -1$ if the i-th preferred value is demoted, and $prom(o_i) = 0$ otherwise. In Eq. (2) we multiply u and v for assigning equal weight to utility and values regardless of their range. Then, all the co-owners share their individual scores with the uploader, who aggregates them in an *overall score* for each audience $aud \in \mathbb{A}$:

$$s_{aud} = \sum_{k \in Ag} s_{k,aud}. \tag{3}$$

3.1 Computing the solution

In this section we describe how the ELVIRA uploader agent computes the solution to an MPC based on argumentation techniques. By completing the abductive reasoning process

[3] Note that we discuss MPCs over one item for simplicity but without loss of generality, as one could define a preferred audience over a collection of items too. The fundamental way in which ELVIRA works would be the same.

that we describe below, not only ELVIRA uploader 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 [9]. We detail how ELVIRA uses this information to generate the explanations in Sect. 4.

We present ELVIRA's abductive reasoning process as an adaptation of the work on practical reasoning by Atkinson and Bench-Capon [31, 32]. First, we consider that an agent can propose, attack and defend justifications for a given action by relying on an argument scheme (AS) and its associated critical questions (CQs) [31]. AS can be expressed as: "I should offer the audience *aud*′, that will be accepted by the co-owners, 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) [31]: (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 AS; (3) it considers objections (the 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*.

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.). We perform this task by building an Action-Based Alternating Transition Systems with Values (AATS+V) [31]. 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.[4] We adapt Atkinson's definition of an AATS+V [31] to MPCs as follows:

Definition 8 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 $k = 1 \dots n$, where:

- $Q = \{\text{conflict}, \text{agreement}_{aud} \quad \forall aud \in \mathbb{A}\}$ is a finite, non-empty set of states;
- $q_0 = \text{conflict}$ is the initial state;
- $Ag = \{up_1, co_2, \dots, co_n\}$ is the set of agents involved in the MPC, with the roles of uploader or co-owners;
- $Ac_1 = \{\text{offer}_{aud} \quad \forall aud \in \mathbb{A}\}$ are the actions available to the agent up_1;
- $Ac_k = \{\text{accept}_{k,aud}, \text{reject}_{k,aud} \quad \forall aud \in \mathbb{A}\}$ are the actions available to the agent co_k, for $k = 2 \dots n$;
- $\rho : Ac_{Ag} \rightarrow 2^Q$ 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;

[4] As in [32], we assume the offer and the response to be "simultaneous" actions, despite their sequentiality.

Table 4 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

J_{Ag}	τ
$j_{1-8} = \langle \text{offer}_{aud_i}, \text{reject}_{2,aud_i}, \text{reject}_{3,aud_i} \rangle$	$\tau(\text{conflict}, j_{1-8}) = \text{conflict}$
$j_{9-16} = \langle \text{offer}_{aud_i}, \text{accept}_{2,aud_i}, \text{reject}_{3,aud_i} \rangle$	$\tau(\text{conflict}, j_{9-16}) = \text{conflict}$
$j_{17-24} = \langle \text{offer}_{aud_i}, \text{reject}_{2,aud_i}, \text{accept}_{3,aud_i} \rangle$	$\tau(\text{conflict}, j_{17-24}) = \text{conflict}$
$j_{25-32} = \langle \text{offer}_{aud_i}, \text{accept}_{2,aud_i}, \text{accept}_{3,aud_i} \rangle$	$\tau(\text{conflict}, j_{25-32}) = \text{agreement}_{aud_i}$

- $S = \{0, s_{aud} \quad \forall aud \in \mathbb{A}\}$ 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 2).

Epistemic Stage The epistemic stage consists of determining what the agent believes about the current situation, given the previous problem formulation. As we mentioned earlier, based on empirical evidence [5], the ELVIRA agents have a collaborative behaviour. From this underlying assumption we can further imply two epistemic assumptions:

- *EA1*: all agents share the same interpretation of the world and have the same knowledge;
- *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 ($s_{aud'} = \max_{\mathbb{A}} s_{aud}$).

EA1 allows the agent to discard any CQs related to the problem formulation and its truthfulness; EA2 allows the agent to evaluate appropriately the expectations regarding the other agents' actions.

Choice of Action Finally, we 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 AS, 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.

Table 5 Utility, value promotion and score generated by each audience for each user in the example

	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

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

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.

Running Example Considering the same MPC as in the previous examples, let us discuss how the ELVIRA uploader agent, acting on behalf of Alice, performs the practical reasoning process. Figure 4 shows the representation of the AATS+V, Table 4 reports all the available joint actions, and Table 5 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 Eq. (3)].

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 $\text{AS}_{\langle 2,2\rangle}$: "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 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 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, $\text{AS}_{\langle 2,2\rangle}$ is discarded.

The agent proceeds similarly to consider all the other possible desirable outcomes, until it eventually formulates $\text{AS}_{\langle 2,3\rangle}$: "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, $\text{AS}_{\langle 2,3\rangle}$ is accepted and the agent identifies $aud' = \langle 2,3\rangle$ as the solution to the MPC.

3.2 Formal properties

We now formally prove how ELVIRA, presenting some properties such as soundness, completeness, anonimity and neutrality, fulfills 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 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 1 (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 Sect. 3.1), 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 2 (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 3 (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 Eq. (3) and in the critical question CQ2 during the practical reasoning, where the order of aggregation of the considered elements is irrelevant. In fact, in Eq. (3), 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 4 (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.

□

4 Generating explanations

According to [33], 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 [9], 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 Sect. 3.1 we 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. We now describe the steps that led us to the definition of the ELVIRA's social process. First we discuss, from a theoretical point of view, the elements that should be included in the explanation for an MPC solution; then, we suggest some different explanation designs and we evaluate them through a user study.

4.1 Design of the explanations

Both [9, 34] 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. Similarly to what we discussed in [35], in the following we outline how the design of ELVIRA's explanations meets these requirements.

Conflict Description In order to explain the solution to a conflict, it is useful to first provide details about the *detection* and *representation* [36] of the conflict. This fits the necessity for an explanation to present causal attribution [9]: 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 allows 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, we include 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, we 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, we 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 [9] 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 are 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, we 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, we 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 6 we report the details of the information included in each of these two types of explanation.

Note that our decision of what to include in the explanations in this paper is not a limitation of the model: if a dialogue between the user and the agent was developed, the agent would be able to reply to any user's objection regarding the selection of alternative solutions based on our model in Sect. 3.1. This is, in fact, a very interesting follow-up future work.

In the user study which we describe next, we comparatively evaluate these explanation structures with a baseline, namely *no explanation*, where the recommended solution is suggested without motivation after the description of the conflict.

Running Example Still considering the same MPC scenario as before, we present here as an example how would the three explanations 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.

Table 6 Detailed design of the suggested structures for an explanation which includes the conflict description

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}*}.

{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 2. For the contrastive explanation, when the *if*-conditions are verified (which is optional), then the corresponding sentences are added to the explanation

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.

4.2 Evaluation of the explanations

We now present the within-subjects user study[5] that we designed and conducted in order to evaluate the structure of the explanations that ELVIRA can generate: the baseline

[5] For the full specification of the experiment design, including the scenarios and questions presented to participants, the generated explanations and the collected data, see https://osf.io/ngs27/.

explanation (*exp0*), the general explanation (*exp1*), and the contrastive explanation (*exp2*). The results of this study informed the final design of ELVIRA, which we evaluated in another user study against other models suggested in the literature (see Sect. 6). Participants were recruited through Prolific[6] and the study received ethical approval by the Ethical Board of our university.

4.2.1 User study design

We 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 the three alternative outputs from Table 6, that the participants were required to comparatively evaluate.

Values elicitation We relied on the Portrait Value Questionnaire (PVQ) designed by Schwartz [27] to elicit the value preferences of the users. Among the tools suggested by Schwartz, this is the most appropriate for a broad audience and can easily be delivered online. We used the PVQ-21 version, which includes 21 sentences describing behaviours of people and asks users how similar are those people to themselves, and which has been very commonly used in social studies and as part of the European Social Survey [28] since 2002. The output of this part informs the ELVIRA agent about the participants' value preference.

MPCs We followed an immersive scenario approach [37], which was successfully used in previous work in MPCs [24, 38], in order to elicit the participant's behaviour in MPC situations. Each participant was shown three scenarios, each consisting of a photo and a short description, and for each scenario the participant was asked to put herself in the shoes of one of the depicted people and provide the following: (i) their preferred sharing policy[7] 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/under-sharing. 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). The MPC was then presented to the participant together with the three alternative explanation types, listed in a random order. For each participant, the scenarios were selected randomly among six pairs of pictures/descriptions taken from [38], which were representative of different sensitivities (low/high) and relationship types (colleagues, friends and family). 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.

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

[7] For simplicity, we used group-based policies, which, as aforementioned and shown in [24], are equivalent to the policies we used in earlier parts of our paper, and *as-it-is* modality, because they are both (policies and modalities) more familiar and intuitive for users, as that is what they currently see in mainstream online social networks [6].

Table 7 Demographics of participants

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 8 Tests of between-subjects effects

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 9 The satisfaction scale [39]

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

Satisfaction For each MPC that was presented to the participant, we asked about their satisfaction with the alternative explanation types. To measure satisfaction, we used the Satisfaction Scale proposed in [39] (see Table 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, we 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, we employed two well-known methods: attention check questions, and participants' previous performance [40–43]. We recruited

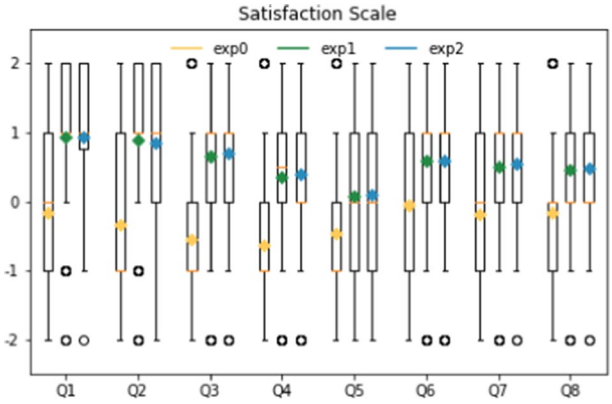


Fig. 5 Satisfaction Scale considering all the conflicts

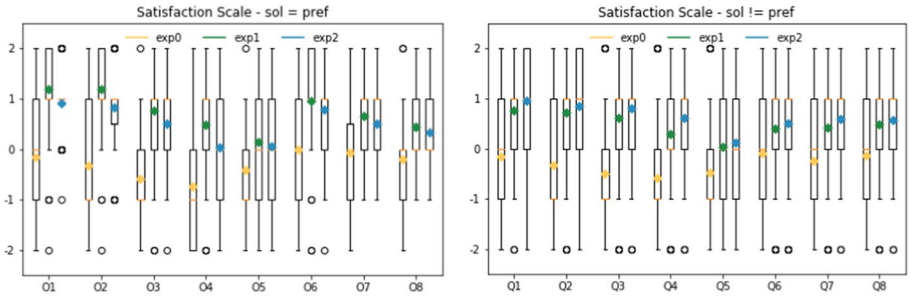
participants from Prolific with at least 100 submissions and an approval rate of 95% according to [41]. Also, during the experiment, the application presented participants with four attention check questions.

4.2.2 User study results

We recruited a total of 68 participants, who were rewarded £3.00 for completing the survey, which took on average 25.3 min (median 20.9 min). We 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. We conducted the analyses on the remaining 64 participants, for a total of 192 MPCs. Table 7 reports the demographic distribution of the participants, including their privacy attitudes measured with the IUIPC scale [44] and social media use.

Overall Satisfaction Figure 5 shows the evaluation through the Satisfaction Scale [39] of the three types of explanations when considering the total of 192 MPCs. We performed a Multivariate ANalysis Of Variance (MANOVA) to compare differences 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 8). More than 80% of the variance is associated with the first four questions, which we conclude being the most important main effects. Furthermore, we are 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 versus Contrastive In order to identify situations where one type of explanations may be preferred over another, we 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 (b) the solution was different from the participant's preference (121 conflicts) (see Fig. 6). Similarly as before,



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

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

Fig. 6 Satisfaction Scale on subsets of the dataset

MANOVA tests showed significantly different distributions in both subsets: (a) $F = 6.73$, p value $< .05$; Wilk's $\Lambda = .625$, partial $\eta^2 = .21$; (b) $F = 7.694$, p value $< .05$; Wilk's $\Lambda = .725$, partial $\eta^2 = .148$. Tukey tests proved that both the general and the contrastive explanations still outperformed significantly the baseline in both subsets (p value $< .05$). We noted 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.". We 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*.

4.2.3 Conclusions of the user study

We summarise the above findings with three intuitions. 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 preference, participants seem to appreciate the description of the positive consequences of selecting that policy, almost as a way of reinforcing its 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. Therefore, we select a *hybrid tailored explanation* structure for the final evaluation of ELVIRA (see Sect. 6), where the agent typically provides a contrastive explanation whenever the solution does not coincide with the user's preference, and a general explanation otherwise.

Finally, based on feedback that we received, we opted for (i) simplifying 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}."); and (ii) labelling the components of the output that ELVIRA generates ("*Conflict:*" followed by the conflict description and "*Solution:*" followed by the hybrid tailored recommendation).

5 Evaluation through software simulations

Having shown above how ELVIRA meets the explainability, role-agnosticism, and adaptability requirements, we now examine experimentally the performance of ELVIRA agents in terms of the utility and adherence to values of the solutions to MPCs they generate. Recall, as explained in Sect. 1, that considering both utility and values to compute a solution to MPC is informed by empirical evidence [5, 14]. In particular, we present a comparative evaluation of ELVIRA (EL) and three other models inspired by the related work approaches (see Sect. 7) that either consider utility, values, or none of them:

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

We compared the performance of EL, UB, VB and FB in two different types of experiments: (i) experiments on synthetic data, which allow us 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 us to compare the models in realistic social networks. In particular, we consider different social networks (in terms of size N and connectivity d), the number of users involved (n) in an MPC, the number of MPCs (T), and the parameters α and β, and varying users' preferred audiences, appreciation for the content to be shared, and values.

To compare the models, we use the individual average variation of utility (*iauc*), normalised over the size of the network, and the individual average of value promotion (*iavc*) per each conflict, generated by each model M:

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

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

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 which the user k gets when selecting the solution suggested by the model M in the conflict t. We also consider the cumulative increment of social utility (*csu*) and of value promotion (*csv*) generated by each model M in order to compare the performance of the different models. They are defined as follows:

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

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

We implemented the models in Python 2.7.10 (*numpy* 1.16.2; *networkx* 2.2) and we ran all our 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] as in [22], where 1 represents a mere acquaintance and 5 a very close relationship. Regarding

the value preference, we 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 [30]. 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.[8] Also, to generate audience aud_f, we 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:

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

We studied also 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.

5.1 Experimental settings

Here we report the settings of our experiments. Experiments I–IV regard synthetic networks, which we generated according to the *scale-free network* model by Barabasi-Albert with preferential attachment [45], where the degrees of the nodes follow a power-law distribution, in order to reproduce scenarios that would resemble as much as possible to real online social networks [46]. Experiment V involves portions of a real social network (Facebook).

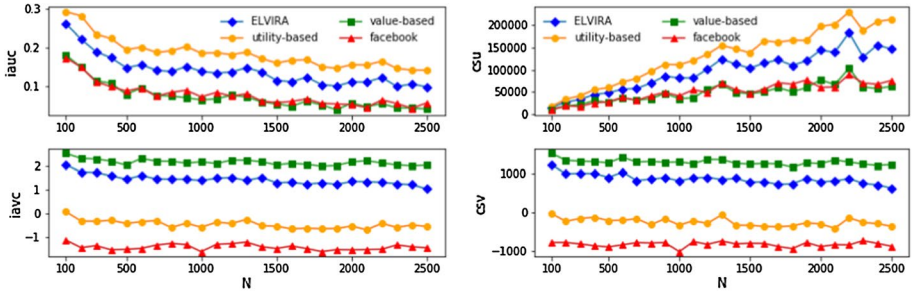
Experiment I In this experiment we studied the performance of EL, UB, VB and FB after solving $T = 300$ conflicts when increasing the size of the network from $N = 100$ up to $N = 2500$ while maintaining $d = 10, n = 3, app = \pm 1, \alpha = 0.9$ and $\beta = 0.1$.

Experiment II In order to see the effect of other parameters in addition to the size of the network, in this experiment, we 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 $app = \pm 1$.

Experiment III In this experiment we 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, we compared the utility generated in Exp. II when selecting randomly only the extreme values of appreciation ($app \in \{+1, -1\}$) with the utility generated when selecting also intermediate values ($app \in [-0.9, -0.45, 0, 0.45, 0.9]$). We maintained all the other settings as in Exp. II.

Experiment IV In this experiment we studied the impact of selecting the audiences *as-it-is* or *modified*, by varying the parameters α and β. We considered $\langle \alpha = 0.9, \beta = 0.1 \rangle$,

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



(a) *iauc* and *iavc* (b) *csu* and *csv*

Fig. 7 Performance of the four models in Experiment I

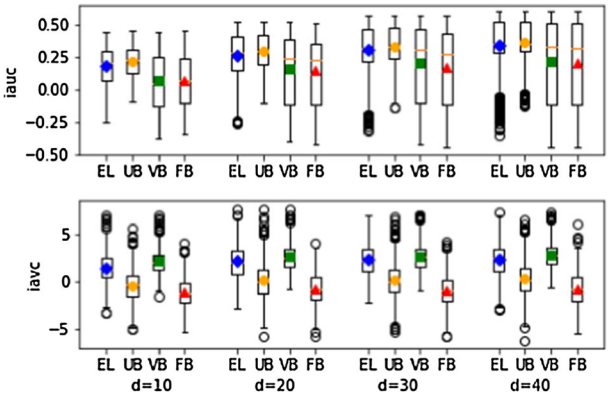


Fig. 8 Performance of the four models in terms of *iauc* and *iavc* when varying d, with $N = 500, n = 5, T = 1000$

$\langle \alpha = 0.5, \beta = 0.5 \rangle$ 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 5) has different influence on the utility. We 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 we used graphs from real portions of Facebook—number of nodes and edges in parenthesis: $G_1 = (769, 16656)$ and $G_2 = (1446, 59589)$ from [47], and $G_3 = (4039, 88234)$ from [48]. Maintaining $\alpha = 0.9$ and $\beta = 0.1$ and $app = \pm 1$, we generated $T = 500$ MPCs among $n = 3$ random users on each graph—considering that, as shown in Exp. 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 [5, 25].

5.2 Experimental results

Here we report the results of the experiments described above.

Experiment I Figure 7a shows the *iauc* and the *iavc* generated by each model after solving $T = 300$ MPCs. Figure 7b 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

Table 10 ELVIRA's performance in Experiment II and III: better (>), worse (<), or not significantly different (n.s.) from the other models; *Marks the significant pairwise t-test with p value < .05

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.	>*	>*	>*	<*	>*

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 we focus only on *iauc* and *iavc*, which we consider more significant to evaluate the performance of the models.

Experiment II Figure 8 shows, as an example of the performance of the four models, the results when varying d and keeping $N = 500$ and $n = 5$. Over a more connected graph, users can in general achieve higher utilities. Table 10 reports an overview of the results and their level of significance. The results show the same similar constant trend as in Fig. 7. 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 very low individual utilities; and EL represents the best utility-value trade-off. By increasing n, we noticed that the distance between the utility generated by EL and UB decreased, while the gap with VB and FB increased. This suggests that EL might reach optimal utilities if increasing further the number of conflicting users.

Experiment III As reported in Table 10, when considering $app \in [-0.9, -0.45, 0, 0.45, 0.9]$, the models performed in the same way as in Experiment II: 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 after $T = 1000$ the *iauc* generated by ELVIRA with $app \in \{-1, +1\}$ and with intermediate values of appreciation, we noticed that the intermediate values of appreciation tended to provide higher utilities, but no significant differences were observed.

Experiment IV We simulated $T = 500$ conflicts for different N, d and n combinations 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 9 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

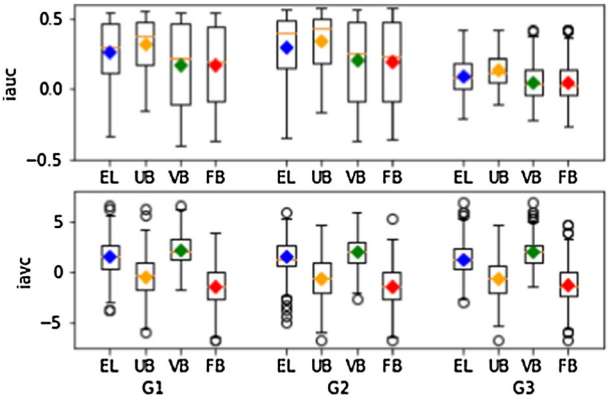


Fig. 9 Comparison of the performance of the four models in terms of *iauc* and *iavc* generated on G_1, G_2 and G_3

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 graph.

5.3 Conclusions of the software simulations

Across all the experiments, we can clearly see that the 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). 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. The Facebook approach selected the solutions with the least generation of utility and the worse value promotion. ELVIRA represented the best utility-value trade-off, by producing utilities very close to UB and value promotion close to VB.

Given the similarity of the models' performance across scenarios and settings, we decided to maintain $n = 3$, $\langle \alpha = 0.9, \beta = 0.1 \rangle$ and the appreciation in a range of possible values for the evaluation through user study that we present in the next section.

6 Evaluation through user study

We now discuss the between-subjects user study[9] that we designed and conducted with a double 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 Sects. 3.1 and 4 allow ELVIRA to *convey* the

[9] For the full specification of the experiment design, including the scenarios and questions presented to participants, and the collected data, see https://osf.io/ngs27/.

Table 11 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

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}

recommendations in a more satisfactory way than existing approaches. Similarly to the software simulations in Sect. 5, we compared the performance of ELVIRA (EL) with three other models inspired by related work approaches: *utility-based* (UB), *value-based* (VB) and *Facebook* (FB). Participants were recruited via Prolific, and the study received ethical approval by the Ethical Board of our university.

6.1 User study design

In order to conduct this experiment, we developed another web application in Python similar to the one whose design is described in Sect. 4.2.1; hence, for the unchanged design details we refer the reader to that section. 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) participants were presented with MPCs automatically generated by our tool, given the recommendations suggested by the model used in the particular treatment, and asked about the acceptability of the recommendations; (ii) after all 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 step to elicit the value preferences of participants through the Schwartz questionnaire PVQ-21 [28] (see Sect. 4.2.1). We now describe the different steps further.

MPCs We followed the same immersive scenario approach as described in Sect. 4.2.1. The application considered the same six scenarios (picture and description,—see [38]) and presented all of them in a random order to each participant. After eliciting sharing preference and appreciation, the application randomly generated the preferences and appreciation of 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 11). The output generated by ELVIRA corresponds to the hybrid tailored explanation described in Sect. 4.2.3. 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 (cf. Sect. 7). 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).

Table 12 Demographics of participants

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%

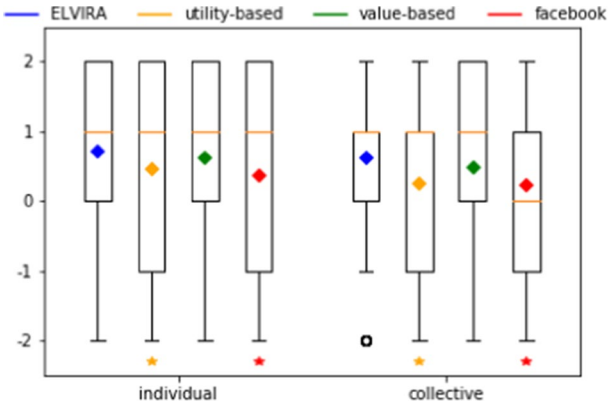


Fig. 10 Individual and collective acceptability of the recommendations presented by each model

Satisfaction After all the MPCs were presented to the participant, and as a final step, we asked about their satisfaction with the model of their treatment 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, we used again the Satisfaction Scale proposed by Hoffman et al. [39] (see Sect. 4.2.1 for details).

Data Quality Measures Similarly to Sect. 4.2.1, in order to maximise data quality, we employed attention check questions and participants' previous performance. We 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.

6.2 User study results

We recruited 470 participants, who were rewarded £2.50 for completing the survey, which took on average 23.71 min (median 20.58 min). We discarded participants who failed at least one attention check question (28.7%), and analysed the remaining 335 participants. Table 12 reports the demographic distribution of the participants, including their privacy attitudes, measured with the IUIPC scale [44], and social media use.

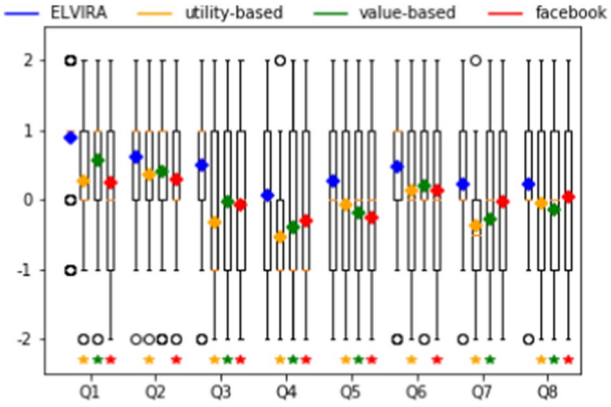


Fig. 11 Evaluation of the outputs provided by each model, according to the Satisfaction Scale [39]

The final split per treatment (recall this was done randomly) was: 85 ELVIRA, 82 utility-based, 85 value-based, and 83 Facebook.

Acceptability of recommendation Figure 10 shows the distribution of individual and collective acceptability for each model (2 = 'Very likely', −2 = 'Very unlikely'). The stars (⋆) on the bottom mark the distributions that are significantly worse 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).

Satisfaction of the output Regarding the quality of the generated output, ELVIRA achieved by far the best performance. Figure 11 shows the distribution of the answers to the Satisfaction Scale (2 = 'Strongly agree', −2 = 'Strongly disagree'), 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, we note ELVIRA's dominant results 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".

6.3 Motivations for accepting a recommendation

When asking the users about the acceptability of each recommendation, we 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 we collected, we 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 :)").

We analysed the remaining *1945 responses* by applying *thematic analysis* (TA) [49], 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. We followed an inductive and semantic approach to TA [49]: starting from the explicit meaning of the data, we 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?", we identified the following main themes. Together with the description of the theme, we 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: *oth-*

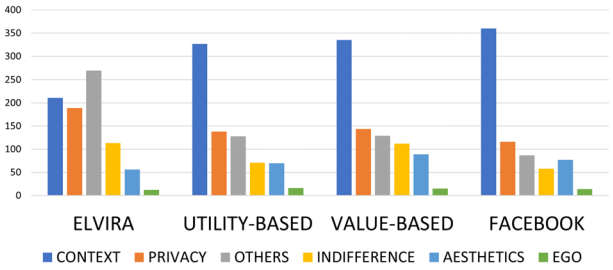


Fig. 12 Themes distribution across the treatments

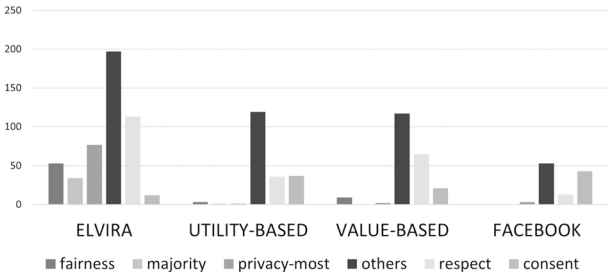


Fig. 13 Codes distribution within the theme "Others"

ers, 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."]
- *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 [5].

Being the thematic analysis purely qualitative and exploratory in nature, we do not draw any confirmed conclusions, but we discuss some interesting trends that have emerged and may be worthy of future confirmatory studies. Figure 12 reports a comparative overview of the themes occurrence in the participants' answers[10] across the treatments. Given that the theme "Others" presents the most diverse distribution across the treatments, we show in Fig. 13 the distribution of the codes that are included in this theme. In any case, the distribution of codes seems to suggest that ELVIRA nudged the participants to be more conscious of the co-owners and more privacy-aware than the other models. 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. Still within the consideration of the others' preferences, the interactions with ELVIRA encouraged the participants to reflect more upon the *fairness* of the recommendation and, more generally, to be more *respectful* of the others' wishes. On the other hand, the other treatments made the participants wonder more often whether *consent* was given by the co-owners. With ELVIRA, the participants were already taking into account what the others would like and how the suggestion recommended was the get considering that.

6.4 Conclusions of the user study

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. 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

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

that are acceptable for users independently of their demographics, their privacy awareness and their OSN experience; (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 Related work

OSNs users can encounter a variety of privacy threats [50] and have reported to be mostly worried by the *insider threat* [51], 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 working on the definition of more usable access control mechanisms to help users better manage their online privacy, independently of their privacy understanding and experience, and prevent inadvertent disclosure on OSNs.

In the following, we present an overview of these mechanisms, focusing in particular on the agent-based models. In fact, autonomous agents such as *privacy personal assistants* [13] have been advocated for helping users make privacy decisions in a variety of contexts. First, we describe some exemplar agent-based models that aim to enhance and protect the individual user's privacy. Then, we focus on multiuser privacy, and discuss to what extent previous work is able to fulfil the requirements introduced in Sect. 1.1.

7.1 Agent-based models to support individual privacy

Plenty of mechanisms have been suggested to help users manage their individual online privacy. Most of them recommend sharing policies based on image features [52–54], some consider also social graphs properties [54–57], similar characteristics among users [58, 59], or the user's sharing history [60–62]. We refer the reader to extensive literature reviews on the topic (such as [50]) and focus next on agent-based approaches to individual privacy management, which have shown some promise, particularly considering that agents could help users, or even pro-actively act on their behalf [13], to protect their privacy.

Kurtan and Yolum [63] 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.

Similarly, in [64] 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 [65, 66] introduce PACMAN 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.

Finally, Criado and Such [67] 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. [68] 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.

The above models are complementary to the work we present in this paper: they help the users identify their *individual* privacy preferences; then, if a conflict is detected, ELVIRA can support them to identify the optimal collective sharing policy.

7.2 Collaborative privacy management

We now discuss the main approaches suggested so far to solve MPCs in OSNs, but refer the reader to reviews on the topic for more details and references [1, 3, 8]. Similarly to ELVIRA, the models that we discuss here 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 [23, 69, 70], whose detecting mechanisms could be preliminarily applied in combinations with other resolutive models.

Most of the methods to solve MPCs in OSNs that have been suggested recently are based on preference-aggregation techniques: in [69, 71–74] the solution is identified mostly by majority voting; [24, 75, 76] 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. [77] 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. [78] 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 [79] present a similar auction system, enriched with an abuse control feature and with agents that can learn the users' bidding strategies. In [80, 81] 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 [82], 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 [38], where different argument schemes prove to be very influential when identifying the optimal sharing solution. Ruiz-Dolz et al. [83] propose a model similar to the one for individual privacy mentioned earlier [68], 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 [84], 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. [85] introduce an iterative negotiation mechanism where, through semantic rules, the co-owners can justify the eventual rejection of sharing offers to help the uploader

Table 13 Summary of the properties satisfied by previous approaches in the literature; *Marks partial fulfilment of the property

Approaches	RA	AD	UD	VD	EX
game-theory	[23, 78–81]	[23, 78–81]	[23, 78–81]	–	–
Aggregation	[24, 69, 71–77]	[24, 69, 75–77]	[69, 77]	–	–
Human values	[87, 88]	[87, 88]	–	[87, 88]	–
Learning	[82, 89, 90]	[82, 89, 90]	–	–	–
Argumentation	[38, 84]	[38]	–	–	[83], *
Semantic rules	[85, 86]	[85, 86]	[86]	–	*
Norms	[88–90]	[88–90]	–	[88]	*
Obfuscation	[25, 92, 93]	[25, 92, 93]	–	–	–
Cryptography	[91, 92]	[91, 92]	–	–	–

suggest an acceptable policy. Kekulluoglu et al. [86] 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 [23].

Mosca et al. [87] introduce a multi-step negotiation protocol where the strategies are driven by moral values. A value-based component in the context of data sharing is also defined by Ajmeri et al. [88], 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. [89], where privacy norms based on the social identity theory are learnt adaptively for different contexts, and by Ulusoy and Yolum [90], where privacy decisions are made according to social and individual norms emerged from previous activities.

Finally, cryptography [91, 92] and obfuscation through image processing techniques [25, 92, 93] offer more fine-grained solutions to MPCs, where 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.

All the approaches that we mentioned above 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 MPCs. However, if we consider the requirements introduced in Sect. 1.1, then all these models reveal evident weaknesses and none of them presents all the required properties, as we summarise in Table 13.

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 context-dependent. Some aggregation-based models, such as [71–74], are generally not adaptive because of their rigid and static way of aggregating the users' preferences. Argumentation approaches [83, 84] 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 fulfilled by the game-theoretical approaches, and by some other specific models [69, 77, 86] 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 are only [87, 88]. However, there have been efforts directed towards modelling real-world dynamics that may occur in MPC, where the users often concede and try to accommodate each other's preferences [5], such as reciprocity [86] and bounded rationality [80].

Finally, approaches based on argumentation [38, 84], or that use semantic rules [85, 86] or normative systems [88] have the potential to support some type of *explainability* of the system, but none of these works autonomously generate explanations for their outputs and share them with the users. Finally, approaches based on argumentation [38, 83, 84], or that use semantic rules [85, 86] or normative systems [88] 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 [83], where explanations are explicitly defined in the model, but in a static way, offering limited information, and without empirical validation.

8 Discussion

In this paper we presented ELVIRA, an agent-based model that supports multiuser privacy in OSNs. Our approach satisfies all the main requirements that have been previously suggested in the literature in order to obtain satisfying solutions, namely being role-agnostic, adaptive, utility-driven, value-driven, and explainable.

As we discussed in Sect. 7.2, most of the models in the related literature are role-agnostic and many reach an adequate level of adaptability. However, to the best of our knowledge, no solution has been presented so far that explicitly takes into account both the gain or loss in utility that OSNs users can experience and the promotion or demotion of their moral values. Furthermore, ELVIRA is the first approach to provide full explainability, i.e. the agent is able to generate and convey explanations for its optimal recommendations.

We thoroughly and extensively evaluated ELVIRA. By combining utility and values in the computation of the solution, the agent is able to identify solutions that better mimic the real dynamics of collaborative decision making in privacy, where every user is mainly self-interested but often cares about others. This is known to happen from the empirical literature about real conflicts in OSN [5] and clearly transpired in our thematic analysis of the motivations that support the acceptance or rejection of a recommendation. Software simulations proved the benefits of computing a solution that considers both utility and values and showed how considering only one of these factors leads to a poor performance in the other.

The design of the explanations was informed by previous studies in Explainable AI, Social Sciences and Cognitive Sciences and aimed at fostering the users' trust in the agent, by helping them recognise that it always recommended the optimal solution to the conflict. The agent may earn the user's trust mainly by aligning its outputs with the user's needs and beliefs: this can happen by accounting for the user's values [33] and by offering tailored and contrastive explanations [9]. For this reason, we suggested that an explanation includes first a complete description of the conflict and then a summary of the benefits of the solution from the user's point of view, that is considering their preferences and values. This design was positively evaluated by users, who reported to be generally satisfied with the level of detail and accuracy of the agent's output (see Sect. 6.2). Furthermore, ELVIRA managed to nudge the users to be more conscious of the co-owners and of the impact that their online privacy decisions may bring upon them. This is a very important result which suggests how beneficial would be the deployment of ELVIRA in real-world platforms where, as we mentioned in the Introduction to this paper, the vast majority of MPCs is caused by the difficulty faced by the uploaders to identify suitable sharing policies for their co-owned contents.

8.1 Limitations and future directions

Despite the positive results obtained in the evaluation of ELVIRA, we are aware of some limitations of the model and foresee directions of improvement.

First, while the benefits of considering moral values in identifying and explaining the solution of an MPC are evident in terms of mimicking real dynamics and fostering users' trust, less evident is which and how moral values should be represented and accounted for by autonomous agents. As reported in Sect. 2.1, ELVIRA relies on the Schwartz theory of basic values [27] because of its extensive validation and previous applications. This design decision seems successful, given the encouraging users' feedback, but ELVIRA could be easily adapted to other sets of values and/or correspondences of values and behaviours.

Second, the nature of relationships on OSNs can impact in several ways the management of multi-user privacy. In ELVIRA, we take this into account during the definition of sharing policies, by considering the weight (distance and intimacy) of the relationship that the co-owners have with their audience (cf. Sect. 2). Nonetheless, the relationship between co-owners themselves might also have an impact during conflict resolution. As a future direction, it would be interesting to enrich ELVIRA's model of interpersonal behaviour, which is currently based on the Schwartz values, with elements dependent on the nature of relationships between co-owners. Unfortunately, how the strength of a relationship influences the emergence and the resolution of an MPC is still unclear [5] and further empirical evidence should be gathered in order to better understand, and consequently appropriately model, these dynamics.

Finally, ELVIRA generates explanations for one-shot interactions with the users. Tailored and contrastive explanations are designed according to *assumptions* on what

the user might be interested in, i.e., the comparison between the optimal solution and their original preference. Although this proved satisfactory in practice with users, sometimes users may desire more or different information. As an interesting line of future research, ELVIRA's social capabilities could be extended to allow dialogical explanations. During a conversation with the user, ELVIRA could answer any relevant question on the MPC thanks to the knowledge of the MPC that it has gathered in the practical reasoning process. Furthermore, by interacting with the user across different MPCs, the agent could learn what the user is more interested about and provide better tailored explanations over time.

9 Conclusion

The management of multiuser privacy, particularly in online social networks, has been raising concerns in recent years and scholars have focussed on defining and identifying acceptable solutions. Informed by previous research and by empirical evidence on multiuser privacy [5, 8], solutions should be acceptable when they are role-agnostic, adaptive, utility- and value-driven, and explainable. None of the previously presented models satisfies all these requirements. In this paper, we presented ELVIRA, an autonomous agent whose design is completely aligned with these requirements. ELVIRA takes into account both utility and moral values in the computation of the optimal solution to the MPC, by following a practical reasoning process which enables it to autonomously generate explanations. We explored different designs of the explanations in a user study (see Sect. 4) which helped us define ELVIRA's explanation format, where tailored and contrastive explanations are offered according to the circumstances. Then, we performed an extensive evaluation of ELVIRA, by comparing it against other approaches that have been previously suggested in the literature. Software simulations showed the benefits of combining both utility and values in the computation of the solutions (Sect. 5), which were overall more appreciated by the users of a second user study (Sect. 6). Finally, role-agnosticism and adaptivity were proven formally (Sect. 3.2). As future work, one of the aspects we would like to explore is the applicability of ELVIRA to other domains where multiuser privacy is known to occur, such as cloud computing [94] and the smart home [95].

Appendix

Appendix A Notation

See Table 14.

Table 14 Summary of symbols and notation

Section	Symbol	Description
Preliminaries (Sect. 2)	$G = (V, R)$	Social graph where V are the users and R their relationships
	i	Intimacy, weight of a relationship on G
	X	Set of digital contents that could be shared online
	Ag	Set of the co-owners of an item $x \in X$
	$sp = \langle d, i \rangle$	Sharing policy where d is the distance between two users
	$aud_{sp_k,k}$	Individual preferred audience of user $k \in Ag$, defined by k's preferred sharing policy sp_k
	aud_{sp}	Collective audience (intersection of individual audiences defined by sp)
	$\mathcal{A} \Leftrightarrow \mathcal{B} \Leftrightarrow \mathcal{C} \Leftrightarrow \mathcal{D}$	Allowed audience, allowed extra audience, excluded audience, excluded extra audience
	app	Appreciation
	$u_{k,aud}$	Utility of the audience aud for user $k \in Ag$
	α, β	Parameters that determine whether to share as-it-is ($\alpha = 1, \beta = 0$) or modified $0 < \alpha, \beta < 1$
	$\mathcal{V} = \{CO, OTC, SE, ST\}$	Set of Schwartz hypervalues (conservation, openness-to-change, self-enhancement, self-transcendence)
ELVIRA (Sect. 3)	n	Number of co-owners (Ag) involved in the MPC
	aud_f	Arbitrary audience, different from all the individual preferences
	$\mathbb{A}$	Finite set of candidate solutions (audiences) for the MPC
	$s_{k,aud}$	Score of the audience aud for user $k \in Ag$
	$v_{k,aud}$	Value promotion of the audience aud for user $k \in Ag$
	J_{Ag}	Set of joint actions
	q_0	Initial state representing the MPC
	aud'	Candidate optimal solution
Generating explanations (Sect. 4)	exp0	Baseline structure, no explanation
	exp1	General explanation structure
	exp2	Contrastive explanation structure
	$\{O\}$	Optimal solution audience to be recommended in the explanation
	$\{P\}$	The user's preferred audience in the explanation
	Q1–Q8	Questions in the Satisfaction Scale

Table 14 (continued)

Section	Symbol	Description
Simulations (Sect. 5)	N	Size of the network (number of nodes)
	d	Connectivity of thenetwork
	T	Number of simulated MPCs
	$iauc$, $iavc$	Individual average variation of utility and value promotion per each conflict
	csu, csv	Cumulative increment of social utility and of value promotion

Acknowledgements The authors would like to thank the anonymous reviewers of the AAMAS and JAAMAS community for their insightful comments on previous versions of this manuscript.

Funding Mosca's work was partly funded through a J.P. Morgan AI Research Fellowship.

References

1. Such, J., & Criado, N. (2018). Multiparty privacy in social media. *Communications of the ACM, 61*(8), 74–81.
2. Besmer, A., & Lipford, H. R. (2010). Moving beyond untagging: Photo privacy in a tagged world. In *CHI* (pp. 1563–1572). ACM.
3. Humbert, M., Trubert, B., & Huguenin, K. (2019). A survey on interdependent privacy. *ACM Computing Surveys, 52*(6), 1.
4. Wisniewski, P., Lipford, H., & Wilson, D. (2012). Fighting for my space: Coping mechanisms for SNS boundary regulation. In *CHI* (pp. 609–618). ACM.
5. Such, J., Porter, J., Preibusch, S., & Joinson, A. (2017). Photo privacy conflicts in social media: A large-scale empirical study. In *CHI* (pp. 3821–3832). ACM.
6. Misra, G., & Such, J. (2016). How socially aware are social media privacy controls? *IEEE Computer, 49*(3), 96–99.
7. Liang, K., Liu, J. K., Lu, R., & Wong, D. S. (2014). Privacy concerns for photo sharing in online social networks. *IEEE Internet Computing, 19*(2), 58–63.
8. Paci, F., Squicciarini, A., & Zannone, N. (2018). Survey on access control for community-centered collaborative systems. *ACM Computing Surveys, 51*(1), 1–38.
9. Miller, T. (2018). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38.
10. Cherubini, M., Niksirat, K., Boldi, M.-O., Keopraseuth, H., Such, J., & Huguenin, K. (2021). When forcing collaboration is the most sensible choice: Desirability of precautionary and dissuasive mechanisms to manage multiparty privacy conflicts. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW1), 1–36.
11. Lampinen, A., Lehtinen, V., Lehmuskallio, A., & Tamminen, S. (2011). We're in it together: Interpersonal management of disclosure in social network services. In *CHI* (pp. 3217–3226). ACM.
12. Acquisti, A., Brandimarte, L., & Loewenstein, G. (2015). Privacy and human behavior in the age of information. *Science, 347*(6221), 509–514.
13. Such, J. (2017). Privacy and autonomous systems. In *Proceedings of the 26th international joint conference on artificial intelligence (IJCAI)* (pp. 4761–4767).
14. Krasnova, H., Spiekermann, S., Koroleva, K., & Hildebrand, T. (2010). Online social networks: Why we disclose. *JIT, 25*(2), 109–125.
15. Mosca, F. (2020). Value-aligned and explainable agents for collective decision making: Privacy application: Doctoral consortium. In *Proceedings of the 19th international conference on autonomous agents and multiagent systems (AAMAS 2020)*.
16. Mosca, F., Such, J. M., & McBurney, P. (2020). Towards a value-driven explainable agent for collective privacy. In *Proceedings of the 19th international conference on autonomous agents and multiagent systems*.
17. Nissenbaum, H. (2011). A contextual approach to privacy online. *Daedalus, 140*(4), 32–48.
18. Winikoff, M. (2017). Towards trusting autonomous systems. In *International workshop on engineering multi-agent systems* (pp. 3–20). Springer
19. Cranefield, S., Oren, N., & Vasconcelos, W. W. (2018). Accountability for practical reasoning agents. In *International conference on agreement technologies* (pp. 33–48). Springer

20. Dignum, V. (2019). *Responsible artificial intelligence: How to develop and use AI in a responsible way*. Springer.
21. Mosca, F., & Such, J. (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 (AAMAS)*.
22. Fogues, R., Such, J., Espinosa, A., & Garcia-Fornes, A. (2014). Bff: A tool for eliciting tie strength and user communities in social networking services. *Information Systems Frontiers, 16*(2), 225–237.
23. Such, J., & Rovatsos, M. (2016). Privacy policy negotiation in social media. *ACM TAAS, 11*(1), 1–29.
24. Such, J., & Criado, N. (2016). Resolving multi-party privacy conflicts in social media. *IEEE TKDE, 28*(7), 1851–1863.
25. Ilia, P., Polakis, I., Athanasopoulos, E., Maggi, F., & Ioannidis, S. (2015). Face/off: Preventing privacy leakage from photos in social networks. In *CCS* (pp. 781–792). ACM Press
26. Ramokapane, K. M., Misra, G., Such, J., & Preibusch, S. (2021). Truth or dare: Understanding and predicting how users lie and provide untruthful data online. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1–15).
27. Schwartz, S. H. (2012). An overview of the Schwartz theory of basic values. *Online Readings in Psychology and Culture, 2*(1), 11.
28. Schwartz, S. H. (2003). A proposal for measuring value orientations across nations. *Questionnaire Package of the European Social Survey, 259*(290), 261.
29. Rokeach, M. (1973). *The nature of human values*. Free Press.
30. Bardi, A., & Schwartz, S. H. (2003). Values and behavior: Strength and structure of relations. *Personality and Social Psychology Bulletin, 29*(10), 1207–1220.
31. Atkinson, K., & Bench-Capon, T. (2007). Practical reasoning as presumptive argumentation using action based alternating transition systems. *Artificial Intelligence, 171*(10–15), 855–874.
32. Atkinson, K., & Bench-Capon, T. (2018). Taking account of the actions of others in value-based reasoning. *Artificial Intelligence, 254*, 1–20.
33. Chander, A., & Srinivasan, R. (2018). Evaluating explanations by cognitive value. In *International cross-domain conference for machine learning and knowledge extraction* (pp. 314–328). Springer
34. Langley, P. (2019). Explainable, normative, and justified agency. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, pp. 9775–9779).
35. Mosca, F., Sarkadi, Ş., Such, J. M., & McBurney, P. (2020). Agent EXPRI: Licence to explain. In *International workshop on explainable, transparent autonomous agents and multi-agent systems* (pp. 21–38). Cham: Springer.
36. Tessier, C., Chaudron, L., & Müller, H.-J. (2006). *Conflicting agents: Conflict management in multi-agent systems* (Vol. 1). Springer.
37. Mancini, C., Rogers, Y., Bandara, A. K., Coe, T., Jedrzejczyk, L., Joinson, A. N., Price, B. A., Thomas, K., & Nuseibeh, B. (2010). Contravision: Exploring users' reactions to futuristic technology. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 153–162).
38. Fogues, R., Murukannaiah, P., Such, J., & Singh, M. (2017). Sharing policies in multiuser privacy scenarios: Incorporating context, preferences, and arguments in decision making. *ACM TOCHI, 24*(1), 5–1529.
39. Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. arXiv preprint arXiv:1812.04608
40. Mason, W., & Suri, S. (2012). Conducting behavioral research on amazon's mechanical turk. *Behavior Research Methods, 44*(1), 1–23.
41. Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on amazon mechanical turk. *Behavior Research Methods, 46*(4), 1023–1031.
42. Hauser, D. J., & Schwarz, N. (2016). Attentive turkers: Mturk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods, 48*(1), 400–407.
43. Paas, L. J., & Morren, M. (2018). Please do not answer if you are reading this: Respondent attention in online panels. *Marketing Letters, 29*(1), 13–21.
44. Malhotra, N. K., Kim, S. S., & Agarwal, J. (2004). Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information Systems Research, 15*(4), 336–355.
45. Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*(5439), 509–512.
46. Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., & Bhattacharjee, B. (2007). Measurement and analysis of online social networks. In *ICM* (pp. 29–42). ACM

47. Viswanath, B., Mislove, A., Cha, M., & Gummadi, K. P. (2009). On the evolution of user interaction in Facebook. In *Proceedings of the 2nd ACM workshop on online social networks* (pp. 37–42). ACM.
48. Leskovec, J., & Mcauley, J. J. (2012). Learning to discover social circles in ego networks. In *NIPS* (pp. 539–547).
49. Terry, G., Hayfield, N., Clarke, V., & Braun, V. (2017). Thematic analysis. In: *The Sage handbook of qualitative research in psychology* (pp. 17–37). Sage.
50. Fogues, R., Such, J., Espinosa, A., & Garcia-Fornes, A. (2015). Open challenges in relationship-based privacy mechanisms for social network services. *International Journal of Human-Computer Interaction, 31*(5), 350–370.
51. Johnson, M., Egelman, S., & Bellovin, S. M. (2012). Facebook and privacy: It's complicated. In *Proceedings of the eighth symposium on usable privacy and security* (pp. 1–15).
52. Squicciarini, A., Caragea, C., & Balakavi, R. (2017). Toward automated online photo privacy. *ACM Transactions on the Web (TWEB), 11*(1), 1–29.
53. Zerr, S., Siersdorfer, S., Hare, J., & Demidova, E. (2012). Privacy-aware image classification and search. In *Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval* (pp. 35–44).
54. Fogues, R., Such, J., Espinosa, A., & Garcia-Fornes, A. (2018). Tie and tag: A study of tie strength and tags for photo sharing. *PLoS ONE, 13*(8), 1–22.
55. Reinhardt, D., Engelmann, F., & Hollick, M. (2015). 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* (pp. 347–356).
56. Shehab, M., & Touati, H. (2012). Semi-supervised policy recommendation for online social networks. In *2012 IEEE/ACM international conference on advances in social networks analysis and mining* (pp. 360–367). IEEE
57. Misra, G., Such, J., & Balogun, H. (2016). 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)* (pp. 49–56). https://doi.org/10.1109/ASONAM.2016.7752212
58. Amershi, S., Fogarty, J., & Weld, D. (2012). Regroup: Interactive machine learning for on-demand group creation in social networks. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 21–30).
59. Misra, G., Such, J., & Balogun, H. (2016). Improve-identifying minimal profile vectors for similarity based access control. In *IEEE Trustcom* (pp. 868–875).
60. Squicciarini, A. C., Lin, D., Sundareswaran, S., & Wede, J. (2014). Privacy policy inference of user-uploaded images on content sharing sites. *IEEE Transactions on Knowledge and Data Engineering, 27*(1), 193–206.
61. Albertini, D. A., Carminati, B., & Ferrari, E. (2016). Privacy settings recommender for online social network. In *2016 IEEE 2nd international conference on collaboration and internet computing (CIC)* (pp. 514–521). IEEE.
62. Li, Q., Li, J., Wang, H., & Ginjala, A. (2011). Semantics-enhanced privacy recommendation for social networking sites. In *2011 IEEE 10th international conference on trust, security and privacy in computing and communications* (pp. 226–233). IEEE.
63. Kurtan, A. C., & Yolum, P. (2021). Assisting humans in privacy management: An agent-based approach. *Autonomous Agents and Multi-Agent Systems, 35*(1), 1–33.
64. Kepez, B., & Yolum, P. (2016). Learning privacy rules cooperatively in online social networks. In *Proceedings of the 1st international workshop on AI for privacy and security* (pp. 1–4).
65. Misra, G., & Such, J. (2017). Pacman: Personal agent for access control in social media. *IEEE Internet Computing, 21*(6), 18–26.
66. Misra, G., & Such, J. (2017). React: Recommending access control decisions to social media users. In *IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)* (pp. 421–426).
67. Criado, N., & Such, J. (2015). Implicit contextual integrity in online social networks. *Information Sciences, 325*, 48–69.
68. Ruiz-Dolz, R., Alemany, J., Heras, S., & García-Fornes, A. (2019). *Automatic generation of explanations to prevent privacy violations*.
69. Hu, H., Ahn, G. J., & Jorgensen, J. (2011). Detecting and resolving privacy conflicts for collaborative data sharing in online social networks. In *ACSAC* (pp. 103–112). ACM.

70. Zhong, H., Squicciarini, A., & Miller, D. (2018). Toward automated multiparty privacy conflict detection. In *Proceedings of the 27th ACM international conference on information and knowledge management* (pp. 1811–1814).
71. Thomas, K., Grier, C., & Nicol, D. (2010). Unfriendly: Multi-party privacy risks in social networks. In *PET* (pp. 236–252). Springer.
72. Carminati, B., & Ferrari, E. (2011). Collaborative access control in on-line social networks. In *CollaborateCom* (pp. 231–240). IEEE
73. Hu, H., Ahn, G.-J., & Jorgensen, J. (2012). Multiparty access control for online social networks: Model and mechanisms. *IEEE Transactions on Knowledge and Data Engineering, 25*(7), 1614–1627.
74. Ratikan, A., & Shikida, M. (2014). Privacy protection based privacy conflict detection and solution in online social networks. In *International conference on human aspects of information security, privacy, and trust* (pp. 433–445). Springer.
75. Shetty, N. P., Muniyal, B., & Mowla, S. (2020). Policy resolution of shared data in online social networks. *International Journal of Electrical& Computer Engineering, 10*, 3767.
76. Akkuzu, G., Aziz, B., & Adda, M. (2020). Towards consensus-based group decision making for co-owned data sharing in online social networks. *IEEE Access, 8*, 91311–91325.
77. Xu, L., Jiang, C., He, N., Han, Z., & Benslimane, A. (2018). Trust-based collaborative privacy management in online social networks. *IEEE Transactions on Information Forensics and Security, 14*(1), 48–60.
78. Squicciarini, A., Shehab, M., & Paci, F. (2009). Collective privacy management in social networks. In *WWW* (pp. 521–530). ACM.
79. Ulusoy, O., & Yolum, P. (2020). Agents for preserving privacy: Learning and decision making collaboratively. In *Multi-agent systems and agreement technologies* (pp. 116–131). Springer.
80. Rajtmajer, S., Squicciarini, A., Griffin, C., Karumanchi, S., & Tyagi, A. (2016). Constrained social-energy minimization for multi-party sharing in online social networks. In *Proceedings of the international conference on autonomous agents & multiagent systems (AAMAS)* (pp. 680–688).
81. Rajtmajer, S., Squicciarini, A., Such, J., Semonsen, J., & Belmonte, A. (2017). An ultimatum game model for the evolution of privacy in jointly managed content. In *GAMESEC* (pp. 112–130). Springer.
82. Fogues, R., Murukannaiah, P., Such, J., & Singh, M. (2017). Sosharp: Recommending sharing policies in multiuser privacy scenarios. *IEEE Internet Computing, 21*(6), 28–36.
83. Ruiz-Dolz, R., Heras, S., Alemany, J., & García-Fornes, A. (2019). Towards an argumentation system for assisting users with privacy management in online social networks. In *CMNA@ PERSUASIVE* (pp. 17–28).
84. Kökciyan, N., Yaglikci, N., & Yolum, P. (2017). An argumentation approach for resolving privacy disputes in online social networks. *ACM TOIT, 17*(3), 27.
85. Mester, Y., Kökciyan, N., & Yolum, P. (2015). Negotiating privacy constraints in online social networks. In *International workshop on multiagent foundations of social computing* (pp. 112–129). Springer.
86. Kekulluoglu, D., Kökciyan, N., & Yolum, P. (2018). Preserving privacy as social responsibility in online social networks. *ACM TOIT, 18*(4), 42.
87. Mosca, F., Such, J., & McBurney, P. (2019). Value-driven collaborative privacy decision making. In *AAAI PAL symposium*.
88. Ajmeri, N., Guo, H., Murukannaiah, P. K., & Singh, M. P. (2020). Elessar: Ethics in norm-aware agents. In *Proceedings of the international conference on autonomous agents and multi-agent systems (AAMAS)* (pp. 16–24).
89. Calikli, G., Law, M., Bandara, A. K., Russo, A., Dickens, L., Price, B. A., Stuart, A., Levine, M., & Nuseibeh, B. (2016). 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)* (pp. 47–56). IEEE.
90. Ulusoy, O., & Yolum, P. (2020). Norm-based access control. In *Proceedings of the 25th ACM symposium on access control models and technologies* (pp. 35–46).
91. Beato, F., & Peeters, R. (2014). Collaborative joint content sharing for online social networks. In *2014 IEEE international conference on pervasive computing and communication workshops (PERCOM WORKSHOPS)* (pp. 616–621). IEEE.
92. Olteanu, A.-M., Huguenin, K., Dacosta, I., & Hubaux, J.-P. (2018). Consensual and privacy-preserving sharing of multi-subject and interdependent data. In *Proceedings of the 25th network and distributed system security symposium (NDSS)* (pp. 1–16). Internet Society.

93. Vishwamitra, N., Li, Y., Wang, K., Hu, H., Caine, K., & Ahn, G.J. (2017). Towards pii-based multiparty access control for photo sharing in online social networks. In *SACMAT* (pp. 155–166). ACM.
94. Ramokapane, K. M., Rashid, A., & Such, J. M. (2017). "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)* (pp. 241–256).
95. Abdi, N., Zhan, X., Ramokapane, K. M., & Such, J. (2021). Privacy norms for smart home personal assistants. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1–14).