ELVIRA: an Explainable Agent for Value and Utility-driven Multiuser Privacy

Francesca Mosca
King's College London
London, UK
francesca.mosca@kcl.ac.uk

Jose M Such
King's College London
London, UK
jose.such@kcl.ac.uk

ABSTRACT

Online social networks fail to support users to adequately share co-owned content, which leads to privacy violations. Scholars proposed collaborative mechanisms to support users, but they did not satisfy one or more requirements needed according to empirical evidence in this domain, such as explainability, role-agnosticism, adaptability, and being utility- and value-driven. We present ELVIRA, an agent that supports multiuser privacy, whose design meets all these requirements. By considering the sharing preferences and the moral values of users, ELVIRA identifies the optimal sharing policy. Furthermore, ELVIRA justifies the optimality of the solution through explanations based on argumentation. We prove via simulations that ELVIRA provides solutions with the best trade-off between individual utility and value adherence. We also show through a user study that ELVIRA suggests solutions that are more acceptable than existing approaches and that its explanations are also more satisfactory.

KEYWORDS

Multiuser Privacy; Explainable Agents; Value-based Agents; Human-Agent Interaction; Practical Reasoning

ACM Reference Format:
Francesca Mosca and Jose M Such. 2021. ELVIRA: an Explainable Agent for Value and Utility-driven Multiuser Privacy. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), Online, May 3–7, 2021*, IFAAMAS, 9 pages.

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 [45]. In particular, whenever the content to be shared involves more people, they should all have a say on with whom the content is shared. Otherwise, a *Multiuser Privacy Conflict* (MPC) is likely to occur. MPCs are frequent, and have been suffered by the majority of OSN users [46, 54].

A commonly-studied example is the case of a photo depicting a group of friends, where each one of them would assign different degrees of publicity/privacy to the picture on OSN. Currently, OSN platforms lack collaborative privacy controls [54] and the responsibility to decide the sharing policy for some content is generally left solely to its uploader. If the other involved users (the *co-owners*) are unhappy with the uploader's choice, they can only resort to unsatisfactory reparative solutions (e.g. untagging), which do not guarantee avoiding a privacy violation [22, 46].

Although some MPCs occur in adversarial settings (e.g., revenge porn), the vast majority of MPCs happen in *non-adversarial settings* [46], where it is simply too difficult for uploaders to identify the optimal sharing policy for an item that involves other co-owners [6, 20, 54].

In order to tackle these common non-adversarial MPCs, research was conducted on how to design collaborative models to support adequate multiuser privacy management in OSN. Importantly, previous research and empirical evidence suggest that models should meet the requirements below to help resolve MPCs collaboratively [1, 29, 33, 36, 46].

First, *role-agnosticism* (RA), i.e., models should treat all the users involved in an MPC in the same way regardless of whether they are uploaders or co-owners. In fact, the asymmetry between uploaders and co-owners in their ability to influence access to content in OSNs is among the main causes for MPCs [54].

Second, *adaptability* (AD), i.e., a model should behave differently depending on the users' subjective preferences, because individuals manage privacy in different ways depending on the context [1].

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 [18]; and negatively experiencing privacy violations [20].

Fourth, *value-driven* (VD), i.e., models should consider moral values, because empirical evidence suggests that users do so in MPCs [46], e.g., 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 [26], is crucial for users to know why a solution is suggested and its effects [36], and to align the differences between uploaders and co-owners [46].

Although models for better supporting users to collaboratively deal with MPCs have been proposed in the related literature (see Sec. 2), none of them satisfies all these requirements together. In this paper, we present ELVIRA, an agent that satisfies the above requirements to recommend the best sharing policy for an MPC. In particular, ELVIRA is both utility-driven and value-driven (see Sec. 3), as it computes the sharing policy in terms of: (i) the utility that each user gains/loses from sharing the content with each particular audience; and (ii) the promotion of moral values, i.e. the degree of coherency with the user's values of choosing each possible policy. When comparing with existing models, we show experimentally that ELVIRA recommends solutions which present

Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (eds.), May 3–7, 2021, Online.

Table 1: Comparison of the requirements met by models representative of different approaches in the literature.

Approach	EX	RA	AD	UD	VD
Game-theoretic [43]	-	✓	✓	✓	-
Aggregation-based [49]	-	✓	-	-	-
Learning-based [9]	-	✓	✓	-	-
Argumentation-based [17]	✓	✓	-	-	-
Value-driven [32]	-	✓	✓	-	✓
Fine-grained [15]	-	✓	✓	-	-
No support (Facebook)	-	-	-	-	-

better utility-value trade-offs (see Sec. 5) and are generally more accepted by users (see Sec. 6). In addition, ELVIRA is able to justify its recommendations by providing explanations, which we show are satisfactory to users. Finally, formal properties such as soundness, completeness, anonymity and neutrality guarantee ELVIRA to be adaptive and role-agnostic.

2 RELATED WORK

We now discuss briefly 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 [14, 36, 45]. For each approach, we comment how it meets the requirements previously introduced; Tab. 1 shows a summary, with a representative work from each approach.

In game-theoretic approaches, the protocols and strategies that the users can follow to find an agreement are analysed according to game theory solution concepts [16, 38, 39, 43, 47]. In general, they are role-agnostic, utility-driven and adaptive. Despite their elegant formal frameworks building upon well-studied analytic tools, users' behaviour does not seem to be always rational in MPC [20, 54] (as assumed in these approaches), and even if some models are starting to consider bounded rationality [38] and other factors like reciprocity [16], they are far from considering a general value-driven approach.

Aggregation-based approaches combine the users' preferences in various ways, e.g. by applying voting rules such as majority and veto [7, 13, 49]. These models are mostly role-agnostic, but their rigid way of aggregating the preferences makes them generally not adaptive. In [44] users' preferences are aggregated more adaptively according to the context, but only in a limited number of situations.

Learning-based approaches [9, 50, 51] look at decisions made in the past to come up with the optimal sharing policy in the future, making these models role-agnostic and adaptive in theory. However, the lack of explainability about why the particular recommendation is made, beyond that a similar decision was made in the past, can hinder the user's endorsement of the proposed solution [36].

There are also argumentation-based approaches [8, 17], which partially address explainability, by being able to generate an explanation for the optimal sharing policy; and approaches that are mostly value-driven [2, 32], as they generate the sharing policy that adheres the most to the values of the users involved. However, both approaches only offer limited support in terms of the other requirements.

Finally, there are fine-grained approaches that allow individualised decisions about whether some personally-identifying objects (e.g. faces) in photo are shown or blurred [15, 34, 52]. They are role-agnostic and adaptive but do not meet the rest of requirements.

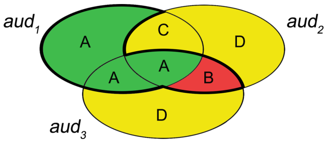

Figure 1: MPC between 3 users, a possible solution aud_s (represented with bold borders), and the A,B,C,D sets for user1.

3 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 [10]. In line with previous work [47], but noting that this is equivalent and can be translated to and back from the group-based access control models used in OSN platforms [44], we define a *sharing policy* as follows:

Definition 1. *A **sharing policy** for an item from user k is $sp_k = \langle d, i \rangle$, 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, and that it can be elicited automatically (e.g. see [19, 28]). In addition, each sharing policy sp_k defines an *audience* aud_k, i.e. a set of users who satisfy the conditions of sp_k. A MPC occurs when users that are involved in the same item have contradictory preferred sharing policies.

Definition 2. *A **MPC** regarding an item co-owned by users k and j occurs when k and j's preferred audiences do not coincide, i.e. $aud_k \neq aud_j$.*

Users are known to benefit from sharing in social media [18], 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 [46].

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. In addition, 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 [15]. In fact, empirical evidence [46] suggests sharing modified content is sometimes an acceptable compromise among co-owners. 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_k, considering the audience aud_s as a potential solution to a MPC where k is involved (see Fig.1 for a graphical representation), then the

Table 2: Variation of the individual utility for item x, considering audience sets, appreciation and mode of sharing.

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

appreciation function capturing the tendencies to under/over-share, and finally the utility function.

Definition 3. *The* ***allowed audience*** *A is the set of users who k desires to grant access to x and that are part of the solution audience, i.e., $A = aud_k \cap aud_s$. The* ***allowed extra audience*** *B is the set of users who k desires to forbid access to x but that are part of the solution audience, i.e., $B = (\bigcup_{l\neq k} aud_l \setminus aud_k) \cap aud_s$. The* ***excluded audience*** *C is the set of users who k desires to grant access to x but that are forbidden to access or allowed to access only a modified version, i.e., $C = aud_k \setminus aud_s$. The* ***excluded extra audience*** *D is the set of users who k desires to forbid access to x and that are either forbidden to access or allowed to access only a modified version of the item, i.e., $D = \bigcup_{l\neq k} aud_l \setminus aud_s$.*

Definition 4. *Given a set of pictures X, the function* ***appreciation****, $app : X \longrightarrow [-1, 1]$, maps a picture $x \in X$ into a positive value if the user is happy to overshare, and to a negative value if the user prefers to undershare.*

Definition 5. *Given an audience aud, its* ***utility*** *for user k is:*

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

For the sake of clarity, Tab. 2 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, parameters which determine whether to share the content *as-it-is* ($\alpha = 1$ and $\beta = 0$) or *modified* ($0 < \alpha, \beta < 1$). However, experiments showed (see [31]) 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.

3.1 Schwartz Basic Values

The *theory of basic values* by Schwartz [42] is one of the most well-known and established theories of human values backed by strong empirical evidence. In this theory, values are socially desirable concepts that represent the mental goals which drive human behaviour and influence any people's decision.

Schwartz identifies two dimensions which summarise the main value tendencies, defining four directions which pull apart while defining the behaviours. On one axis, *openness to change* (OTC) is opposed to *conservation* (CO), representing dynamic and independent ways of acting versus conservatory 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.

Table 3: 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 the same
	-	with some user's pref	
CO	+	with most private option	preserving everyone's privacy
	-	with a more public option	
ST	+	with the other's pref	making others happy
	-	ignoring the other user's pref	
SE	+	with own pref	getting your way
	+	gaining better utility	
	-	gaining worse utility	

The individual preferred order over the values is proven to be relatively stable over the lifetime [5], making sufficient to elicit it –through validated tools [42]– from the users just once. Such tools are more reliable than the ones offered by other value theories (see for instance [40]), which do not provide an overall value architecture or direct insights on the behavioural impact of the values.

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 doing what is good for the other people; and SE, meant as getting the own's 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 Tab. 3. In the remaining part of the paper, we refer to these value-directions as $\mathcal{V}$.

Example. Let us consider the following situation, where the individually preferred sharing policies, considered *as-it-is*, imply different audiences and generate an MPC:

Users	Sharing Policy	Values	$app(x)$
Kay (uploader)	$\langle 3, 6\rangle$	$ST > OTC > CO > SE$	+1
Dan (co-owner1)	$\langle 2, 6\rangle$	$CO > SE > OTC > ST$	+1
Zoe (co-owner2)	$\langle 2, 8\rangle$	$SE > CO > ST > OTC$	-1

By selecting $\langle 2, 6\rangle$ as sharing policy, Kay would promote ST, because is selecting Dan's preference; however, Zoe would demote both CO, as $\langle 2, 6\rangle$ is not the most restrictive policy, and SE, because oversharing would get her a lower utility (the appreciation for the item is negative).

4 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 Sec. 1: *explainability* is given by the practical reasoning approach (Sec. 4.1) and the process to describe MPCs and their recommended solution (4.2), which are evaluated in Sec. 6; *role-agnosticism* and *adaptability* are guaranteed by its formal properties (Sec. 4.3); and, finally, both individual utility and moral values are explicitly considered to compute the solution to the MPC as described below (and evaluated in Sec. 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 the majority of MPCs happen in non-adversarial settings [6, 20, 54]. 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 because of lack 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[1], the uploader can offer to the co-owners an audience aud, chosen *as-it-is* or *modified*, from a finite set of options $\mathcal{A}$ which includes the n audiences $aud_1, ...aud_n$ deriving from the users' preferred sharing policies, and aud_f, where f is some function identifying a subset of the union of all the individually preferred audiences, such that $aud_f \neq aud_k \quad \forall k \in Ag$.

For each possible audience aud, each agent k computes its *individual score*, which represents its appreciation of the particular option in terms of utility and value promotion:

$$s_{k,aud} = u_{k,aud} \cdot v_{k,aud}. \tag{2}$$

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:

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

where $I = |\mathcal{V}| + 1$, 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:

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

4.1 Computing the Solution

In this section we describe how the ELVIRA uploader agent computes the solution to an MPC based on argumentation techniques, similarly to [30]. By completing the abductive reasoning process 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 [26]. We detail how ELVIRA uses this information to generate the explanations in Sec. 4.2.

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) [3]. AS can be expressed as: *"I should offer the audience $\widehat{aud}$, that will be accepted by the co-owners, that will generate the score $s_{\widehat{aud}}$ and that will promote the values V"*.

In order to identify the best solution to offer, ELVIRA uploader follows a practical reasoning process (PR)[3]: (1) it identifies the most desirable outcome, e.g. the audience $\widehat{aud}$; (2) it argues in favour of offering $\widehat{aud}$, 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) [3]. 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[2]. In the MPC context, a joint action is composed of the uploader's offer and the co-owners' response. We adapt Atkinson's definition of an AATS+V [3] to MPCs as follows:

DEFINITION 6. *In the context of an MPC among n users, an* ***AATS+V*** *is a $2n+8$ tuple $\Sigma = \langle Q, q_0, Ag, Ac_k, \rho, \tau, S, \mathcal{V}, Av_k, \delta \rangle$, with $k = 1...n$, where:*

- *$Q = \{conflict, agreement_{aud_{a,m}} \quad \forall aud \in \mathcal{A}\}$ is a finite, non-empty set of states, where each audience is considered as-it-is (aud_a) and modified (aud_m);*
- *$q_0 = conflict$ is the initial state;*
- *$Ag = \{up_1, co_2, ..., co_n\}$ is the set of agents involved in the MPC, with the roles of uploader or co-owners;*
- *$Ac_k = \{offer_{aud}, accept_{aud}, reject_{aud} \quad \forall aud \in \mathcal{A}\}$ are the actions available to the agents;*
- *$\rho : Ac_{Ag} \rightarrow 2^Q$ is the action-precondition function; here, every action can be executed just from q_0;*
- *$\tau : Q \times J_{Ag} \rightarrow Q$ defines what state results from performing the joint action j in the state q, where possible; here, only the joint actions where all the co-owners accept the uploader's offer end up in an agreement state, the others stay in q_0;*
- *$S = \{0, s_{aud} \quad \forall aud \in \mathcal{A}\}$ 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 Tab. 3).*

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 [46], 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.

[1] 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.

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

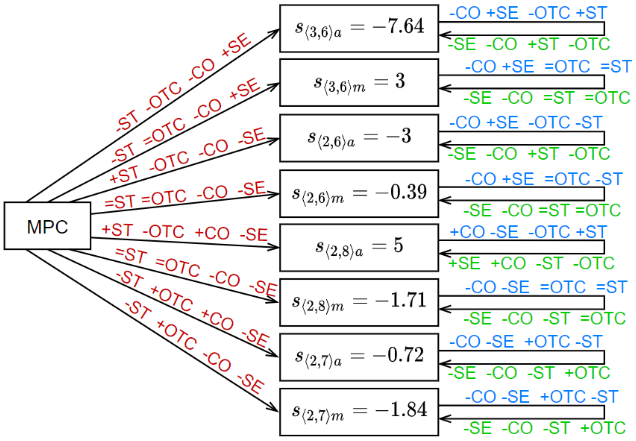

Figure 2: The AATS+V representing the PR performed by the agents in the example ($aud_f = \langle 2, 7\rangle$).

when the offered audience guarantees either (i) the individual maximum score ($s_{k,\widehat{aud}} = \max_{\mathcal{A}} s_{k,aud}$), or (ii) the collective maximum score ($s_{\widehat{aud}} = \max_{\mathcal{A}} s_{aud}$). 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 instantiating an appropriate argument scheme, and its consequent evaluation according to the preference over the values. Starting from AS, the agent discusses the CQs which contest the desirability of the audience $\widehat{aud}$:

- **CQ1** Would another audience guarantee a better score?
  i.e. $\exists aud \in \mathcal{A} : s_{aud} > s_{\widehat{aud}}$
- **CQ2** Would another audience with the same score promote better values?
  i.e. $\exists aud \in \mathcal{A} : s_{aud} = s_{\widehat{aud}} \wedge v_{Ag,aud} > v_{Ag,\widehat{aud}}$, where $v_{Ag,aud} = \sum_{k \in Ag} v_{k,aud}$
- **CQ3** Would any co-owner reject this offer?
  i.e. $\exists j \in J_{Ag}, k \in Ag : j_1 = \text{offer}_{\widehat{aud}} \wedge j_k = \text{reject}_{\widehat{aud}}$

If $\widehat{aud}$ 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. By reasoning on the AATS+V in Fig.2, $\widehat{aud} = \langle 2, 8\rangle_a$ is identified as a desirable audience, because $s_{\langle 2,8\rangle a} = 5$ is the maximum score and its selection would promote ST and CO. Next, the uploader discusses the CQs, which all get negative answers because (i) there is no audience which would give a better score; (ii) there is no audience with the same score promoting better values; and (iii) for EA2, the co-owners are believed to accept this offer, which would promote CO and ST for Dan and SE and CO for Zoe.

4.2 Generating Explanations

Following Miller's definition of *explainability* [26], an *explanation* is composed by a *cognitive process*, i.e. the process of abductive inference determining the causal attribution for a given event which we presented in the previous section, and a *social process*, i.e. the process of transferring knowledge between the explainer and the explainee.

We now discuss how ELVIRA conveys such information to the users. First, we identify the main features that the explanation for an MPC solution should present; here we keep referring just to the uploader agent, but an equivalent explanation structure can be applied for the co-owners as well. Then, in Sec. 6, we discuss the feedback we received from users regarding their satisfaction towards such designed explanations.

Conflict Description. If we aim to explain the solution to an MPC, we need to provide details also about other components of the conflict, such as its *detection* and *representation* [48]. This fits the necessity for an explanation to present causal attribution [26]: 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.

Tailored and contrastive 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. Also, Miller [26] clearly highlights the importance of contrastive explanations, because people may in general be not as interested in the causes of selecting the solution $\widehat{aud}$ per se, as they are in the causes of not selecting their initial preference aud_k. Therefore, given the redundancy of reporting ELVIRA's entire PR process, we suggest that the agent includes in the explanation only the elements that regard $\widehat{aud}$ in relation to aud_k. Whenever $\widehat{aud} = aud_k$, ELVIRA simply reports the AS instantiated for aud_k. On the other hand, whenever $\widehat{aud} \neq aud_k$, ELVIRA includes in the explanation the positive answers to the CQs discussed during the choice of action, stressing in particular on the differences between $\delta(q_0, \text{agreement}_{\widehat{aud}}, Av_k)$ and $\delta(q_0, q_0, Av_k)$. Note that our decision of what to include in the explanation 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 Sec. 4.1. This is, in fact, a very interesting follow-up future work.

Running Example. ELVIRA presents the following output to Kay: '*Conflict*: The sharing preferences of the other people involved do not coincide with yours. You suggested to share with $\langle 3, 6\rangle_a$; Dan opted for sharing with $\langle 2, 6\rangle_a$ and Zoe would like to share with $\langle 2, 8\rangle_a$. *Solution*: To share with $\langle 2, 8\rangle_a$ is the best compromise that solves the conflict because it satisfies as much as possible everyone's preferences. Notice that, by sharing with $\langle 3, 6\rangle_a$ (your initial sharing choice), other users may experience negative consequences, you would not make other users happy and you would not preserve everyone's privacy'. This is a *contrastive* explanation, which highlights how the user's preference is worse than the recommended solution because it would demote values (in this case, ST and CO) that would not be demoted if selecting the recommended solution. ELVIRA shows a similar output to Zoe, where the last part would be: 'Notably, by selecting to share with $\langle 2, 8\rangle_a$, you would get your way and preserve everyone's privacy', which is a *tailored*, not contrastive explanation for selecting $\langle 2, 8\rangle_a$.

4.3 Formal Properties

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 users and their preferences are treated equally, thus satisfying *role-agnosticism*.

LEMMA 1 (**SOUNDNESS**). *The audience recommended by ELVIRA is always optimal, i.e. the one which is the most coherent with everyone's utility and value preferences.*

PROOF SKETCH. Disregarding the roles of uploader or co-owner, ELVIRA identifies the audience to recommend by going through the PR described in Sec. 4.1. The agent can recommend only an audience that has collected only negative answers to the CQs: therefore, such audience must present the maximum overall score and the best individual value promotion. This means that there is no other action that is more coherent with everyone's preferences, and this makes the recommendation optimal. □

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 SKETCH. If the optimal audience exists, i.e. has the maximum overall score and the best individual value promotion, then ELVIRA will collect only negative answers to the CQs in the choice of action. Hence, the optimal audience will be the successful output of the PR and ELVIRA will recommend it 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 SKETCH. Anonimity is provided by Eq. (3), where the commutative property guarantees the sum of the individual scores to be independent of their order of aggregation, and by CQ2 of the PR process, where the values of all users are considered equally. □

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 SKETCH. When performing PR, ELVIRA instantiates the AS for every possible audience, and all the audiences are considered when discussing the CQs. Therefore, the order of consideration of the audiences is irrelevant. □

5 EVALUATION THROUGH 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 Sec. 1, that considering both utility and values to compute a solution to MPC is informed by empirical evidence [18, 46]. In particular, we present a comparative evaluation of ELVIRA (EL) and three other models inspired by the related work approaches (see Sec. 2) 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 only utility-driven;
- *Value-based* (VB): selects the audience that maximises the promotion of values for all the involved users, similar to works only value-driven;
- *Facebook* (FB): selects the uploader's preferred audience, i.e., neither utility- or value-driven.

We analysed the performance of these models on real data (portions of Facebook, as detailed later). To compare the models, we use[3] 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:

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

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 implemented the models in Python 2.7.10 (*numpy* 1.16.2; *networkx* 2.2) and we ran all our simulations on Windows 10. In each network, users were allocated to the nodes with a random value ordering, which was static for all the simulations, as informed by [5]; and intimacies were also generated randomly, in the range $[1, 5]$ as in [10], where 1 represents a mere acquaintance and 5 a very close relationship. 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[4]. Also, to generate audience aud_f, we randomly selected a distance from the range identified by the minimum and the maximum distance among the users' preferences; and we selected the intimacy similarly.

Simulation Settings. We report here the experiment with real portions of Facebook, but we conducted extensive experiments varying all parameters with similar results (see [31]). In particular, we used (number of nodes and edges in parenthesis): $G_1 = (769, 16656)$ and $G_2 = (1446, 59589)$ from [53], and $G_3 = (4039, 88234)$ from [21], and we generated T=500 MPCs among $n = 3$ random connected users on each of these graphs, with $\alpha = 0.9$ and $\beta = 0.1$.

Simulation Results. Fig. 3 displays the performance of the models in terms of *iauc* and *iavc*. Pairwise t-tests of EL with the other three models show significant differences between the distributions with p-value< .05 (as marked with '★' in the figure). The effect size of the comparison between the models is medium or large in all cases (average over the three graphs): (i) w.r.t. *iauc*, ELvsUT: -.4, ELvsVA: .36, ELvsFB: .23; w.r.t. *iavc*, ELvsUT: 1.45, ELvsVA: -.59, ELvsFB:

[3] For lack of space, we show only these metrics, but we also studied social, instead of individual, measures that provided equivalent results (cf. [31]).

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

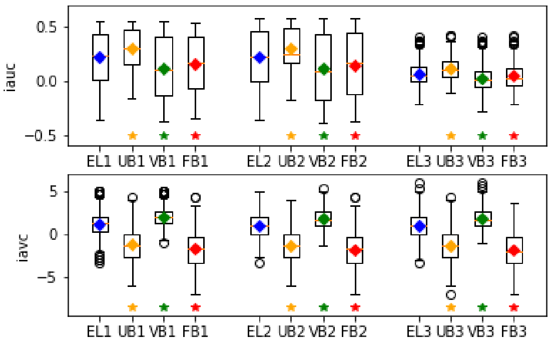

Figure 3: Models performance on real portions of OSN.

1.58. We can clearly see that UB always generated the maximum *iauc*, but guaranteed a poor promotion of moral values; VB always generated the maximum *iavc*, but with very low utilities; and EL represented the best utility-value trade-off, very close to UB in utility and to VB in values.

6 EVALUATION THROUGH USER STUDY

We now discuss the between-subjects user study 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 Sec. 4.1,4.2 allow ELVIRA to *convey* the recommendations in a more satisfactory way than existing approaches. Participants were recruited via Prolific[5], and the study received ethical approval by the Ethical Board of our university.

6.1 User Study Design

We developed a web application in Python to conduct the experiment. The application randomly assigned each participant to one treatment (as in the previous section): 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. We now describe the steps further. For the full specification of the experiment design, including the scenarios and questions presented to participants, see [31].

Values elicitation. We relied on the Portrait Value Questionnaire (PVQ) designed by Schwartz [42] to elicit the value preferences of the users. 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 [41] since 2002. The output of this part informs the ELVIRA and the value-based models about the participants' value preference.

MPCs. We followed an immersive scenario approach [24], which was successfully used in previous work in MPCs [8, 44]. For this, our application created six scenarios for each participant and presented them in a random order. Scenarios were composed of photos and descriptions taken from [8], in which the scenarios were proved to be representative of different sensitivities (low/high) and relationship types (colleagues, friends and family), and a randomly generated MPC as detailed next. For each scenario, the participant was asked to provide: (i) their preferred sharing policy[6] 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). Note therefore that even if the photos and descriptions were the same, the conflicts changed every time randomly, in practice meaning that there were many more than just six scenarios (for the same photo and description, each of two non-participant users could have one of 4 policies, one of 5 different appreciation levels, and one of 24 orders over values –see Supp.Mat. for details). 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. 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.

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). The output generated by ELVIRA correspond to the explanations discussed in Sec. 4.2. 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. Sec. 2). The Facebook model simulates what happens in Facebook: an uploader, randomly selected among the involved users, shares the picture with the uploader's preference. For further details about each model refer to [31]. Finally, to measure satisfaction, we used the Satisfaction Scale proposed in [12]. 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.).

Data Quality Measures. To maximise data quality, we employed two well-known methods: attention check questions, and participants' previous performance [11, 25, 35, 37]. We recruited participants from Prolific with at least 100 submissions and an approval rate of 95% according to [37]. Also, during the experiment, the application presented participants with three attention check questions (see [31]).

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

[6] Note that we used group-based policies, which, as aforementioned and shown in [44], 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 [27].

Table 4: Demographics of participants.

Age	‘18-25’: 32.6%, ‘26-35’: 31.3%, ‘36-45’: 18.9%, ‘46-55’: 9%, ‘55+’: 8.2%
Gender	‘Male’: 56.2%, ‘Female’: 43.3%, ‘Rather not say’: 0.5%
Country	‘UK’: 44.2%, ‘USA’: 15%, ‘Poland’: 9.4%, ‘Greece’: 5.6%, ‘Portugal’: 5.2%, ‘Canada’: 2.6%, other: 18%
Highest education	‘Grad degree’: 21%, ‘Undergrad degree’: 35.6%, ‘Tech/community college’: 10.3%, ‘Secondary education’: 28.4%, other: 4.7%
Social media use	‘Daily’: 85.4%; ‘2-3 times/week’: 9.9%; ‘Once a week’: 2.1%; ‘Less than once a week’: 2.6%
Privacy	‘Not concerned’: 3%; ‘Concerned’: 46.4%; ‘Very concerned’: 50.6%

6.2 User Study Results

We recruited 321 participants, who were rewarded £2.50 for completing the survey, which took on average 23.1 minutes (median 20.3 minutes). We discarded participants who failed at least one attention check question (27.4%), and analysed the remaining 233 participants. Tab. 4 reports the demographic distribution of the participants, including their privacy attitudes, measured with the IUIPC scale [23], and social media use. The final split per treatment (recall this was done randomly) was: 60 ELVIRA, 57 utility-based, 60 value-based, and 56 Facebook.

Acceptability of recommendation. Fig. 4 shows the distribution of individual and collective acceptability for each model (2=‘Very likely’, -2=‘Very unlikely’). The symbols on the bottom mark the distributions that are significantly worse than ELVIRA, when considering pairwise t-tests with p-value< .05 (★) and p-value< .1 (-) (effect size for individual acceptability: ELvsUT: .13, ELvsFB: .22; for collective acceptability: ELvsUT: .32, ELvsFB: .34). We can see that the recommendations generated by ELVIRA were significantly more accepted than those generated with utility-based or Facebook models. The value-based model shows a performance not significantly different from ELVIRA’s, but with a wider interquartile range for individual acceptability, including negative acceptability. For this reason, we sought to understand whether this higher proportion of participants providing negative acceptability for value-based was related to demographics. In particular, for older people (age≥46), for participants with at most secondary education, for users accessing social media less than daily, and for less privacy concerned individuals (IUIPC<.4), we found that ELVIRA’s recommendations were significantly (p-value< .05) more accepted than the value-based ones (differences in all other groups remained not significant).

Satisfaction of the output. Regarding the quality of the generated output, ELVIRA achieved by far the best performance. Fig. 5 shows the distribution of the answers to the Satisfaction Scale (2=‘Strongly agree’, -2=‘Strongly disagree’) , with significant differences marked as above (minimum effect size when marked with ‘★’ is .44; when marked with ‘-’ is .32). 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’, Q6: ‘The output that the tool provided are *useful to my goals*’, and Q7: ‘The output showed me how *accurate* the tool is’.

Discussion. Considering both the acceptability of the recommendations and the satisfaction with the model output, ELVIRA outperforms all 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, less educated, less privacy concerned and less active social media users, providing recommendations that are significantly less acceptable than ELVIRA’s for these groups.

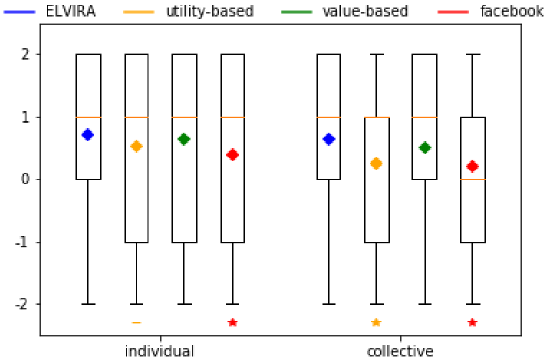

Figure 4: Individual and collective acceptability of the recommendations presented by each model.

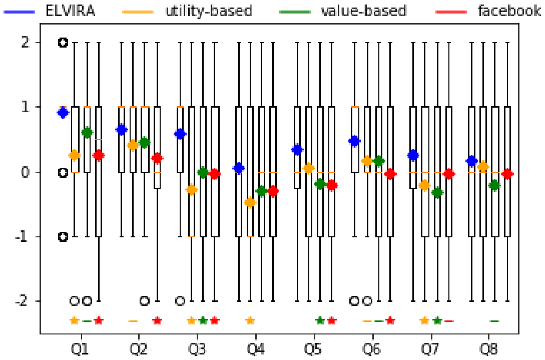

Figure 5: Evaluation of the outputs provided by each model, according to the Satisfaction Scale [12].

7 CONCLUSION

We introduced ELVIRA, the first agent to support collaborative multiuser privacy that meets all the requirements suggested by previous research and empirical evidence on multiparty privacy [36, 46]. As we proved in Sec. 4.3, ELVIRA is *role-agnostic* and *adaptive.* Then, through software simulations (see Sec. 5), we showed how the combination of a *utility-driven* component with a *value-based* one allows to reach a trade-off in terms of utility gain and value promotion that is better than the other models that have been suggested in the literature so far. The benefits of such a combination were also evident in a user study (see Sec. 6), where the solutions recommended by ELVIRA were considered generally more acceptable than for the other models. Finally, ELVIRA is able to generate and convey explanations which are confirmed to be satisfying by users.

Regarding future work, we plan to also tackle the much less frequent but severe cases of MPCs where malicious, non-collaborative behaviour may be present, e.g., revenge porn and cyber-bullying.

REFERENCES

[1] A. Acquisti, L. Brandimarte, and G. Loewenstein. 2015. Privacy and human behavior in the age of information. *Science* 347, 6221 (2015), 509–514.

[2] N. Ajmeri, H. Guo, P. K Murukannaiah, and M. P Singh. 2020. Elessar: Ethics in Norm-Aware Agents. In *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. 16–24.

[3] K. Atkinson and T. Bench-Capon. 2007. Practical reasoning as presumptive argumentation using action based alternating transition systems. *Artificial Intelligence* 171, 10-15 (2007), 855–874.

[4] K. Atkinson and T. Bench-Capon. 2018. Taking account of the actions of others in value-based reasoning. *Artificial Intelligence* 254 (2018), 1–20.

[5] A. Bardi and S.H Schwartz. 2003. Values and behavior: Strength and structure of relations. *Personality and social psychology bulletin* 29, 10 (2003), 1207–1220.

[6] A. Besmer and H.R. Lipford. 2010. Moving beyond untagging: photo privacy in a tagged world. In *CHI*. ACM, 1563–1572.

[7] B. Carminati and E. Ferrari. 2011. Collaborative access control in on-line social networks. In *CollaborateCom*. IEEE, 231–240.

[8] R. Fogues, P. Murukannaiah, J. Such, and M. Singh. 2017. Sharing Policies in Multiuser Privacy Scenarios: Incorporating Context, Preferences, and Arguments in Decision Making. *ACM TOCHI* 24, 1, Article 5 (2017), 29 pages.

[9] R. Fogues, P. Murukannaiah, J. Such, and M. Singh. 2017. SoSharP: Recommending Sharing Policies in Multiuser Privacy Scenarios. *IEEE IC* 21, 6 (2017), 28–36.

[10] R. Fogues, J. Such, A. Espinosa, and A. Garcia-Fornes. 2014. BFF: A tool for eliciting tie strength and user communities in social networking services. *Information Systems Frontiers* 16, 2 (2014), 225–237.

[11] D. J Hauser and N. Schwarz. 2016. Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior research methods* 48, 1 (2016), 400–407.

[12] R. R Hoffman, S. T Mueller, G. Klein, and J. Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[13] H. Hu, G.J. Ahn, and J. Jorgensen. 2011. Detecting and resolving privacy conflicts for collaborative data sharing in online social networks. In *ACSAC*. ACM, 103–112.

[14] M. Humbert, B. Trubert, and K. Huguenin. 2019. A Survey on Interdependent Privacy. *Comput. Surveys* (2019), 35.

[15] P. Ilia, I. Polakis, E. Athanasopoulos, F. Maggi, and S. Ioannidis. 2015. Face/Off: preventing privacy leakage from photos in social networks. In *CCS*. ACM Press, New York, New York, USA, 781–792.

[16] D. Kekulluoglu, N. Kökciyan, and P. Yolum. 2018. Preserving privacy as social responsibility in online social networks. *ACM TOIT* 18, 4 (2018), 42.

[17] N. Kökciyan, N. Yaglikci, and P. Yolum. 2017. An argumentation approach for resolving privacy disputes in online social networks. *ACM TOIT* 17, 3 (2017), 27.

[18] H. Krasnova, S. Spiekermann, K. Koroleva, and T. Hildebrand. 2010. Online social networks: Why we disclose. *JIT* 25, 2 (2010), 109–125.

[19] A. C Kurtan and P. Yolum. 2018. PELTE: Privacy estimation of images from tags. In *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*.

[20] A. Lampinen, V. Lehtinen, A. Lehmuskallio, and S. Tamminen. 2011. We're in it together: interpersonal management of disclosure in social network services. In *CHI*. ACM, 3217–3226.

[21] J. Leskovec and J.J. Mcauley. 2012. Learning to discover social circles in ego networks. In *NIPS*. 539–547.

[22] K. Liang, J. K Liu, R. Lu, and D. S Wong. 2014. Privacy concerns for photo sharing in online social networks. *IEEE Internet Computing* 19, 2 (2014), 58–63.

[23] N. K Malhotra, S. S Kim, and J. Agarwal. 2004. Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information systems research* 15, 4 (2004), 336–355.

[24] C. Mancini, Y. Rogers, A. K. Bandara, T. Coe, L. Jedrzejczyk, A. N. Joinson, B. A. Price, K. Thomas, and B. Nuseibeh. 2010. Contravision: exploring users' reactions to futuristic technology. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 153–162.

[25] W Mason and S Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods* 44, 1 (2012), 1–23.

[26] T. Miller. 2018. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* (2018).

[27] G. Misra and J. Such. 2016. How socially aware are social media privacy controls? *IEEE Computer* 49, 3 (2016), 96–99.

[28] G. Misra and J. Such. 2017. PACMAN: Personal Agent for Access Control in Social Media. *IEEE Internet Computing* 21, 6 (2017), 18–26.

[29] F. Mosca. 2020. Value-Aligned and Explainable Agents for Collective Decision Making: Privacy Application. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*. 2199–2200.

[30] F. Mosca, Ş. Sarkadi, J. Such, and P. McBurney. 2020. Agent EXPRI: Licence to Explain. In *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*. Springer, 21–38.

[31] F. Mosca and J. Such. 2021. An Explainable Assistant for Multiuser Privacy. *Working Paper* (2021).

[32] F. Mosca, J. Such, and P. McBurney. 2019. Value-driven Collaborative Privacy Decision Making. In *AAAI PAL Symposium*.

[33] F. Mosca, J. Such, and P. McBurney. 2020. Towards a Value-driven Explainable Agent for Collective Privacy. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*. 1937–1939.

[34] A-M. Olteanu, Ké. Huguenin, I. Dacosta, and J-P Hubaux. 2018. Consensual and privacy-preserving sharing of multi-subject and interdependent data. In *Proceedings of the 25th Network and Distributed System Security Symposium (NDSS)*. Internet Society, 1–16.

[35] L. J Paas and M. Morren. 2018. Please do not answer if you are reading this: Respondent attention in online panels. *Marketing Letters* 29, 1 (2018), 13–21.

[36] F. Paci, A. Squicciarini, and N. Zannone. 2018. Survey on access control for community-centered collaborative systems. *Comput. Surveys* 51, 1 (2018).

[37] E. Peer, J. Vosgerau, and A. Acquisti. 2014. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior research methods* 46, 4 (2014), 1023–1031.

[38] S. Rajtmajer, A. Squicciarini, C. Griffin, S. Karumanchi, and A. Tyagi. 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)*. 680–688.

[39] S. Rajtmajer, A. Squicciarini, J. Such, J. Semonsen, and A. Belmonte. 2017. An Ultimatum Game Model for the Evolution of Privacy in Jointly Managed Content. In *GAMESEC*. Springer, 112–130.

[40] M. Rokeach. 1973. *The nature of human values.* Free press.

[41] S. H. Schwartz. 2003. A proposal for measuring value orientations across nations. *Questionnaire package of the european social survey* 259, 290 (2003), 261.

[42] S. H. Schwartz. 2012. An overview of the Schwartz theory of basic values. *Online readings in Psychology and Culture* 2, 1 (2012), 11.

[43] A. Squicciarini, M. Shehab, and F. Paci. 2009. Collective privacy management in social networks. In *WWW*. ACM, 521–530.

[44] J. Such and N. Criado. 2016. Resolving Multi-Party Privacy Conflicts in Social Media. *IEEE TKDE* 28, 7 (2016), 1851–1863.

[45] J. Such and N. Criado. 2018. Multiparty Privacy in Social Media. *Commun. ACM* 61, 8 (2018), 74–81.

[46] J. Such, J. Porter, S. Preibusch, and A. Joinson. 2017. Photo privacy conflicts in social media: a large-scale empirical study. In *CHI*. ACM, 3821–3832.

[47] J. Such and M. Rovatsos. 2016. Privacy Policy Negotiation in Social Media. *ACM TAAS* 11, 1 (2016), 1–29.

[48] C Tessier, L Chaudron, and H-J Müller. 2006. *Conflicting agents: conflict management in multi-agent systems*. Vol. 1. Springer Science & Business Media.

[49] K. Thomas, C. Grier, and D. Nicol. 2010. Unfriendly: Multi-party privacy risks in social networks. In *PET*. Springer, 236–252.

[50] O. Ulusoy and P. Yolum. 2019. Emergent Privacy Norms for Collaborative Systems. In *PRIMA*. Springer, 514–522.

[51] O. Ulusoy and P. Yolum. 2020. Norm-based Access Control. In *Proceedings of the 25th ACM Symposium on Access Control Models and Technologies*. 35–46.

[52] N. Vishwamitra, Y. Li, K. Wang, H. Hu, K. Caine, and G.J. Ahn. 2017. Towards pii-based multiparty access control for photo sharing in online social networks. In *SACMAT*. ACM, 155–166.

[53] B. Viswanath, A. Mislove, M. Cha, and K.P. Gummadi. 2009. On the evolution of user interaction in facebook. In *Proceedings of the 2nd ACM workshop on Online social networks*. ACM, 37–42.

[54] P. Wisniewski, H. Lipford, and D. Wilson. 2012. Fighting for my space: Coping mechanisms for SNS boundary regulation. In *CHI*. ACM, 609–618.