Agent EXPRI: Licence to Explain

Francesca Mosca*, Ştefan Sarkadi, Jose M. Such, and Peter McBurney

King's College London
Department of Informatics
{francesca.mosca,stefan.sarkadi,jose.such,peter.mcburney}@kcl.ac.uk

Abstract. Online social networks are known to lack adequate multi-user privacy support. In this paper we present EXPRI, an agent architecture that aims to assist users in managing multi-user privacy conflicts. By considering the personal utility of sharing content and the individually preferred moral values of each user involved in the conflict, EXPRI identifies the best collaborative solution by applying practical reasoning techniques. Such techniques provide the agent with the cognitive process that is necessary for explainability. Furthermore, the knowledge gathered during the practical reasoning process allows EXPRI to engage in contrastive explanations.

Keywords: Multi-user Privacy · Practical Reasoning · Explainable AI.

1 "*EXPRI, agent EXPRI*": Introduction

Online collaborative platforms have recently generated an increasing concern for individual privacy. 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. Among other platforms, online social networks (OSNs) have proved to be particularly unsuitable to manage access control in a satisfying way for the users [6, 36, 12]. A common example of MPC in the literature is the case of a picture representing a group of friends, where each of them would assign different degrees of publicity/privacy to the picture on the OSN. Currently, most platforms lack built-in mechanisms that allow the users to discuss and agree on a policy in advance [42], and the responsibility of selecting a policy is generally left solely to the uploader. The other involved users, if unhappy with the uploader's choice, can only resort to unsatisfying reparative solutions, such as untagging or asking for the content to be removed.

MPCs happen frequently, with a majority of users having experienced at least one MPC [37]. However, generally users have collaborative attitudes, e.g. in a recent study [37] most uploaders wished to have known in advance the consequences of their decisions in order to tackle the conflicts before they occurr.

Previously, in [21], we have outlined an agent architecture to assist users during MPCs. In this paper we define further this agent architecture, that we

* Corresponding author.

now name EXPRI, especially in regard of its explainable component. EXPRI is an agent that aims to help users solve MPCs in OSNs by computing the optimal sharing policy for all the involved users. Optimality is considered in terms of (i) the utility that each user gains from sharing the content with a particular audience, and (ii) the promotion of moral values, i.e. the degree of coherency with the individual morality of choosing each possible policy. Several previous studies pointed out the necessity for autonomous systems to fully support users in privacy decisions (and not only), they need to be transparent and explainable [34, 23]. We show how EXPRI, which follows practical reasoning techniques to identify the optimal action, is fully equipped with the necessary information to provide a justification for the optimal action to the end-user; then we present a collection of starting points to inspire the development of the social process of EXPRI, e.g. how to best convey a justification to the end-user.

2 "*For Your Eyes Only*": Related Work in Privacy

In recent years, models for better supporting users to collaboratively deal with MPCs have been proposed in the related literature. We refer the interested reader to more comprehensive surveys like [23, 12, 36] for further details on multiuser privacy management. Researchers focused on the achievement of desirable properties [21], such as role-agnosticism, adaptability, explainability, and value- and utility-orientation. Given the aim of this paper, we particularly focus on what previous works have achieved in terms of explainability.

The ability of a system to be able to explain itself and justify its outputs is generally considered crucial for fostering the users' trust towards autonomous systems [34, 23]. Of course, this is also valid in the context of multiuser privacy. The running hypothesis is that, by interacting with explainable systems, users will find it easier to understand the received recommendations and, consequently, to endorse them, notably reducing the occurrence of MPCs.

The approaches suggested for this type of application range from game theoretical solutions [32, 27], to agent-based ones [38, 35, 22], to learning models [10, 40], and more technical, fine-grained systems [13]. Despite the abundance of efforts, none of these approaches can be considered fully explainable. However, some solution-concepts, like argumentation-based models, make it easier than others to meet the explainability requirement. In [14] each user of the OSN is represented by an agent that captures its user's privacy constraints through ontologies and semantic rules. When MPCs occur, the agents interact in persuasion dialogues to defend their privacy preferences. The arguments generated in the dialogue can be reported to the users as a justification of the output, even though the best way to do so is not investigated by the authors. In [9], the authors design a recommendation system, where the prediction of the optimal collective sharing policy is based on the scenario's context, the users' preferences, and their arguments about those preferences. A limited set of arguments is considered, leaving unclear their efficacy in convincing the users, but providing the first steps towards an explanation of the system's decision.

Our work differs from the literature because it presents a design that is explicitly oriented towards the provision of an explanation. We consider the transparency of the process as the main feature of our model, where the provided explanation is crucial and not just an accessory of the model.

3 "*A View to an EXPRI*": the Agent Architecture

In this section we detail the components of our agent architecture, EXPRI. We instantiate an agent EXPRI for each user registered to the OSN: each agent supports a user while taking decisions on multiuser privacy. In order to do so, the agent needs to be aligned with the user's preference, in terms of (i) *utility* and (ii) *moral values*. In fact, users are reported to share content online for personal advantage [15], but they may also consider the consequences of their decisions and transcend their own benefit to accommodate others' preferences [37].

3.1 Utility-driven Component of EXPRI

We represent a OSN as a graph $G = (Ag, R)$, where Ag is the set of all the registered agents/users a_k, and R is the set of all their relationships (a_k, a_j, i_{kj}), with i_{kj} being the *intimacy* of the relationship between the users a_k and a_j. We consider *intimacy* as defined in [11], where the authors present also a way to elicit it automatically.

We assume that every user has an individual preference in terms of publicity/privacy for sharing content, that can be elicited automatically for each item or a collection of them [33, 20]. We define the concepts below for each individual content x, even though for simplifying the notation we do not always report x.

Definition 1. *The user a_k defines the **sharing policy** $sp_k = \langle d, i \rangle$ for the item x, meaning that a_k wants to allow access to x to any other user who is distant at most d and intimate at least i.*

Definition 2. *The **individual audience** for the user a_k is the set $aud_{sp,k}$ of users who satisfy the conditions set by a sharing policy sp for the content x.*

Definition 3. *The **collective audience** is the set $aud_{sp} = \bigcap_{k \in Ag} aud_{sp,k}$, that is the intersection of the individual audiences of all the involved users generated by the sharing policy sp for the content x.*

A multi-user privacy conflict (MPC) occurs whenever two or more users, who are involved in the same content, have contrasting preferred sharing policies, i.e. their preferred individual audiences do not coincide.

As discussed in [15], we believe that users perceive some type of benefit when sharing an appealing photo online, but they can also experience some discomfort whenever a picture is seen by undesired people. We refer to this advantage and disadvantage in terms of gain or loss of *utility*. Furthermore, in order to find a compromise to solve the MPC, users may be more inclined to overshare or undershare, that is to make the content available to a broader or smaller audience than the preferred one.

Definition 4. *The function* ***appreciation*** *determines whether the user prefers to overshare ($app_k(x) = 1$) or undershare ($app_k(x) = -1$) the item x.*

Definition 5. *Given the preferred audience aud_k and a sharing policy sp, if sharing with the collective audience aud_{sp}, the individual* ***utility*** *of user a_k varies according to:*

$$u_{aud_{sp},k} = \sum_{j \in allDesAud} \frac{i_j}{d_j} - \sum_{j \in excDesAud} \frac{i_j}{d_j} + app_k(x) \sum_{j \in allExtAud} \frac{i_j}{d_j}, \quad (1)$$

where allDesAud (allowed desired audience) is the set of users who a_k desires to grant access to x and that are part of aud_{sp}; excDesAud (excluded desired audience) is the set of users who a_k desires to grant access to but that are excluded by aud_{sp}; allExtAud (allowed extra audience) is the set of users who a_k desires to forbid access from but that are part of aud_{sp}.

Users perceive a gain in utility whenever approved people access the content, but they can lose utility if undesired people access the content (if *appreciation* is negative) or desired people are excluded. Also, these effects get amplified with people that are closer and more intimate, as reported in recent user studies [37].

Example Let us consider the simplified OSN in Figure 1. Alice, Bob and Charlie appear together in the picture x. Their preferred sharing policies 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 preferred individual audiences: $aud_A = \{A, B, C, D, E, F, G, I, L\}$, $aud_B = \{A, B, C, D, G\}$ and $aud_C = \{A, B, C, G, I\}$. A conflict occurs, because the three preferred individual audiences do not coincide. The collective audiences resulting from the intersection of the individual ones generated by each policy are $aud_{sp_A} = \{A, B, C, D, E, G, I, L\}$, $aud_{sp_B} = \{A, B, C\}$ and $aud_{sp_C} = \{A, B, C, G, I\}$. Furthermore, Alice and Bob have a positive appreciation for x ($app_A(x) = app_B(x) = 1$), while for Charlie it is negative ($app_C(x) = -1$).
Let us consider $sp' = \langle 2, 3 \rangle$ as a possible sharing policy for x: the collective audience generated by sp' is $aud_{sp'} = \{A, B, C, D, G, I\}$ (that we rename as aud' for brevity). Then, Alice, Bob and Charlie would perceive the following variation in utility:

$$u_{aud',A} = \sum_{j \in \{B,C,D,G,I\}} \frac{i_j}{d_j} - \sum_{j \in \{E,F,L\}} \frac{i_j}{d_j} = \frac{5}{1} + \frac{4}{1} + \frac{3}{1} + \frac{10}{2} + \frac{9}{2} - \frac{2}{1} - \frac{6}{2} - \frac{2}{1} = 14.5$$

$$u_{aud',B} = \sum_{j \in \{A,C,D,G\}} \frac{i_j}{d_j} + \sum_{j \in \{I\}} \frac{i_j}{d_j} = \frac{5}{1} + \frac{3}{1} + \frac{3}{1} + \frac{5}{1} + \frac{8}{2} = 20$$

$$u_{aud',C} = \sum_{j \in \{A,B,G,I\}} \frac{i_j}{d_j} - 1 \cdot \sum_{j \in \{D\}} \frac{i_j}{d_j} = \frac{4}{1} + \frac{3}{1} + \frac{8}{2} + \frac{5}{1} - \frac{7}{2} = 12.5$$

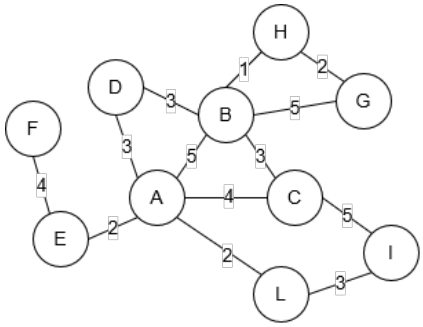

Fig. 1: The simplified online social network discussed in the example.

Table 1: Interpretation of the Schwartz values in the multi-user application and details of their promotion and demotion for a user, comparing different sharing options with own preference.

Value	Interpretation	Condition
OTC	appreciate compromises which differ from anyone's initial preference	+ if sharing with no one's initial preference - if sharing with another user's preference
CO	preserve individual and social security	+ if sharing with a smaller audience - if sharing with a bigger audience
ST	do what is good for the other people	+ if sharing with the other's preference + if compromising with the other user - if ignoring the other user's preference - if rejecting an offer
SE	maintain or increase one's own utility	+ if sharing with own preference + if gaining a better utility - if gaining a worse utility

3.2 Value-Aligned Component of EXPRI

We base the moral component of EXPRI on the Schwartz Theory of Basic Values [30]. This is the most well-known and established theory of human values and combines a complete theoretical architecture with a strong empirical validation.

Values are socially desirable concepts representing the mental goals which drive human behaviour [30], influencing any decision. In particular, the Schwartz theory presents ten main values, organised along four directions (which we refer to as $\mathcal{V}$) that pull apart. 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), which characterises authoritarian and image-conscious conducts. The individual preference over these values, which is considered relatively stable over time [5], can be elicited from the users through validated questionnaires [30].

Given the MPC application, we interpret the Schwartz value-directions as described in Table 1, where we also report how the user's behaviour can promote or demote the values. However, this is just for illustrative purpose, because our

agent architecture for solving conflicts can be adapted to any other value theory or application. We believe that the agent can suggest solutions to the MPC that are more compatible with the user's morale if it is informed about the user's preferred order over the values. Hence, while reasoning about possible solutions to the conflict, EXPRI considers the *value promotion* of a sharing policy.

Definition 6. *Given a user a_k and her preferred order o_k over $\mathcal{V}$, the **value promotion** of an audience aud for the user a_k is given by*

$$v_{aud,k} = \sum_{i=1}^{|\mathcal{V}|} (I - i) \cdot prom_{aud}(o_i), \tag{2}$$

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.

Running Example Alice, Bob and Charlie's preferred orders over the values $\mathcal{V}$ are, respectively:

$$OTC \prec_A SE \prec_A CO \prec_A ST$$
$$CO \prec_B OTC \prec_B ST \prec_B SE$$
$$ST \prec_C CO \prec_C OTC \prec_C SE$$

The selection of $aud' = aud_{sp'} = \{A, B, C, D, G, I\}$ generates the following individual value promotions:

$$v_{aud',A} = +4 - 3 + 2 + 1 = 4$$
$$v_{aud',B} = -4 + 3 + 2 + 1 = 2$$
$$v_{aud',C} = +4 - 3 + 2 - 1 = 2$$

Alice promotes every value but SE, Bob promotes every value but CO, and Charlie promotes only ST and OTC.

3.3 Resolution of MPCs

Each EXPRI agent can cover two roles in the resolution of a MPC: *uploader*, when the user wants to share some content online, and *co-owner*, when the user is involved in some content that another user wants to share. Let us recall that we consider a non-adversarial setting: therefore, we assume that the agents cooperate in order to identify a collectively satisfying solution to the MPC. In fact, empirical studies showed how users are frequently willing to find acceptable compromises; in particular, uploaders reported to wish to have known in advance the preferences of the co-owners, to avoid conflicts before their occurrence [37].

For each conflict involving n users, a set $\mathcal{A}$ of at most $n + 1$ audiences are considered as possible solutions: the collective audiences aud_{sp_k} generated by each individually preferred sharing policy sp_k, and aud', generated in such a way that $aud' \neq aud_{sp_k} \forall k$.

Independently of the role, all agents compute an *individual score* for each possible solution, expressing their appreciation of the option in terms of both utility and value promotion:

$$s_{aud,k} = \begin{cases} +|u_{aud,k}| \cdot |v_{aud,k}| & \text{if } u_{aud,k} > 0 \wedge v_{aud,k} > 0, \\ -|u_{aud,k}| \cdot |v_{aud,k}| & \text{otherwise.} \end{cases} \tag{3}$$

Then, each EXPRI-coowner shares with EXPRI-uploader its individual scores: in this way, each agent collaborates without directly disclosing its potential gain in terms of neither utility nor value promotion. Then, EXPRI-uploader aggregates all the individual scores into a *collective score* for each possible solution:

$$s_{aud} = \sum_{k \in Ag} s_{k,aud}. \tag{4}$$

Finally, EXPRI-uploader identifies the most desirable solution, through the process we describe in the next section, and suggests it to the EXPRI-coowners. The EXPRI-coowners also perform a similar reasoning process to decide whether to accept or reject the offer. At the end of the deliberations, the outcomes can be reported to the users: we discuss possible guidelines to do this in section 5.

Running Example Table 2 reports the individual utilities and value promotions for each agent and each possible audience. The details of the computations are reported, for instance regarding the utilities and the value promotions for sp', in the previous examples.

Table 2: Individual and collective metrics for the scenario in the example.

agents	aud_{sp_A}			aud_{sp_B}			aud_{sp_C}			$aud_{sp'}$		
	u	v	s	u	v	s	u	v	s	u	v	s
A	22.5	0	0	-10.5	-4	-42	8.5	-4	-34	14.5	+4	58
B	27	-4	-108	0	0	0	14	-4	-56	20	+2	40
C	6.5	-2	-13	2	+4	8	16	-5	-64	12.5	+2	25
collective			-121			-34			-154			123

4 "*From Practical Reasoning with Love*": Design of the Cognitive Process

As Miller discusses in [19], an *explanation* is composed by 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 interactive process of transferring knowledge between the explainer and the explainee. In this section we describe how techniques from computational argumentation can be applied in order to provide EXPRI with a cognitive process that allows the agent to gather the

Table 3: Detail of the joint actions J_{Ag}, for each $aud_i \in \mathcal{A}$, and the partial transition function τ when $n = 3$ in a MPC scenario.

J_{Ag}	τ
$j_{1-4} = \langle \text{offer}_{aud_i}, \text{reject}_{2,aud_i}, \text{reject}_{3,aud_i} \rangle$	$\tau(\text{conflict}, j_{1-4}) = \text{conflict}$
$j_{5-8} = \langle \text{offer}_{aud_i}, \text{accept}_{2,aud_i}, \text{reject}_{3,aud_i} \rangle$	$\tau(\text{conflict}, j_{5-8}) = \text{conflict}$
$j_{9-12} = \langle \text{offer}_{aud_i}, \text{reject}_{2,aud_i}, \text{accept}_{3,aud_i} \rangle$	$\tau(\text{conflict}, j_{9-12}) = \text{conflict}$
$j_{13-16} = \langle \text{offer}_{aud_i}, \text{accept}_{2,aud_i}, \text{accept}_{3,aud_i} \rangle$	$\tau(\text{conflict}, j_{13-16}) = \text{agreement}_{aud_i}$

necessary information in order to justify to the user the selection of the optimal solution to the MPC.

We start by considering that an argument scheme (AS) and its associated critical questions can enable an agent to propose, attack and defend justifications for a given action [3]. In the following we adapt the AS, that was introduced by Atkinson, to EXPRI-uploader (AS-U) and EXPRI-coowners (AS-C). For AS-U, this would take the form of *"Given the current conflict, I should offer the audience aud, that will be accepted by the co-owners and therefore will solve the conflict, that will provide the score s_{aud} and that will promote my values V"*. Symmetrically, AS-C results to be: "*Given the current conflict, I should accept the audience aud and solve the conflict, to get the score s_{aud} and to promote my values V*" [1]. An agent who does not accept this presumptive argument, can challenge it by presenting *critical questions* (CQs), formally described in [3]. Unfavourable answers to the CQs provide attacks to the original argument. Attacks can be directed to different elements of the argument, i.e. to the different stages of the *practical reasoning* (PR) which led to such conclusion. In line with [2, 3], in the remaining part of this section we present the three stages of the practical reasoning (PR) process for the agent EXPRI, namely (i) the *problem formulation*, (ii) the *epistemic stage*, and (iii) the *choice of action*.

4.1 Problem Formulation

The first step of PR consists of representing the relevant elements of the situation (i.e. conflict occurrence, possible solutions, involved users' preferences, etc.). We perform this task by building an Action-Based Alternating Transition Systems with Values (AATS+V) [3]. This structure provides the underlying semantics that we use to describe the world and formulate the arguments about action, in particular when the outcome of an individual action (e.g. for the uploader to offer some particular solution) depends on what the other agents decide to do (e.g. whether the co-owners accept or reject the uploader's offer). We refer to this case as *joint actions* (J_{Ag}), i.e. actions that are performed at the same time[2] by a set of agents. For clarity, in Table 3 we show the possible joint actions in

[1] Note that the definition of "values" in [3] is based on [26], which is different from that of Schwartz [31] that we use in this paper.

[2] Similarly to [4], we assume the offer and the response to be a "simultaneous" action, despite its sequentiality.

the MPC scenario with $n = 3$ agents involved: we reported all the combinations of individual actions that are available to each agent, i.e. $offer_{aud_i}$ for the uploader, and $accept_{aud_i}$ and $reject_{aud_i}$ for the co-owners, referring to all the possible audiences $aud_i \in \mathcal{A}$. We adapt for the MPC scenario the definition of AATS+V given in [3].

Definition 7. *In the context of a MPC among n users, an* ***AATS+V*** *is a* $2n + 8$ *tuple* $\Sigma = \langle Q, q_0, Ag, Ac_1, ..., Ac_n, \rho, \tau, S, \mathcal{V}, Av_1, ..., Av_n, \delta \rangle$*, where:*

- $Q = \{conflict,\ agreement_{aud} \quad \forall aud \in \mathcal{A}\}$ *is a finite, non-empty set of states;*
- $q_0 = conflict$ *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;*
- $Ac_1 = \{offer_{aud} \quad \forall aud \in \mathcal{A}\}$ *are the actions available to the agent* up_1*;*
- $Ac_k = \{accept_{k,aud}, reject_{k,aud} \quad \forall aud \in \mathcal{A}\}$ *are the actions available to the agent* co_k*, for* $k = 2...n$*;*
- $\rho : Ac_{Ag} \rightarrow 2^Q$ *is the action-precondition function, which defines the set of states from which an action* $ac \in Ac_{Ag}$ *can be executed:* $\rho(offer_{aud}) = \rho(accept_{aud}) = \rho(reject_{aud}) = conflict$*;*
- $\tau : Q \times J_{Ag} \rightarrow Q$ *is the partial system transition function, which defines what state results from performing the joint action* j *in the state* q*, where possible (see the case with* $n = 3$ *in Table 3);*
- $S = \{0, s_{aud} \quad \forall aud \in \mathcal{A}\}$ *is the set of collective scores characterising each state, where* $s_{conflict} = 0$*;*
- $\mathcal{V} = \{SE, ST, CO, OTC\}$ *is the set of values considered by each agent;*
- $Av_k = o_k(\mathcal{V})$ *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 of each agent (see Table 1).*

Running Example Considering the scenario described in the previous examples and Table 3, the first step of the reasoning process for EXPRI-uploader, that represents Alice, consists of the problem formulation given by the AATS+V in Figure 2. Note that each agent knows only its own value preference (therefore the evaluation of δ); however, in the figure we represent all the promoted and demoted values for completeness: δ_A is in red, δ_B in blue, and δ_C in green.

4.2 Epistemic Stage

The epistemic stage consists of determining what the agent believes about the current situation, given the previous problem formulation. Let us recall our assumption, based on empirical evidence [37], that the EXPRI agents have a collaborative and non-adversarial behaviour. From this underlying assumption we can further imply other two fundamental epistemic assumptions:

- EA1 (for all the agents): all agents interpret the world in a similar manner: hence, all the agents have the same knowledge regarding all the components of the AATS+V, the only exception being Av_k. In fact, in order to preserve

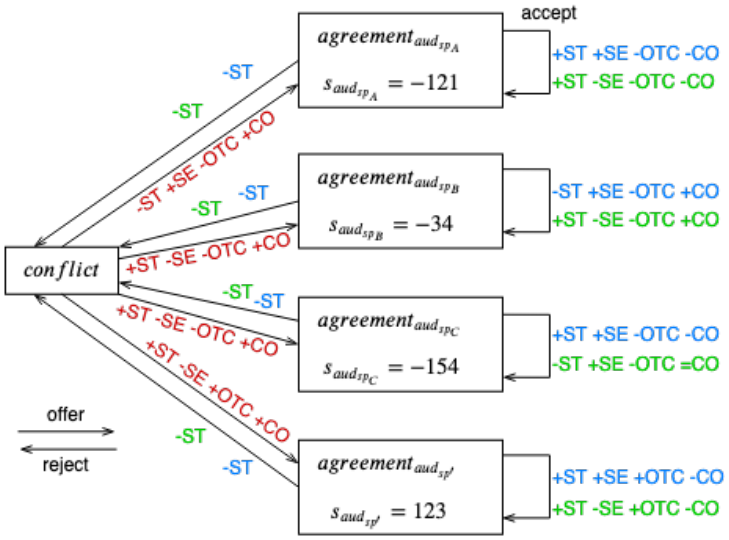

Fig. 2: Problem formulation through AATS+V for the scenario in example.

even further the privacy of the users involved in the MPC, we assume that each agent k only knows its own preferred order Av_k over the values and is uninformed about any other Av_j for $j \neq k$.

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

With reference to the CQs in [3], because of EA1, we are not interested in the CQs that are related to the problem formulation (CQ2-4 and CQ12-16) and its truthfulness (CQ1). Because of EA2, we are able to evaluate appropriately CQ17 when instantiated for each possible argument.

4.3 Choice of Action

The last step of the PR is the choice of action, that is the development of a value-based argumentation framework (VAF), which instantiates an appropriate argumentation scheme, and the consequent evaluation of the arguments according to the preference of values.

We focus in particular on AS-U and AS-C, and the critical questions which contest the optimality of the identified action, i.e. to offer *aud* for the EXPRI-uploader, and to accept or reject *aud* for the EXPRI-coowners. We refer to the questions CQ5-CQ11 from [3]: CQ5, CQ6 and CQ7 offer alternative actions that realise the same consequence, goal and value promotion; CQ8, CQ9 and CQ10 consider unacknowledged side effects, such as demotion of desired values or promotion of other values; and, finally, CQ11 wonders whether there is any other action that is more desirable in terms of values promotion.

The collection of negative answers to these CQs provides the justification for action. We argue that this abductive reasoning is sufficient to define the causal attribution of the recommended event and, therefore, the practical reasoning process can be equivalent to the cognitive process required for providing an explanation.

Running Example Let us analyse the process of choosing an action for the agent EXPRI acting on behalf of Alice, the uploader. Given the assumption of co-operative behaviour and the common goal of solving the MPC by reaching an agreement, EXPRI-uploader discards immediately the joint actions j_{1-12}: in fact, if at least one of the co-owners does not accept the offer, the conflict is guaranteed to persist. The uploader needs to identify the optimal audience to offer, i.e. the one that, if accepted by the co-owners like in j_{13-16} provides the best agreement. In order to do this, with reference to Figure 2, the uploader examines one by one its possibilities and checks whether they get challenged by any CQ. Note that we do not report a graphical representation of the VAF that would be generated in this process because of the high number of considered arguments; however, we detail the main arguments and all the attacks provided by the CQs (we leave implicit any supporting relationship).

- AS-U for j_{13}: "Given the current conflict, I should offer aud_{sp_A}, that will be accepted by the co-owners, to solve the conflict, to obtain the score $s_{A,aud_{sp_A}}$ and to promote SE and CO."
  - obj13.1: a better score can be achieved by performing alternative actions (CQ5): *successful*, e.g. in j_{14} and j_{16};
  - obj13.2: the agreement is reached also with alternative actions (CQ6): *successful*, e.g. in j_{14}, j_{15} and j_{16};
  - obj13.3: CO is promoted also with alternative actions (CQ7): *successful*, e.g. in j_{14}, j_{15} and j_{16};
  - obj13.4: ST is demoted (CQ9): *rejected*, because Alice cares more about SE and CO (here promoted) than ST;
  - obj13.5: OTC is demoted (CQ9): *successful*, OTC is the most preferred value for Alice;
  - obj13.6: other values can be promoted by performing alternative actions (CQ11): *successful*, e.g. +OTC in j_{16}, which Alice prefers to SE and CO;
  - obj13.7: EXPRI-Bob will not accept the offer (CQ17): *successful*, because $s_{B,aud_{sp_A}} \neq \max_{\mathcal{A}} s_{B,aud}$ and $s_{aud_{sp_A}} \neq \max_{\mathcal{A}} s_{aud}$;
  - obj13.8: EXPRI-Charlie will not accept the offer (CQ17): *successful*, because $s_{C,aud_{sp_A}} \neq \max_{\mathcal{A}} s_{C,aud}$ and $s_{aud_{sp_A}} \neq \max_{\mathcal{A}} s_{aud}$.
- AS-U for j_{14}: "Given the current conflict, I should offer aud_{sp_B}, that will be accepted by the co-owners, to solve the conflict, to obtain the score $s_{A,aud_{sp_B}}$ and to promote ST and CO."
  - obj14.1: a better score can be achieved by performing alternative actions (CQ5): *successful*, e.g. j_{16};
  - obj14.2: the agreement is reached also with alternative actions (CQ6): *successful*, e.g. in j_{13}, j_{15} and j_{16};
  - obj14.3: CO is promoted also with alternative actions (CQ7): *successful*, e.g. in j_{13}, j_{15} and j_{16};
  - obj14.4: ST is promoted also with alternative actions (CQ7): *successful*, e.g. in j_{15} and j_{16};

  - obj14.5: SE is demoted (CQ9): *successful*, SE is the second most preferred value for Alice;
  - obj14.6: OTC is demoted (CQ9): *successful*, OTC is the most preferred value for Alice;
  - obj14.7: other values can be promoted by performing alternative actions (CQ11): *successful*, e.g. +OTC in j_{16}, which Alice prefers to ST and CO;
  - obj14.8: EXPRI-Bob will not accept the offer (CQ17): *successful*, because $s_{B,aud_{sp_B}} \neq \max_{\mathcal{A}} s_{B,aud}$ and $s_{aud_{sp_B}} \neq \max_{\mathcal{A}} s_{aud}$;
  - obj14.9: EXPRI-Charlie will not accept the offer (CQ17): *successful*, because $s_{C,aud_{sp_B}} \neq \max_{\mathcal{A}} s_{C,aud}$ and $s_{aud_{sp_B}} \neq \max_{\mathcal{A}} s_{aud}$.

- AS-U for j_{15}: "Given the current conflict, I should offer aud_{sp_C}, that will be accepted by the co-owners, to solve the conflict, to obtain the score $s_{A,aud_{sp_C}}$ and to promote ST and CO."
  - obj15.1: a better score can be achieved by performing alternative actions (CQ5): *successful*, e.g. j_{16};
  - obj15.2: the agreement is reached also with alternative actions (CQ6): *successful*, e.g. in j_{13}, j_{14} and j_{16};
  - obj15.3: CO is promoted also with alternative actions (CQ7): *successful*, e.g. in j_{13}, j_{14} and j_{16};
  - obj15.4: ST is promoted also with alternative actions (CQ7): *successful*, e.g. in j_{14} and j_{16};
  - obj15.5: SE is demoted (CQ9): *successful*, SE is the second most preferred value for Alice;
  - obj15.6: OTC is demoted (CQ9): *successful*, OTC is the most preferred value for Alice;
  - obj15.7: other values can be promoted by performing alternative actions (CQ11): *successful*, e.g. +OTC in j_{16}, which Alice prefers to ST and CO;
  - obj15.8: EXPRI-Bob will not accept the offer (CQ17): *successful*, because $s_{B,aud_{sp_C}} \neq \max_{\mathcal{A}} s_{B,aud}$ and $s_{aud_{sp_C}} \neq \max_{\mathcal{A}} s_{aud}$;
  - obj15.9: EXPRI-Charlie will not accept the offer (CQ17): *successful*, because $s_{C,aud_{sp_C}} \neq \max_{\mathcal{A}} s_{C,aud}$ and $s_{aud_{sp_C}} \neq \max_{\mathcal{A}} s_{aud}$.

- AS-U for j_{16}: "Given the current conflict, I should offer $aud_{sp'}$, that will be accepted by the co-owners, to solve the conflict, to obtain the score $s_{A,aud_{sp'}}$ and to promote ST, OTC and CO."
  - obj16.1: the agreement is reached also with alternative actions (CQ6): *successful*, e.g. in j_{13}, j_{14} and j_{15};
  - obj16.2: CO is promoted also with alternative actions (CQ7): *successful*, e.g. in j_{13}, j_{14} and j_{15};
  - obj16.3: ST is promoted also with alternative actions (CQ7): *successful*, e.g. in j_{14} and j_{15};
  - obj16.4: SE is demoted (CQ9): *rejected*, because Alice cares more about OTC (here promoted) than SE;
  - obj16.5: other values can be promoted by performing alternative actions (CQ11): *rejected*, because SE (promoted in j_{13}) is less important to Alice than OTC, here promoted;
  - obj16.6: EXPRI-Bob will not accept the offer (CQ17): *rejected*, because $s_{B,aud_{sp'}} = \max_{\mathcal{A}} s_{B,aud}$ and $s_{aud_{sp'}} = \max_{\mathcal{A}} s_{aud}$;
  - obj16.7: EXPRI-Charlie will not accept the offer (CQ17): *rejected*, because $s_{C,aud_{sp'}} = \max_{\mathcal{A}} s_{C,aud}$ and $s_{aud_{sp'}} = \max_{\mathcal{A}} s_{aud}$.

AS-U for j_{13} is rejected, because all the attacks provided by the CQs are successful (note that obj13.4 is considered irrelevant because of Alice's values preference); similarly, the arguments for j_{14} and j_{15} are rejected. Regarding AS-U for j_{16}, we reject in a subsequent moment obj16.1, obj16.2 and obj16.3, because the suggested alternative actions are proved not to be as desirable as the current action (all their objections are successful).

In conclusion, EXPRI-uploader identifies j_{16} as the most desirable joint action and therefore suggests Alice to offer $aud_{sp'}$. The EXPRI-coowners go through a similar reasoning process, which we do not report in detail for lack of space, to identify the best individual action upon the uploader's offer.

5 "*The Cognitive Process is not enough*": Challenges for Designing the Social Process

So far we have showed how EXPRI is able to solve an MPC by identifying through practical reasoning the optimal solution for an MPC in OSNs. According to [25], abductive reasoning provides the best explanation given all available information. This means that, EXPRI's practical reasoning being an abductive form of reasoning, by reporting it, the agent can provide the best explanation for the given recommended action.

However, considering the social nature of explanations in AI [19], we have to address the very important distinction between *explainable* AI and *self-explainable* AI. An artificial agent can be explainable in the sense that humans can follow and understand its cognitive process, and by following this process, humans are able to explain why the agent is doing what it is doing. A self-explainable artificial agent, on the other hand, is a socially aware agent which has the capability of communicating explanations to the human that it interacts with. For reasons of trustworthiness [41], accountability [7], and responsibility [8], that have been mentioned in the literature, it is desirable for an agent to be self-explainable.

Both [19] and [16] 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 engage in counterfactual explanations, e.g. justifying the rejection of possible alternative actions; and (ii) to tailor the explanation according to the individual user's needs. In the following, we outline how EXPRI may be able to meet these requirements.

Contrastive explanations In [19], Miller clearly summarises the importance of providing contrastive explanations. Research shows that people are in general not as interested in the causes of an event per se, as they are in the relation of that event to some other event that did not occur. For instance, a user may wonder why EXPRI suggested action x rather than action y. An answer to this question might provide a more convincing explanation for the user than the simple motivation to choose x. As we detailed earlier, EXPRI's cognitive process

comprises of practical reasoning about alternative options through the discussion of critical questions. The process of accepting or rejecting each objection that arises from the CQs provides EXPRI with the necessary knowledge to justify why the action it suggests is the optimal one and why the alternatives are not as good as the optimal one. It follows that EXPRI is able to answer any interrogation that the user may conduct in terms of contrasting and comparing the other possible options.

Tailored explanations We are planning to give EXPRI the capability of providing explanations that are generated by taking into account the perspective of the interlocutor and/or interlocutors. Continuing on the path of using practical reasoning, it could be feasible to use AATS+V to reason about which is the optimal explanation depending on the social context in which the interaction between EXPRI and the given user takes place. To be able to do this, EXPRI could build an AATS+V taking into account the values and beliefs of the user, in order to be able to reason from the perspective of the user. This additional AATS+V is similar to a Theory-of-Mind (ToM) of the user [1]. However, the formation of this additional AATS+V that is to be used for finding the optimal explanation in social interactions is not as straightforward: the joint actions are not sets of uploader's offers and co-owners' responses anymore, but they represent subsets of dialogues. That is, EXPRI needs to be able to reason about elements such as speech acts, their implicatures, and how these elements change the beliefs and update the knowledge of the interlocutor during a dialogue. Similarly to [4], the epistemic stage now involves uncertainty, because EXPRI does not know what the user's reaction to its explanation may be. Argumentation Dialogue Games (ADGs) [18] provide an elegant way to address this issue. ADGs have recently been used for the formation and use of ToM through speech acts to reach states of shared beliefs with other agents [24], even under conditions of uncertainty [29], for dynamic story generation in interrogation games [28], as well as for providing protocols of interactive explanations to users [17]. Therefore, EXPRI could use ADGs to reason about how and what it communicates to the interlocutor in order to see what kind of explanation might emerge from a hypothetical interaction. Ideally, after going through various alternative dialogues, it would be able to select the dialogue it intends to have with the interlocutor, that will lead to the optimal explanation. EXPRI could also use ADGs to reason about what it tells the interlocutor and what the interlocutor understands in real time, by updating its ToM of the interlocutor based on what the interlocutor tells or asks EXPRI.

Explaining Conflicts Conflict management literature makes the distinction between three main components of a conflict in multi-agent systems, namely *conflict detection*, *conflict representation*, and *conflict resolution* [39]. Perhaps, in most cases it would be useful to give users a general overview of the context evolution of the MPC, explaining why and how a solution has been found or not. From a causal attribution perspective, it seems reasonable that conflict detection represents the cause of whose effect is represented by the conflict's resolution.

Therefore, in the case of an MPC, it would be desirable to have an explanation that not only guides the user from cause to effect, but also that describes to the user the cause and the effect [19]. In this way, the user can assess whether the agent that is providing the explanation has understood the context and has thus grounded the explanation in a realistic representation. Such a causal explanation guides the user from commonly established premises that describe the conflict's detection, to a valid conclusion that represents the solution, or lack thereof.

In conclusion, the explanation for the user should describe the conflict detection and the conflict resolution. On the other hand, the conflict representation, e.g. using AATS+V, does not need to be explicitly included in the explanation, as it is the representation itself that allows the agent to generate explanations. We argue that if the conflict representation is accurate, then the explanation that is generated from it using PR will consist of a valid and sound argument.

Running Example EXPRI-uploader needs to communicate to Alice the optimal output, i.e. to offer $aud_{sp'}$. There are several possibilities to do so. For illustrative purposes, we report a hypothetical dialogue that may happen between EXPRI-uploader (EU) and Alice (A), to show how EXPRI can provide contrastive and tailored explanations.

> EU(1): Given the disagreement with Bob and Charlie about how to share your picture, to offer $aud_{sp'}$ is your most convenient action, because it would allow you to compromise with your friends (remember that openness-to-change is your most preferred value).
> A: Why shouldn't I offer aud_{sp_A} instead?
> EU(2): Because you could get a better score than the one guaranteed by aud_{sp_A} (obj13.1), openness-to-change would be demoted (obj13.5), and because Bob and Charlie would most likely reject your offer (obj13.7 and obj13.8).

Note that EU(1) is a tailored explanation, because openness-to-change is a very important value to Alice and to highlight its promotion would not necessarily be as efficacious when interacting with a different user. Also, EU(2) is a contrastive explanation, that provides justification for the optimal action by reporting the objections to the alternative action that Alice asked about.

6 "*Tomorrow Never Dies*": Discussion and Future Work

In this paper we have presented EXPRI, an agent architecture that aims to assist users for managing multiuser privacy in online social networks. EXPRI identifies for each user, that is involved in a privacy conflict, the best action to collaboratively solve it, by considering both the utility they would gain by sharing the content online and the personal moral values they would promote by compromising with the other users. EXPRI identifies the most desirable solution by applying practical reasoning techniques. This abductive reasoning allows the agent to gather all the necessary knowledge to justify to the user the selection or the rejection of any particular action. To be able to do so is crucial for an

agent to be considered explainable. However, in order for the agent to be self-explainable, EXPRI also requires social awareness, i.e. the ability of efficiently communicating explanations to the user, for instance by providing contrastive and tailored explanations. We hypothesise that, by using a practical reasoning process, EXPRI is already able to engage in dialogues with the user to provide contrastive explanations. Further theoretical and empirical research will allow us to develop the social component of EXPRI, by enabling it to also provide fully customised explanations.

References

1. Albrecht, S.V., Stone, P.: Autonomous agents modelling other agents: A comprehensive survey and open problems. Artificial Intelligence **258**, 66–95 (2018)
2. Atkinson, K., Bench-Capon, T.: Action-based alternating transition systems for arguments about action. In: AAAI. vol. 7, pp. 24–29 (2007)
3. Atkinson, K., Bench-Capon, T.: Practical reasoning as presumptive argumentation using action based alternating transition systems. Artificial Intelligence **171**(10-15), 855–874 (2007)
4. Atkinson, K., Bench-Capon, T.: Taking account of the actions of others in value-based reasoning. Artificial Intelligence **254**, 1–20 (2018)
5. Bardi, A., Schwartz, S.H.: Values and behavior: Strength and structure of relations. Personality and social psychology bulletin **29**(10), 1207–1220 (2003)
6. Besmer, A., Richter Lipford, H.: Moving beyond untagging: photo privacy in a tagged world. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 1563–1572. ACM (2010)
7. Cranefield, S., Oren, N., Vasconcelos, W.W.: Accountability for practical reasoning agents. In: International Conference on Agreement Technologies. pp. 33–48. Springer (2018)
8. Dignum, V.: Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. Springer International Publishing (2019)
9. Fogues, R.L., Murukannaiah, P.K., Such, J.M., Singh, M.P.: Sharing policies in multiuser privacy scenarios: Incorporating context, preferences, and arguments in decision making. ACM Transactions on Computer-Human Interaction (TOCHI) **24**(1), 5 (2017)
10. Fogues, R.L., Murukannaiah, P.K., Such, J.M., Singh, M.P.: Sosharp: Recommending sharing policies in multiuser privacy scenarios. IEEE Internet Computing **21**(6), 28–36 (2017)
11. Fogues, R.L., Such, J.M., Espinosa, A., Garcia-Fornes, A.: Bff: A tool for eliciting tie strength and user communities in social networking services. Information Systems Frontiers **16**(2), 225–237 (2014)
12. Humbert, M., Trubert, B., Huguenin, K.: A survey on interdependent privacy. ACM Computing Surveys p. 35 (2019)
13. Ilia, P., Polakis, I., Athanasopoulos, E., Maggi, F., Ioannidis, S.: Face/Off: preventing privacy leakage from photos in social networks. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security - CCS '15. pp. 781–792. ACM Press, New York, New York, USA (2015)
14. Kökciyan, N., Yaglikci, N., Yolum, P.: An argumentation approach for resolving privacy disputes in online social networks. ACM Transactions on Internet Technology (TOIT) **17**(3), 27 (2017)

15. Krasnova, H., Spiekermann, S., Koroleva, K., Hildebrand, T.: Online social networks: Why we disclose. Journal of information technology **25**(2), 109–125 (2010)
16. Langley, P.: Explainable, normative, and justified agency. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 9775–9779 (2019)
17. Madumal, P., Miller, T., Sonenberg, L., Vetere, F.: A grounded interaction protocol for explainable artificial intelligence. In: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems. pp. 1033–1041. International Foundation for Autonomous Agents and Multiagent Systems (2019)
18. McBurney, P., Parsons, S.: Dialogue games for agent argumentation. In: Argumentation in artificial intelligence, pp. 261–280. Springer (2009)
19. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence (2018)
20. Misra, G., Such, J.M.: Pacman: Personal agent for access control in social media. IEEE Internet Computing **21**(6), 18–26 (2017)
21. Mosca, F., Such, J.: Towards an explainable and value-driven agent for collective privacy. In: Extended Abstract accepted at AAMAS (2020)
22. Mosca, F., Such, J.M., McBurney, P.: Value-driven collaborative privacy decision making. In: Proceedings of the AAAI Symposium on Privacy-Enhancing Artificial Intelligence and Language Technologies (PAL) (2019)
23. Paci, F., Squicciarini, A., Zannone, N.: Survey on access control for community-centered collaborative systems. ACM Computing Surveys **51**(1) (2018)
24. Panisson, A.R., Sarkadi, Ş., McBurney, P., Parsons, S., Bordini, R.H.: On the formal semantics of theory of mind in agent communication. In: International Conference on Agreement Technologies. pp. 18–32. Springer (2018)
25. Paul, G.: Approaches to abductive reasoning: an overview. Artificial intelligence review **7**(2), 109–152 (1993)
26. Perelman, C., Olbrechts-Tyteca, L.: Traité de l'argumentation. la nouvelle rhétorique (1971)
27. Rajtmajer, S., Squicciarini, A., Such, J.M., Semonsen, J., Belmonte, A.: An ultimatum game model for the evolution of privacy in jointly managed content. In: International Conference on Decision and Game Theory for Security. pp. 112–130. Springer (2017)
28. Sarkadi, S., McBurney, P., Parsons, S.: Deceptive storytelling in artificial dialogue games. In: Proceedings of the AAAI 2019 Spring Symposium Series on Story-Enabled Intelligence (2019)
29. Sarkadi, Ş., Panisson, A.R., Bordini, R.H., McBurney, P., Parsons, S.: Towards an approach for modelling uncertain theory of mind in multi-agent systems. In: International Conference on Agreement Technologies. pp. 3–17. Springer (2018)
30. Schwartz, S.H.: An overview of the schwartz theory of basic values. Online readings in Psychology and Culture **2**(1), 11 (2012)
31. Schwartz, S.H., Bilsky, W.: Toward a theory of the universal content and structure of values: Extensions and cross-cultural replications. Journal of personality and social psychology **58**(5), 878 (1990)
32. Squicciarini, A.C., Shehab, M., Paci, F.: Collective privacy management in social networks. In: Proceedings of the 18th international conference on World wide web. pp. 521–530. ACM (2009)
33. Squicciarini, A.C., Sundareswaran, S., Lin, D., Wede, J.: A3p: adaptive policy prediction for shared images over popular content sharing sites. In: Proceedings of the 22nd ACM conference on Hypertext and hypermedia. pp. 261–270. ACM (2011)

34. Such, J.M.: Privacy and autonomous systems. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence. pp. 4761–4767. AAAI Press (2017)
35. Such, J.M., Criado, N.: Resolving Multi-Party Privacy Conflicts in Social Media. IEEE Transactions on Knowledge and Data Engineering **28**(7), 1851–1863 (2016)
36. Such, J.M., Criado, N.: Multiparty privacy in social media. Communications of the ACM **61**(8), 74–81 (2018)
37. Such, J.M., Porter, J., Preibusch, S., Joinson, A.: Photo privacy conflicts in social media: a large-scale empirical study. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. pp. 3821–3832. ACM (2017)
38. Such, J.M., Rovatsos, M.: Privacy policy negotiation in social media. ACM Transactions on Autonomous and Adaptive Systems **11**(1), 1–29 (2016)
39. Tessier, C., Chaudron, L., Müller, H.J.: Conflicting agents: conflict management in multi-agent systems, vol. 1. Springer Science & Business Media (2006)
40. Ulusoy, O., Yolum, P.: Emergent privacy norms for collaborative systems. In: International Conference on Principles and Practice of Multi-Agent Systems. pp. 514–522. Springer (2019)
41. Winikoff, M.: Towards trusting autonomous systems. In: International Workshop on Engineering Multi-Agent Systems. pp. 3–20. Springer (2017)
42. Wisniewski, P., Lipford, H., Wilson, D.: Fighting for my space: Coping mechanisms for sns boundary regulation. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 609–618. ACM (2012)