Persuasion-enhanced computational argumentative reasoning through argumentation-based persuasive frameworks

Ruiz-Dolz, Ramon; Taverner, Joaquin; Barberá, Stella M. Heras; García-Fornes, Ana

Published in:
User Modeling and User-Adapted Interaction

DOI:
10.1007/s11257-023-09370-1

Publication date:
2024

User Modeling and User-Adapted Interaction (2024) 34:229–258
https://doi.org/10.1007/s11257-023-09370-1

Persuasion-enhanced computational argumentative reasoning through argumentation-based persuasive frameworks

Ramon Ruiz-Dolz[1] · Joaquin Taverner[1] · Stella M. Heras Barberá[1] · Ana García-Fornes[1]

Received: 31 May 2022 / Accepted in revised form: 15 May 2023 / Published online: 19 June 2023

Abstract

One of the greatest challenges of computational argumentation research consists of creating persuasive strategies that can effectively influence the behaviour of a human user. From the human perspective, argumentation represents one of the most effective ways to reason and to persuade other parties. Furthermore, it is very common that humans adapt their discourse depending on the audience in order to be more persuasive. Thus, it is of utmost importance to take into account user modelling features for personalising the interactions with human users. Through computational argumentation, we can not only devise the optimal solution, but also provide the rationale for it. However, synergies between computational argumentative reasoning and computational persuasion have not been researched in depth. In this paper, we propose a new formal framework aimed at improving the persuasiveness of arguments resulting from the computational argumentative reasoning process. For that purpose, our approach relies on an underlying abstract argumentation framework to implement this reasoning and extends it with persuasive features. Thus, we combine a set of user modelling and linguistic features through the use of a persuasive function in order to instantiate abstract arguments following a user-specific persuasive policy. From the results observed in our experiments, we can conclude that the framework proposed in this work improves the persuasiveness of argument-based computational systems. Furthermore, we have also been able to determine that human users place a high level of trust in decision support systems when they are persuaded using arguments and when the reasons behind the suggestion to modify their behaviour are provided.

Keywords Computational argumentation · User modelling · Human–computer interaction · Computational persuasion

✉ Ramon Ruiz-Dolz
raruidol@dsic.upv.es

[1] Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Valencia, Spain

1 Introduction

Computational argumentation is a multidisciplinary area of research that investigates every phase of human argumentation from the computational viewpoint (Atkinson et al. 2017; Ruiz-Dolz 2020). Research in this area is done from different perspectives, such as natural language processing (NLP) (Lawrence and Reed 2019; Gleize et al. 2020; Khatib et al. 2021), knowledge representation and automated reasoning (KRR) (Dung 1995; Baroni et al. 2011), and human–computer interaction (HCI) (Ruiz-Dolz 2019; Chalaguine and Hunter 2020). However, most of the research carried out on this topic focuses on a very specific perspective taken from each area and does not explore the potential existing synergies among the advances performed in different areas. Taking the human argumentative reasoning process as a reference (Walton 2009), we consider that transversal computational argumentation research is of utmost importance in leveraging the findings and proposals made in each specific area of research, for example, by integrating the algorithms proposed for modelling the human argumentative reasoning from a computational point of view (e.g. in KRR), with user modelling (e.g. in HCI) and predictive techniques (e.g. in NLP). Therefore, in this paper, we propose an extension for formal argumentation frameworks and their semantics that enables argument-based computational persuasion. Our main objective is to bridge the gap between argument-based KRR (i.e. formal computational argumentation) and HCI research.

A recent trend in computational argumentation research has been focused on how the computational approaches to the different aspects of human argumentation (e.g. identification, analysis, evaluation, or invention (Walton 2009)) can benefit from combining the advances contributed independently in each specific domain (e.g. NLP, formal logic, HCI, persuasion, etc.). Approaches that extend the specific tasks of argument mining have been investigated in search of a convergence between natural language argument structures and argumentation frameworks (Cocarascu and Toni 2017). Furthermore, recent research reports the benefits of combining argumentation semantics with NLP algorithms for improving the automatic evaluation of argumentative debates (Ruiz-Dolz et al. 2022). However, argument-based computational persuasion research has not explored such synergies in depth yet. Most of the research aimed at (computationally) persuading human users using arguments independently explores the use of machine learning for estimating the most persuasive argument (Donadello et al. 2021), analyses human behaviour with empirical studies (Thomas et al. 2019), or explores the use of interactive chatbots for behaviour change (Chalaguine et al. 2019). A common feature in all of these independent approaches is the modelling of human users, which plays a major role in the personalisation of computational persuasion systems (Hunter 2018).

Following this trend, we introduce the argumentation-based persuasive frameworks (APF), which rely on the argumentative reasoning provided by any underlying abstract argumentation framework and generates user-tailored natural language arguments. This goal is achieved through user modelling, which plays a fundamental role in our proposal and enables a personalised interaction between the human user and the argumentative system. We model our users using their personality and their online behaviour (e.g. number of friends, comments, or likes). Then, natural lan-

guage arguments are created taking into account the logical principles of admissibility and conflict-freeness (Baroni et al. 2011) of abstract arguments encoded in the argumentation framework. The abstract arguments are instantiated into natural language arguments using a set of linguistic features that allow the perceived persuasiveness of the produced arguments to be increased for each different user profile. In addition to the formalisation, we do a complete integration of the APF in the online social network (OSN) domain for the prevention of privacy violations. Furthermore, we evaluate the performance of an argumentation system with an underlying APF when trying to persuade human users not to disclose specific potential privacy threatening publications. We observed a significant improvement in the persuasiveness of arguments when using the proposed APF to engage human–computer interaction instead of relying exclusively on an argumentation framework without any type of explicit personalisation. Furthermore, we have also observed a high level of trust from human users towards the argumentation system when modifying their initial decisions after reading the arguments.

The rest of the paper is structured as follows: Section 2 reviews the previous work done on the intersection of computational argumentation and computational persuasion; Sect. 3 introduces the formal background and provides a formal definition of the argumentation-based persuasive framework; Sect. 4 presents a use case of our framework in the online privacy domain and proposes a complete implementation of the proposed framework in a real argumentation system; Sect. 5 evaluates our proposal in terms of behaviour change and human persuasion; Sect. 6 discusses the obtained results; and Sect. 7 summarises the most important conclusions of this paper.

2 Related work

Persuasion represents one of the most important goals of human argumentation. When engaging in an argumentative dialogue, a common goal is to persuade other participants (McBurney and Parsons 2002). From a computational perspective, persuasion is typically studied as a cornerstone of HCI systems. In computational argumentation research specifically, persuasion has been investigated from different viewpoints (Hunter 2018; Khatib et al. 2020).

The automatic estimation of the persuasiveness of a natural language argument has been widely studied in the NLP area of research. In Gleize et al. (2020), the authors present a corpus that is specifically designed for determining the most persuasive argument from a given pair of arguments. A neural network architecture is trained to learn linguistic features and solve the task of predicting and modelling persuasion from natural language input. Another approach is proposed in Baff et al. (2020), where the authors focus on the analysis of the impact of style on the persuasive power of news editorial arguments. For that purpose, five different NLP features are used to model style: Linguistic Inquiry and Word Count, a lexicon of emotions (i.e. anger, disgust, and fear) and sentiments (i.e. positive and negative), argumentative discourse units features (i.e. anecdotal, statistical, and testimonial evidence) (Khatib et al. 2017), arguing elements (i.e. assessments, doubt, authority, and emphasis) (Somasundaran et al. 2007), and text subjectivity (i.e. subjective or objective) (Wiebe and Riloff 2005).

These features are used to train a support vector machine (SVM) (Vapnik 1998) on a task aimed at predicting whether or not a message will be persuasive. Finally, we can observe a combination of NLP and user modelling in Khatib et al. (2020). The authors propose an approach that uses users' beliefs, interests, and personality traits, along with NLP feature engineering on natural language inputs to predict the persuasiveness of arguments and users' resistance to persuasion. However, the analysed research only takes into consideration natural language and user models and does not take argumentative reasoning into account.

A different approach aimed at understanding specific aspects of the persuasive properties of computationally generated arguments comes by the hand of empirical studies. In Thomas et al. (2019), the authors propose a scale to measure the persuasive power of different argumentative messages in the health and security domains. The scale is developed after conducting a study where users were asked to provide information related to three different factors of the perceived persuasiveness of different messages: their effectiveness, their quality, and their capability. A study of the impact of the personality, the age, and the gender of human users on their susceptibility to persuasive messages is done in Ciocarlan et al. (2019). Combined with the results presented in Ruiz-Dolz et al. (2022), we can learn more about the persuasion of arguments when used in an argumentative interaction with a human user based on personal characteristics. Another interesting approach is presented in Ruiz-Dolz et al. (2021), where the authors propose a metric for measuring the persuasive power of different reasoning patterns and arguments based on a study with human participants. The study makes an analysis of how human features (i.e. personality and social interaction) are related to perceived persuasive power. Finally, in Hadoux and Hunter (2019), the authors present a series of empirical studies that are designed to measure how different preferences and concerns of human users can be a factor of influence in perceived persuasion when reading specific arguments.

Persuasion has also been studied as the utility function of argumentation dialogues and negotiation. In Hadoux et al. (2018), the authors present a framework for argumentation-based decision-making assistance. This framework relies on decision trees for modelling the dialogue and improves its persuasiveness when the user model is combined with emotional features. In a dialogue, choosing which argument are more persuasive can be modelled as an optimisation of a strategy learning problem. With regard to this, reinforcement learning (RL) (Sutton and Barto 1998) is a promising technique for learning persuasive dialogue strategies. In Monteserin and Amandi (2013), persuasion is defined as the effectiveness of arguments when used in a negotiation for reaching a satisfactory agreement. In that work, an argumentative agent learns to use the most persuasive argument in a given step of the dialogue through RL. Similarly, RL is used for learning dialogue strategies in Alahmari et al. (2019). Furthermore, in Hadoux et al. (2021), the authors retake the belief-concern user model of Hadoux and Hunter (2019) and propose a Monte Carlo tree search for finding the optimal persuasive policies for specific user models. The belief-concern user model was also considered in Hunter et al. (2019), where a general framework for computational persuasion is presented. This framework is instantiated into an argumentative chatbot for the purpose of behaviour change in the domains of cycling and university fees. In a recent work, a machine learning approach to argument-based persuasion was

proposed in Donadello et al. (2021), where bi-party decision trees are used for predicting an argument's utility (i.e. persuasiveness) in a dialogue. The proposed model is evaluated in a simulated environment. Finally, in a recent work, a visual interactive system for making persuasive analyses of online discussions has been proposed (Xia et al. 2022). This system makes it possible to improve the persuasive strategies of users through a complete visualisation of different persuasive features of arguments when used in a dialogue.

From the previous literature review, two major limitations are identified. First, there is only limited research on how computational argumentative reasoning can be extended to a persuasive argumentative system. Research on this topic is relevant for deepening computational persuasion research, where a system could perform argumentative reasoning before interacting with a human user. Second, there are not many evaluations of behaviour change with real humans. Even though argument-based computational persuasion has been explored from many different viewpoints, only a few works have conducted a complete evaluation of their proposal when trying to persuade human users. Furthermore, it has not been possible to identify many works where concepts from computational argumentation theory are combined with HCI and argument-based persuasion such as Hadoux and Hunter (2019) and Rosenfeld and Kraus (2016). In Hadoux and Hunter (2019), argumentation frameworks are used for computationally representing arguments as a graph. However, this work only considers this concept as a data structure, and the automatic argumentative reasoning is not carried out using argumentation semantics. In contrast, in Rosenfeld and Kraus (2016), the authors propose an argumentative agent that uses a formal argumentation framework and its semantics for approaching argumentative reasoning, together with a partially observable Markov decision process for learning persuasive strategies. This agent is evaluated when interacting with real human users, but only a very small population is used. Our research extends this line of work by providing a formal framework for generalising the integration of argumentation frameworks with persuasive systems, combined with user modelling for personalising the interactions. We present an implementation of our proposal with a complete evaluation of its persuasiveness when interacting with human users, which has been evaluated in a sample population of 50 participants.

3 Formalisation

In this section, we present all of the formal definitions that support the research conducted in this paper. First, we introduce all of the required background concepts in order to have a complete understanding of the scope of our proposal and our experimentation. Second, we formalise our argument-based persuasive framework.

3.1 Background

Before defining our proposal for an argument-based persuasive framework, it is of the utmost importance to introduce some fundamental formal aspects of the computa-

tional abstract argumentation theory. The concept of *argumentation frameworks* can be considered as a cornerstone in this topic, from which most of the research in computational argumentation and logic has been based. As proposed in Dung (1995), an *argumentation framework* makes it possible to computationally represent the logical aspects behind human argumentation from an abstract perspective:

Definition 1 *(Abstract Argumentation Framework).* An abstract argumentation framework (AAF) is a tuple $AAF = \langle A, R \rangle$ where A is a set of arguments, and R is the attack relation on A such that $A \times A \rightarrow R$.

Thus, an *argumentation framework* can be instantiated as a directed graph, where nodes are arguments and edges are attack relations between arguments. This representation eases the computational encoding of an argument-based reasoning. However, *argumentation frameworks* are just data structures and representations and do not enable an analysis of their underlying reasoning *per se*. The set of (topo)logical rules or conditions that make it possible to carry out the analysis of an argument that is instantiated into an *argumentation framework* are the *argumentation semantics*. Through the semantics, it is possible to determine the set of *acceptable* (and *defeated*) arguments. In this paper, we emphasise the fundamental properties behind *argumentation semantics*, but a thorough review of the most important semantics is conducted in Baroni et al. (2011). This way, the *argumentation semantics* defines the conditions required to determine the set of *acceptable* (and *defeated*) arguments belonging to an *argumentation framework*. These conditions rely on two basic properties that are related to sets of (abstract) arguments: the conflict-free principle and the principle of admissibility.

Definition 2 *(Conflict-free).* Let $AF = \langle A, R \rangle$ be an argumentation framework and $Args \subseteq A$. The set of arguments $Args$ is conflict-free *iff* $\neg\exists\alpha_i, \alpha_j \in Args$: $(\alpha_i, \alpha_j) \in R$.

Definition 3 *(Admissible).* Let $AF = \langle A, R \rangle$ be an argumentation framework and $Args \subseteq A$. The set of arguments $Args$ is admissible *iff* $Args$ is conflict-free, and $\forall\alpha_i \in Args$, $\neg\exists\alpha_k \in A$: $(\alpha_k, \alpha_i) \in R$, or $\exists\alpha_k \in A$: $(\alpha_k, \alpha_i) \in R$ and $\exists\alpha_j \in Args$: $(\alpha_j, \alpha_k) \in R$ (i.e. defends $Args$).

This way, it is possible to define a conflict-free set of arguments whenever no attack relations can be observed among the arguments included in the set, and an admissible set of arguments whenever the arguments belonging to a conflict-free set also defend themselves from external attacks. It is important to point out that admissible sets of an AF are always among the conflict-free sets of the same AF. Let us illustrate these formal definitions with the example shown in Fig. 1. Assuming a situation where four different arguments are encoded in an AF, i.e. $\alpha_1, \alpha_2, \alpha_3, \alpha_4 \in A$, and the relations $(\alpha_1, \alpha_2), (\alpha_2, \alpha_3), (\alpha_3, \alpha_4), (\alpha_4, \alpha_3) \in R$; it could be possible to define two groups of acceptable arguments depending on which principle is brought into consideration. The conflict-free sets of arguments are $\{\alpha_1, \alpha_3\}$, $\{\alpha_1, \alpha_4\}$, and $\{\alpha_2, \alpha_4\}$ since there are no attack relation among the arguments included in these sets. In contrast, the admissible set of arguments would only be $\{\alpha_1, \alpha_4\}$ because only these two arguments are conflict-free and are able to defend themselves from external attacks.

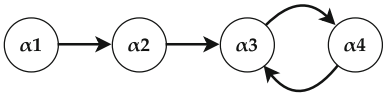

Fig. 1 Abstract argumentation framework

From these properties, two major families of semantics for abstract *argumentation frameworks* arise, conflict-free and admissibility-based semantics. Some significant examples of these semantics are complete, preferred, grounded, and ideal for admissibility-based semantics, and Naïve, Stage, and CF2 for conflict-free based semantics (see Baroni et al. 2011 for more detail in their formalisation and properties). Depending on each domain and/or the nature of the encoded argument, the suitability of *argumentaton semantics* can differ. However, in general, the admissibility principle is of the utmost importance when defining consistent sets of arguments from a framework since they can defend themselves.

Finally, in order to completely understand the experimentation carried out in this work, it is important to introduce the argumentation framework for online social networks (AFOSN). This framework was originally proposed in Ruiz-Dolz et al. (2019) as the basis of an argumentation system aimed at the prevention of privacy threats in online environments. Its underlying mechanism is based on the theory behind the QBAFs (Baroni et al. 2015) and allows the acceptability of an abstract argument to be determined depending on a quantitative feature. For this purpose, in addition to abstract arguments and attacks, the AFOSN relies on information that is extracted from the social network (i.e. publication features and user profiles) and on an argument scoring function for determining the acceptability of the arguments. The AFOSN is formally defined as follows:

Definition 4 *(Argumentation Framework for Online Social Networks)***.** We define an argumentation framework for online social networks as a tuple $AFOSN = \langle A, R, P, \tau \rangle$, where A is a set of n arguments $[\alpha_1, \ldots, \alpha_n]$; R is the attack relation on A such that $\mathrm{A} \times \mathrm{A} \rightarrow \mathrm{R}$; P is the list of e profiles involved in an argumentation process $[p_1, \ldots, p_e]$; and τ is a function $A \times P \rightarrow [0, \ldots, 1]$ that determines the score of an argument α for a given profile p.

An argument $\alpha \in A$ is instantiated by the framework as a 3-tuple $\alpha = (\beta, T, D)$: β represents the claim (i.e. +1 if the argument is in favour and -1 if the argument is against sharing); T indicates the type of the argument (i.e. privacy, risk, trust and content); and D encodes the support of the argument (i.e. a numerical value distilled from the online social network environment). Each user profile $p \in P$ is also instantiated as a 3-tuple $p = (\nu, \rho, M)$, where the preference values ν, the personality of a user profile ρ and a set of general information M (e.g. age, likes, statistics) are used to model human users. Finally, the argument scoring function τ is defined as follows:

$$\tau(\alpha, p) = \alpha_\beta \cdot \alpha_D \cdot p_{v_i} \tag{1}$$

The resulting product of the claim, the support of the argument, and the preference value of a specific human user towards each topic will determine the strength of an argument in the AFOSN. Then, it is possible to define defeat for an argument as follows:

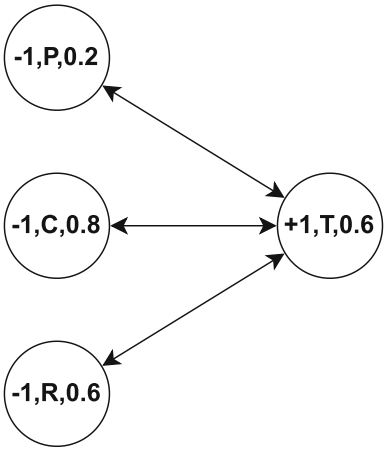

Fig. 2 Example of an AFOSN. Each node represents an argument in favour or against sharing a given publication generated from the social network information

Definition 5 *(Defeat (AFOSN))***.** An argument $\alpha_i \in A$ defeats another argument $\alpha_j \in A$ in a context determined by a user profile p iff $(\alpha_i, \alpha_j) \in R \wedge |\tau(\alpha_i, p)| > |\tau(\alpha_j, p)|$.

The collective defeat for a set of arguments w.r.t. another set of arguments is defined as follows:

Definition 6 *(Collective Defeat (AFOSN))***.** The set of arguments $Args_i \subset A$ defeats the set of arguments $Args_j \subset A$ in a context determined by a user profile p iff $\forall \alpha_i \in Args_i$, $\forall \alpha_j \in Args_j$, $(\alpha_i, \alpha_j) \in R \wedge \sum_{\forall \alpha_i \in Args_i} |\tau(\alpha_i, p)| > \sum_{\forall \alpha_j \in Args_j} |\tau(\alpha_j, p)|$.

Thus, from these defeat definitions, it is possible to define acceptance (considering defeat) and collective acceptance (considering collective defeat) in an AFOSN:

Definition 7 *(Acceptance (AFOSN))***.** An argument $\alpha_i \in A$ is acceptable iff $\forall\, \alpha_j \in A \wedge defeat(\alpha_j, \alpha_i) \rightarrow \exists \alpha_k \in A \wedge defeat(\alpha_k, \alpha_j)$ or $\forall\, \alpha_j \in A \wedge \nexists\, defeat(\alpha_j, \alpha_i)$.

Definition 8 *(Collective Acceptance (AFOSN))***.** The set of arguments $Args_i \subset A$ is acceptable iff $\neg \exists\ Args_j \subset A; Args_i \cap Args_j = \emptyset \wedge defeat(Args_j, Args_i)$.

It is important to emphasise that collective defeat and collective acceptance are the core of an AFOSN since there will always be two sets of arguments, one in favour of sharing and one against doing it. Let us illustrate this second framework with the example depicted in Fig. 2. Imagine that User A shares a post saying "Looking forward our trip to London next week", tags his friend User B, and shares it with the public configuration setting (i.e. visible by the whole network). In this case, the AFOSN will generate three arguments against sharing the publication and one in favour. After analysing the network data (i.e. the post, and A and B user profile preferences), the framework will generate a privacy argument against this publication (because User A typically shares posts considering more restricted configurations), a content argument against the publication (because the author is revealing his future location), a risk argument against sharing the publication (because the post-propagation in the social network will reach unexpected users), and a trust argument in favour of sharing this post

(because based on the previous social interactions, users A and B present an elevated degree of trust). This way, the AFOSN will result in a bipartite graph, granting the properties of conflict-freeness and admissibility to the acceptable arguments defined under collective acceptance.

3.2 Argument-based persuasive framework

Abstract argumentation frameworks and semantics provide the formal tools to encode human argumentative reasoning from a computational viewpoint. However, most of the research in formal argumentation focuses on proposing models for approaching argumentative reasoning instead of deepening the focus on how the output of the underlying reasoning could be used in a direct human–computer interaction. In this paper, we formalise the argument-based persuasive framework as a higher-level framework that enables human–computer interaction and that can be instantiated on top of any abstract argumentation framework. Our proposal brings into consideration any underlying formal argumentation framework that is in charge of approaching the argumentative reasoning, a human user model for personalising and adapting the interaction, and a set of linguistic features to concretise the abstract arguments. All of these features are combined by a persuasive function as described below:

Definition 9 *(Argument-based Persuasive Framework)***.** We define an argument-based persuasive framework as a tuple $APF = \langle AF, U, L, \gamma \rangle$, where AF is the underlying argumentative framework; U is the human user model; L is a set of linguistic features; and γ is a persuasive function that produces a persuasive natural language argument (NLA) such that $U \times Args \times L \rightarrow NLA$.

Each user model U contains a set of user descriptive features (e.g. personality, behavioural patterns, or emotions) that may vary depending on the application environment and domain, and the availability of such features. The set of linguistic definitions L (e.g. argumentation schemes, argument templates or databases, or logical structures) contains different non-abstract representations of the arguments that are included in the argumentation process. Finally, the γ function is aimed at estimating the most persuasive natural language argument given a set of arguments and natural language features for a specific user profile:

$$\gamma(U, \text{Args}, L) = \hat{Ar} \tag{2}$$

which takes as input the user descriptive features associated with a human profile U, the set of acceptable arguments $Args \in A$ (where A is the argument set of the underlying AF), and the set of linguistic features L, to produce a persuasive argument $\hat{Ar} \in |NLA|$ belonging to the domain of natural language arguments. Using the APF, a new dimension to formal computational argumentation research can be unlocked. This framework makes it possible to leverage the computational argumentative reasoning provided by any argumentation framework (which may vary depending on our needs, the application domain, or the available information) for defining better informed persuasive strategies through the use of arguments and, thus enable an effective argument-based HCI.

4 Implementation of the argument-based persuasive framework

To validate our formal proposal and to depict how the argument-based persuasive framework can be instantiated and implemented in a real situation, we have chosen the domain of privacy management in online social networks (OSNs). Privacy violations in OSNs are a threat of major concern that has been thoroughly researched in the literature. Different viewpoints and approaches can be identified when dealing with this problem, e.g. automatic agent-based negotiations (Kökciyan et al. 2017), privacy nudges (Acquisti et al. 2017), persuasive argumentation systems (Ruiz-Dolz et al. 2019), and the multi-party privacy conflict (Mosca and Such 2022), among others. As discussed in Sect. 3, an AFOSN provides the underlying reasoning mechanism of an argumentation system that is aimed at identifying and preventing privacy violations in OSNs (Ruiz-Dolz et al. 2019). In this paper, we retake this domain to instantiate the argument-based persuasive framework (APF) on top of the AFOSN and to evaluate its power of behaviour change when preventing privacy violations.

For that purpose, we instantiate the APF (i.e. $\langle AF, U, L, \gamma \rangle$) as follows:

- The computational argumentative reasoning engine (AF) is managed by an AFOSN. Whenever a new post is being shared in the network, it generates a set of abstract arguments from the data retrieved from the OSN (Ruiz-Dolz 2019). For that purpose, user and post information are automatically retrieved from the network. The natural language of the post is analysed to identify sensitive information, the privacy configuration of the post (set in the OSN) is used to determine the potential privacy issues, and the user network is used to determine the post reachability risks and the trust between different users. Then, the AFOSN instantiates a set of abstract arguments (see *Argumentation Framework for Online Social Networks*, Definition 4) in favour and against sharing the publication considering all the retrieved information. Finally, the set of acceptable arguments is defined (see *Collective Acceptance*, Definition 8) to determine if a potential privacy violation is happening.
- The user model (U) is instantiated tacking into account two different helpful aspects for user behaviour modelling: the Big Five personality traits model (Rothmann and Coetzer 2003) (i.e. openness, conscientiousness, extraversion, agreeableness, neuroticism), and their OSN interaction data. As proven in previous research (Ruiz-Dolz et al. 2021), both aspects are helpful in identifying variances in the perceived persuasive power of arguments and reasoning patterns.
- The set of linguistic definitions (L) enables the natural language representation of the abstract arguments provided by the argumentation framework. In our experiments, we consider the four argument types supported by the AFOSN (i.e. privacy, risk, trust, and content) and five different argumentation schemes (Walton et al. 2008) (i.e. patterns of human reasoning) in order to define a database of 45 domain-specific natural language arguments. We selected five commonly used argumentation schemes that suited our application domain and that were researched in previous studies (Ruiz-Dolz et al. 2021): the *Argument from Consequences* (AFCQ), the *Argument from Expert Opinion* (AFEO), the *Argument from*

Popular Practice (AFPP), the *Argument from Popular Opinion* (AFPO), and the *Argument from Witness Testimony* (AFWT).

- The persuasive function (γ) is approached in two steps: persuasive policy learning and natural language argument generation. This way, in our approach, we first estimate a persuasive policy for each specific user, and then we generate natural language arguments by combining the predicted policies and the argumentative linguistic definitions. Both steps are described in the following sections.

4.1 Persuasive policy learning

4.1.1 The persuasive policy learning task

Our first step for approaching the γ function is to learn user-specific persuasive policies. For that purpose, we need to consider both the user model U and the linguistic definitions L. Furthermore, depending on the content and nature of any privacy threatening publication, the set of coherent arguments may vary (e.g. if a publication does not involve more than one person, it would not be acceptable to argue against sharing the publication by reasoning that some other user that appears in the publication could be offended). Our objective is to be able to always use the most persuasive coherent argument for each given author of any conflicting publication. For this purpose, we need to estimate the persuasive policies π^s and π^t for the whole set of argumentation schemes (s) and argument types (t) considered in this work, respectively. We define a persuasive policy $\pi \in \mathbb{R}^l$, where l are argumentative features in L, as follows: $\pi = [\alpha_1, \alpha_2, \ldots, \alpha_{|l|}]$, where $pp(\alpha_1) \geq pp(\alpha_2) \geq \cdots \geq pp(\alpha_{|l|})$ being $pp(\alpha)$ the perceived persuasive power of an argument α by a human user U. We consider two different sets of linguistic features L: five argumentation schemes ($l_s = 5$) and four argument types ($l_t = 4$). Furthermore, we use the persuasive power definition presented in Ruiz-Dolz et al. (2021), where the persuasiveness of an argument is represented as a quantitative score based on the position of each argument in a persuasive ranking indicated by human users. Thus, our persuasive policies are represented as lists with orderings of arguments based on their assigned persuasive power.

In this work, we model the persuasive policy learning as a maximisation of the conditional probability described in Eq. 3. For each user model U, we need to estimate the probabilistic distributions of the persuasive power of both the argumentation schemes π^s and argument types π^t.

$$\hat{\pi}_U^{s,t} = \arg\max_{j \in J} P(\pi_j | U) \tag{3}$$

where J is the total number of possible combinations for a given set of linguistic features (i.e. 5! for the argumentation schemes, and 4! for the argument types). Then, each user U is modelled by combining the two features described above (i.e. Big Five and OSN interaction data), which will be the input for the probabilistic models in our experiments. To sum up, we approach the persuasive policy learning as a pattern recognition task. The goal is to identify any existing pattern in the different user models that allow us to determine the optimal privacy policy for each specific user model.

4.1.2 The OSNAP-400 dataset

To learn user-specific persuasive policies and to approach this task as the probabilistic modelling proposed in Eq. 3, we have developed a new dataset for argument persuasion. A total of 400 adults (194 males, 206 females) from 18 to 76 years old completed a study designed for the creation of the Online Social Network Argument Persuasion (OSNAP-400) dataset.[1] This study was aimed at adult OSN users. The study from which we created the OSNAP-400 dataset consisted of the 50-item personality inventory (Goldberg 1999), two persuasive questionnaires for argumentation schemes (*Questionnaire A*), and argument types (*Questionnaire B*), and an OSN interaction questionnaire (*Questionnaire C*). The configuration of the questionnaires is described in Appendices A.1, A.2, and A.3. In the persuasive questionnaires, the participants had to order the arguments (i.e. schemes and types) displayed in a randomised way based on their perceived persuasiveness. Furthermore, we included attention check questions in all of the questionnaires in order to validate their submissions.

For the elaboration of the OSNAP-400, we first calculated the Big Five personality traits of all of the participants from the results of the 50-item personality test. Then, with the answers provided in *Questionnaires A* and *B*, we also calculated the persuasive power of the five argumentation schemes and the four argument types following the definition presented in Ruiz-Dolz et al. (2021), from which we generated the ground truth persuasive policies for each specific user. Finally, we encoded the OSN interaction answers of *Questionnaire C* to discrete normalised values in the range from 0 to 1. Thus, the OSNAP-400 consists of 400 samples. Each sample of the dataset represents a different OSN user modelled with the Big Five and the OSN interaction data and is associated with two persuasive policies (one for argumentation schemes π^s, and the other for argument types π^t).

Before approaching the persuasive policy learning task, we conducted a descriptive analysis of the OSNAP-400 data. First of all, we analysed the user descriptive features (see Fig. 3). For the OCEAN Big Five personality traits (i.e. OCEAN stands for Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) of our samples, we observed almost all of the possible values in the allowed range for every trait (see Fig. 3a). However, we also were able to observe that extraversion and neuroticism traits tend to have lower values than the rest in our dataset. For the social network interaction data, we included twelve different user modelling features that represent the online behaviour of each human user: the number of friends, the number of status updates, the number of likes, the number of comments, the number of publications shared in private, the number of publications shared in public, the number of publications shared with friends only, the number of publications shared with a specific collection of friends, the number of publications deleted, the number of photos uploaded, the average length of the text in the publications, and the average time spent using OSN. Some interesting insights can be observed: how users prefer to share content with friends rather than the whole network; that it is easier for users to give likes than to comment on other users' publications; and that there is an important number of publication regrets that lead to deleting the previously shared content (see Fig. 3b).

[1] Contact the authors for data availability inquiries.

Furthermore, it is important to emphasise that the age distribution of the samples used in our experiments is not uniform (see Fig. 33); most of the samples are within in the 22–34 age interval. Finally, for the gender distribution, we have a balanced population of 194 male samples and 206 female samples (see Fig. 3d).

We also analysed the distribution of the observed persuasive policies π^s and π^t in the OSNAP-400, in order to describe how balanced the dataset is. Figure 4 depicts the frequency at which each persuasive policy appears in the dataset. We observed that regardless of being argumentation schemes or argument types, there is a very strong imbalance in the data. We found that the most frequent persuasive policy of argumentation schemes (with a total of 22 occurrences) was the following one: AFCQ > AFPO > AFEO > AFWT > AFPP. It was closely followed by the second most frequent persuasive policy for argumentation schemes with (21 occurrences): AFCQ > AFEO > AFPO > AFWT > AFPP. We observed how the arguments from consequences are in general perceived to be the most persuasive pattern of human reasoning in our domain. On the other hand, regarding argument types, we observed that the most frequent persuasive policy (with a total of 60 occurrences) is dominated by the arguments containing content references: *Content* > *Trust* > *Privacy* > *Risk*. The strong imbalance observed between the existing persuasive policies of argumentation schemes and argument types makes the persuasive policy learning a hard task to perform a probabilistic modelling on, as the following section shows.

4.1.3 Experimental results

Finally, we present the results obtained in the proposed persuasive policy learning task. For that purpose, we trained five different models to predict how a given user should perceive the persuasive power of both argumentation schemes and argument types and generate the subsequent user-specific persuasive policies π^s and π^t. Considering the probabilistic modelling defined in Eq. 3, the user modelling features were used as the input for our models, and an optimised persuasive policy was generated as the output. Based on the findings of a previous study on the persuasive power of arguments in the OSN domain (Ruiz-Dolz et al. 2021), we modeled our users by combining their Big Five personality traits together with twelve different features that represent their social behaviour in online environments.

Thus, four classical machine learning algorithms have been used in our persuasive policy learning experiments: support vector regression, stochastic gradient descent linear regression, K-neighbours regression, and random forests. Support vector regression (SVR) (Drucker et al. 1996) is a maximum margin regression model which has shown good performance in a wide variety of tasks. After optimising its hyperparameters, we used the linear kernel, $C = 100$ and 1e-9 tolerance values. Stochastic gradient descent linear regression (SGDLR) (Bottou 2012) is a technique by which a linear model is optimised with stochastic gradient descent on minimising a regularised empirical loss. In our experiments, we obtained the best results minimising the huber loss function with a 1e-3 tolerance value and a 1e-5 alpha. K-nearest neighbours regression (k-NNR) is a regression method that is based on the k-nearest neighbours algorithm (Cover and Hart 1967). The estimated value for an unobserved sample is based on the k samples that are the closest to it. In our experiments, we considered the 32 near-

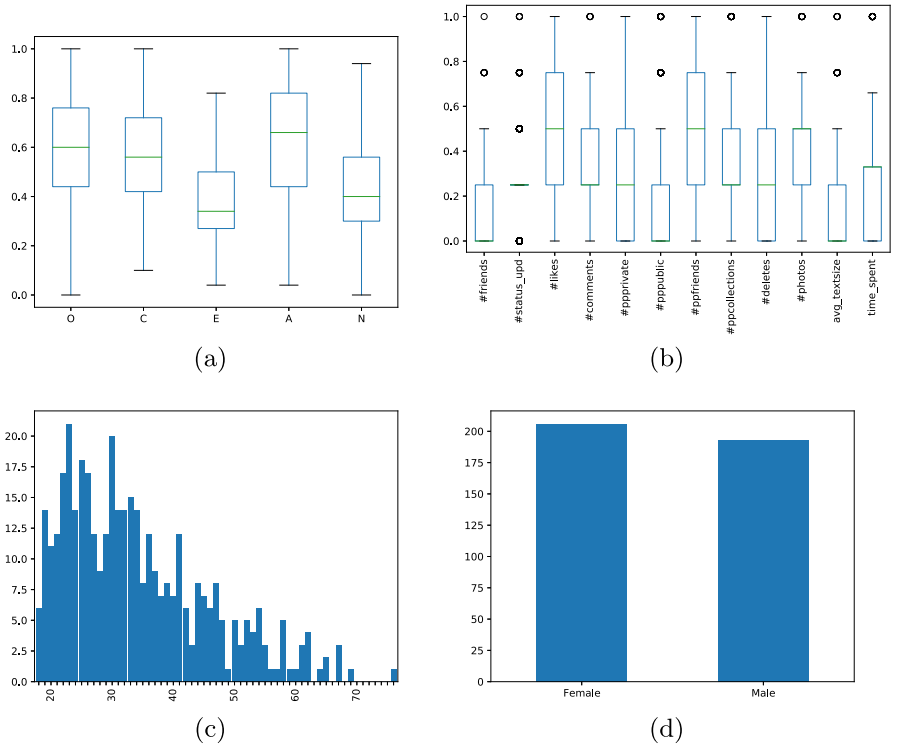

Fig. 3 **a** Box and whiskers diagram of the OCEAN Big Five personality traits observed among the samples of the OSNAP-400 dataset. **b** Box and whiskers diagram of the OSN interaction data observed among the samples of the OSNAP-400 dataset. **c** Age distribution of the OSNAP-400 dataset samples. **d** Gender distribution of the OSNAP-400 dataset samples

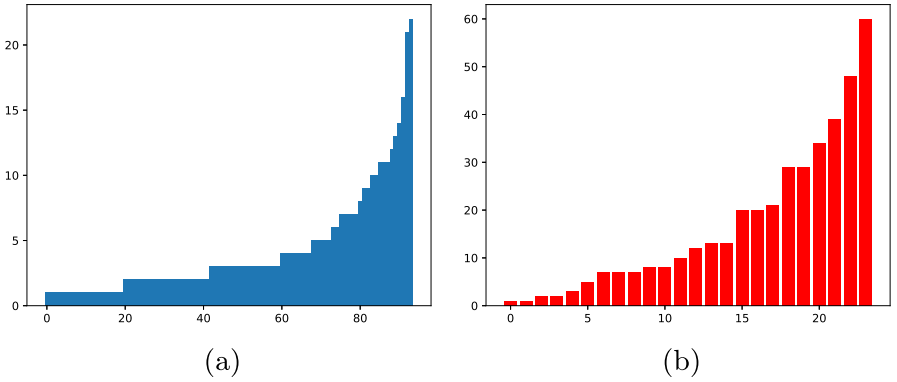

Fig. 4 Distribution of the number of occurrences of the observed persuasive policies. **a** stands for argumentation schemes and **b** for argument types. The *Y* axis represents the number of occurrences of each different persuasive policy. The *X* axis represents each different observed persuasive policy. Each policy is represented by a unique *id* from 0 (the least frequent) to *N*-1 (the most frequent), with N being the number of different persuasive policies observed in our data

est neighbours weighted by their distance to the new observation. The last classical approach considered in this work are random forests (Breiman 2001). A random forest is a meta-learning technique which fits a specific number of decision trees on different subsets of the original dataset. In our experiments, we used 10,000 decision trees to estimate the value that minimises the mean absolute error loss for each tree split. We used the *sklearn*[2] implementations of all of the described classical machine learning algorithms.

In addition to these four classical machine learning models, we also experimented with a neural network model. We implemented a feed-forward multi-layer perceptron (MLP) to approach the persuasive policy learning task. The chosen architecture for our model consists of three hidden layers (32, 32 and 64 units per layer) with *ReLU* activation functions and a total amount of 4196 parameters. The input layer has as many units as the size of our input (i.e. 17 user modelling features). The output layer has 4 or 5 units depending on the persuasive policy being learnt (π^t or π^s, respectively) and a *sigmoid* activation function.

The performance results of the described models on the persuasive policy learning task are shown in Table 1. In addition to the five models, we also considered two baselines: a random baseline and a majority baseline. First, the random baseline assigns a random persuasive power (i.e. a value in range [0,1]) to each one of the arguments and generates a persuasive policy by ordering them by their randomly assigned persuasive power. Second, the majority baseline uses the most common persuasive policy of both argumentation schemes and argument types for all users regardless of their descriptive features. Three different metrics were used to evaluate different aspects of the persuasive policy learning task: the mean absolute error (MAE, lower is better), the hit rate (HR, higher is better), and the Spearman ρ correlation (higher is better). These are common metrics that are used to evaluate recommendation systems with similar requirements (Gunawardana and Shani 2015). The MAE indicates the quality of the model predictions tacking exclusively into account the persuasive power estimations of each individual argument. However, it is not possible to draw significant conclusions about the performance on the persuasive policy learning task considering the MAE alone. The hit rate (HR) measures the number of *hits* observed in the predicted persuasive policies. We consider a *hit* to be whenever an argument (scheme or type) is correctly placed in the predicted persuasive policy compared to the ground truth persuasive policy for a given human user. This metric is most revealing when it comes exclusively to the performance of our models in the persuasive policy learning task. Finally, to complement the previously described metrics, we also considered the Spearman ρ correlation measure between predicted and ground truth persuasive policies. With the Spearman ρ metric, it is possible to evaluate how good the models are at learning partial orderings in the predicted persuasive policies. For example, assuming the ground truth persuasive policy $\pi_u = [\alpha_1, \alpha_2, \alpha_3, \alpha_4]$ and the estimated persuasive policy $\pi'_u = [\alpha_2, \alpha_1, \alpha_4, \alpha_3]$, then $\text{HR}(\pi'_u) = 0$ but $\rho(\pi'_u) = 0.6$, since the estimated persuasive policy does not have any argument placed in its correct position, but the persuasive partial orderings of arguments are decently estimated. This way, it is possible to understand how well the models are performing, not only when predict-

[2] https://scikit-learn.org/stable/index.html.

Table 1 Results obtained on the persuasive policy learning task (Schemes π^s/ Types π^t)

Model	MAE (π^s / π^t)	Hit Rate (π^s / π^t)	Spearman ρ (π^s / π^t
Random Baseline	0.32/0.31	0.22/0.23	−0.02/0.04
Majority Baseline	–	0.20/0.19	**0.22**/−0.01
SVR	**0.16/0.17**	**0.34**/0.38	**0.10/0.09**
SGDLR	0.17/0.17	0.32/0.38	0.06/0.07
k-NNR	0.17/0.17	0.32/**0.40**	0.04/0.07
RandomForest	0.17/0.17	0.33/0.38	0.08/0.06
MLP	0.18/0.18	0.33/0.38	0.09/0.05

Bold indicates the best performing model in general and the best scores in each of the evaluation metrics
The depicted results represent the average of a tenfold evaluation

ing persuasive policies, but also when predicting the individual persuasive power of arguments, and retaining partial ordering dependencies between different arguments.

It can be observed in Table 1, in general, the models perform better than the proposed baselines. Furthermore, it can also be observed that all of the models perform similarly after a tenfold evaluation using the OSNAP-400 dataset. We attribute this behaviour to model convergence and a limited size of training samples. However, the proposed models achieved an improvement with respect to the baselines of 42–50% regarding the prediction of the individual persuasive power of arguments (i.e. MAE), an improvement of 54–110% regarding the accuracy when estimating persuasive policies (i.e. HR), and an improvement of 125% when learning partial orderings in the estimated persuasive policies (i.e. Spearman ρ). These results are reported when learning persuasive policies for both argumentation schemes and argument types (π^s and π^t, respectively). An exception in the Spearman ρ performance of the majority baseline for argumentation scheme persuasive policy estimation can also be observed. It presents outstanding results compared to the rest of approaches. This may be because of the data distribution of ground truth persuasive policies of argumentation schemes (see Fig. 4a), where the most common occurrences are slight variations preserving similar partial orderings. However, it performs significantly worse than the rest of the models regarding the hit rate, even worse than the random baseline. Thus, even though it outperforms our models when learning partial orderings of the persuasive policies, it is not a solid alternative to bring into consideration when approaching the persuasive policy learning task.

4.2 Natural language argument generation

Our second step in this work is the generation of natural language arguments. Once we have computed the user-specific persuasive policies ($\pi_U^{s,t}$), we need to be able to automatically generate a natural language argument for each abstract argument produced by the AFOSN in order to persuade the human user. For that purpose, we defined a database of 45 natural language arguments by combining the four types of arguments supported by the AFOSN with the five argumentation schemes selected

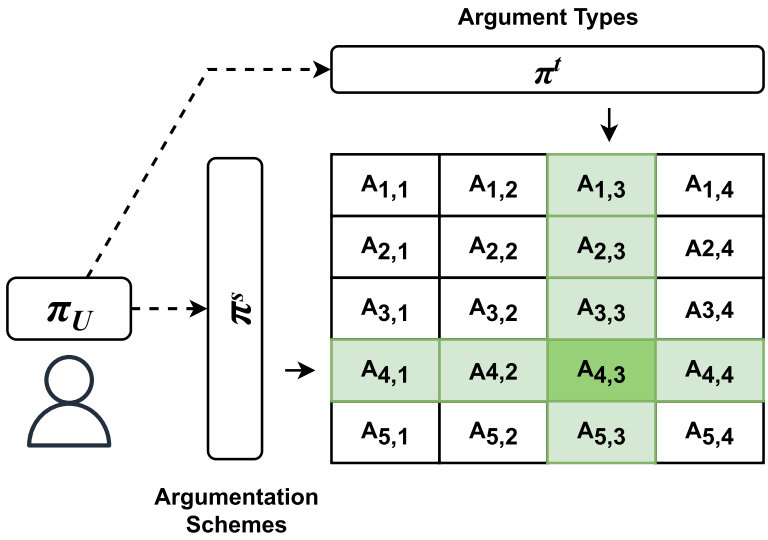

Fig. 5 Scheme of the proposed natural language argument generation method

for the OSN domain. This way, the persuasive function γ takes into account the user model U, the set of acceptable arguments provided by the AFOSN $Args$, and the set of linguistic features L. The list of the arguments included in the database is described in Appendix A.

Our approach is then able to generate a different natural language argument for each user model depending on the predicted privacy policies (both π^s and π^t for argumentation schemes and argument types, respectively). As shown in Fig. 5, when engaging a persuasive interaction with a human user, our system selects from the argument database the most (potentially) persuasive argument considering the persuasive policy estimations. Thus, our argumentation system retrieves the natural language argument tacking into account the most persuasive argumentation scheme (rows) and the most persuasive argument type (columns) from the set of acceptable arguments. Our proposed method for generating natural language arguments only considers arguments that are coherent with each privacy threatening situation. Therefore, the argumentation system will select the most persuasive argument type provided by π^t, from only the ones that are included in the set of acceptable arguments $Args$ produced by the AFOSN (see Definition 8). Thus, we avoid the problem of using arguments that are not coherent with a situation where a potential privacy violation is happening and whose persuasiveness would be nil. The persuasive aspect related to coherence is therefore granted by the underlying computational argumentative reasoning.

5 Persuasive and behaviour change evaluation

To evaluate the persuasive power of the arguments generated by our argument-based persuasive framework w.r.t. behaviour change, we have designed a study that is divided

into two stages. The APF is used to persuade OSN users in order to prevent potential privacy violations. In the first stage, we collect user modelling inputs (i.e. personality traits and OSN behaviour); in the second stage, we measure the persuasive power of the arguments generated by our APF by considering the user modelling inputs and comparing them with a random selection method. For this purpose, a set of abstract arguments is generated for each potential privacy-threatening publication using an AFOSN, and its semantics are used to determine the set of acceptable arguments. Then, the persuasive γ function is used to improve the persuasiveness of the argumentative reasoning provided by the argumentation framework. In view of the results of the persuasive policy learning task, we decided to use the SVR model to estimate the optimal persuasive policies for the users who were participating in our evaluation.

To analyse the influence the content of the post on the persuasive power of the arguments, six different types of content were included in the experiment (see Table 3 of Appendix A.3): location, medical, alcohol/drugs, personal, family/association, and offensive.

5.1 Participants

For this experiment, 50 participants (25 male and 25 female) ranging in age between 18 and 44 years old ($\mu = 25.72, \sigma = 5.18$) were recruited. We required the participants to have experience using at least one social network.

In order to keep parity between age and gender, we divided the participants into two groups: experimental and control. The experimental group consisted of 30 participants (15 males and 15 females) ranging in age between 20 and 33 years old ($\mu = 25.87, \sigma = 4.22$). The control group was composed of 20 participants (10 males and 10 females) ranging in age between 18 and 44 years old ($\mu = 25.5, \sigma = 6.48$).

5.2 Materials

For the first stage concerning the acquisition of user modelling inputs, we designed an online questionnaire that was composed of two sections. In the first section, we asked the participants to answer a set of questions based on the 50-item personality inventory (Goldberg 1999) along with three attention check questions using the same questionnaire as in Sect. 4.1; in the second section, we asked the participants to complete the OSN interaction questionnaire (*Questionnaire C* described in Sect. 4.1 and shown in Appendix A3) along with one attention check question.

In the second stage, in which the persuasive power of the arguments generated by our argument-based persuasive framework was evaluated, we designed an online questionnaire composed of fourteen sections. In each section, a scenario in which a post (consisting of a message and an image) containing sensitive material that could violate the user's privacy was presented (see Fig. 6). The post was followed by an argument that attempted to convince the user to modify the original post in order to preserve his or her privacy. To evaluate the persuasive power of the argument, the participants were asked whether or not they would publish the post after reading the argument and also their degree of trust regarding this decision. To capture the degree

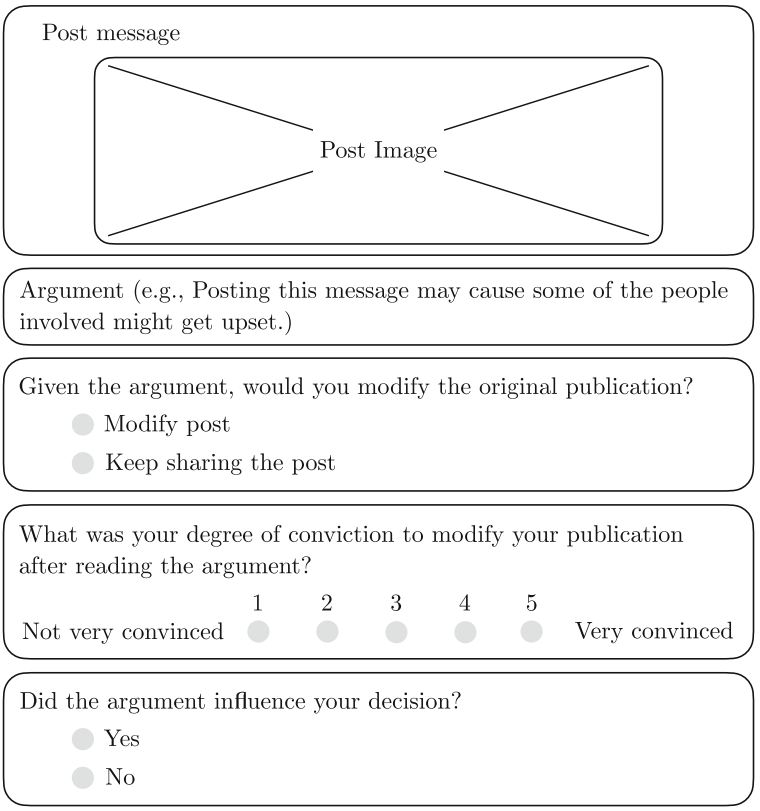

Fig. 6 Experiment layout

of trust, we used a 5-item Likert scale ranked from "not very convinced" to "very convinced". To measure the impact of the arguments on the participants' decisions, at the end of the section, the participants were asked whether or not the argument had influenced their decision.

There were fourteen sections in total: two sections were for attention monitoring, and twelve sections represented the six types of arguments (two sections per type of argument content) that were randomly distributed. The sections dedicated to attention monitoring followed a similar pattern to the twelve sections in order to determine if the participants were actually reading the questions carefully and not answering randomly.

With regard to the selection of the arguments to be presented to the participants during the second stage of the experiment, the experimental group received arguments that were generated by the argument-based persuasive framework. The control group received arguments whose reasoning pattern was randomly chosen and instantiated to natural language. Likewise, the type of argument was also randomly selected, but only those types that made sense with the context of the question were considered.

5.3 Procedure

The two stages of the experiment were performed on different days to avoid biases. At the beginning of each stage of the experiment, the participants were provided with the instructions describing the task to be accomplished. Then, the participants were asked to complete the questionnaires without a time limit.

5.4 Results

The results of the experiment show differences between the control group and the experimental group when making the decision of whether or not to publish a post on a social network. Thus, we observed that, in the control group, the participants who chose to modify the post after reading the argument reported that the argument had influenced their decision (30.41% of the group). This result contrasts with the 37.7% obtained in the experimental group. Therefore, by personalising the arguments to the users' characteristics, we obtained better effectiveness in modifying their behaviour. To analyse the statistical difference in the participants' behaviour according to the arguments in the two groups, we performed a Chi-square test. The results of the analysis show significant statistical evidence between the control group and the experimental group with a Chi-square value of 10.57 and a p-value of 0.014 (for a critical value of 7.82 and 3 degrees of freedom). These results also confirm that arguments that are generated according to user-specific persuasive policies improve the persuasiveness of an argumentation system.

With regard to the type of content of the arguments (see Table 3 of Appendix A.3), we found that, in general, there was a greater change in user behaviour in the experimental group compared to the control group in five of the six types analysed (all except personal content). In the sections related to medical content, 28.33% of the participants in the experimental group modified their behaviour after being influenced by the argument compared to 15% of the control group. The same can be observed for the offensive content, where 66.67% of the participants of the experimental group modified their behaviour compared to 55% of the control group. For family/association and alcohol/drugs, the experimental group was influenced by the argument (26.67% and 35%, respectively), while the control group was only influenced by 17.5% and 22.5%, respectively. However, in the case of personal content, we found that 48.88% of participants in the experimental group modified their behaviour after being influenced by the argument versus 50% in the control group. This may be due to the sensitivity of the content of the post. We observed that the in the experimental group, the posts related to personal content and to offensive content were more sensitive since, in general, the participants modified their behaviour (49% and 62%, respectively). In contrast, the medical content, the family content, and the location content showed less sensitivity and less probability of behavioural change influenced by an argument (23%, 23%, 17%, respectively).

With regard to the level of trust, we found that the mean of the degree of trust that users showed when modifying their behaviour based on an argument was $\mu = 4.23$ (with $\sigma = 0.85$) out of a maximum of 5. In contrast, the mean of the degree of trust of

the participants who decided not to modify their behaviour was only $\mu = 2.58$ (with $\sigma = 0.81$). This is an interesting result which indicates that the use of arguments to persuade users' behaviour reinforces their degree of trust in their decision when modifying their behaviour on a social network. These results highlight the importance of research into the use of persuasive argumentation systems in applications that seek to study, interpret, or modify human behaviour.

6 Discussion

Abstract argumentation frameworks have been extensively used in the field of computational argumentation to encode argumentative data and to approximate argumentative reasoning through the use of argumentation semantics. Research on this topic has been focused on proving and refuting logical properties and formulae, rather than extending their functions to other areas such as natural language processing or computational persuasion.

The ideas of extending formal computational argumentation concepts to the area of computational persuasion have been explored in recent research (Hunter 2018). The authors propose a general framework for computational persuasion for behaviour change applications where computational argumentation is introduced as a promising approach to solve this problem. A complete analysis of the existing research and proposed techniques is done, but no specific proposal or implementation is presented. Some of these ideas are further developed in Hadoux and Hunter (2019). However, argumentation frameworks are considered to be mere graph data structures, and argumentation semantics are removed from the computational argumentative reasoning process. Thus, it is not possible to explore the benefits of combining the coherence and rationality provided by argumentative reasoning together with personalised persuasive interactions that are aimed at behaviour change. A more ambitious effort at combining aspects from formal computational argumentation theory and computational persuasion is done in Rosenfeld and Kraus (2016). The authors propose a persuasive agent that approaches argumentative reasoning through a weighted argumentation framework and its quantitative semantics. Arguments are then used in a dialogue with human users following strategies learnt by a partially observable Markov decision process. The results achieved by the agent show 20% of cases where human users decided to change their behaviour. However, a small population was used to evaluate the argumentative agent (i.e. 15 participants).

In order to overcome the identified limitations, we have proposed a generalised framework for extending formal computational argumentation techniques to the area of computational persuasion. The main contributions of our proposal are twofold. First, we have formalised a general framework for argument-based computational persuasion that is designed to work with any underlying argumentation framework considering different user models. Our APF is not constrained to any specific argumentation framework, semantics, or user model, and it can be instantiated on top of any computational argumentative algorithm that provides a set of acceptable arguments, regardless of the domain or how the algorithm is approached (i.e. quantitative or qualitative). Furthermore, the APF also includes a persuasive function that is not constrained

to any specific implementation. It is important to emphasise that our approach to the persuasive function γ is not the only valid one. Throughout Sects. 4.1 and 4.2, we presented an implementation proposal of the γ of the APF's that is formally defined at the beginning of this paper. However, other approaches for generating a natural language argument from the set of acceptable arguments of an argumentation framework can also be proposed. The only requirement is that the γ function approach must take into account a user model and a set of linguistic features in addition to the acceptable abstract arguments. Second, we provide a complete implementation of the APF in a real case study and a persuasive evaluation with real human users. In our proposal, we model our human users considering two different sets of user modelling features: personality and online behaviour (e.g. number of friends, comments, or likes). Through our implementation, it is possible to observe how the different parameters of the APF need to be instantiated. Furthermore, at the end of our experiments, we validated the proposed persuasive framework since it significantly improves the persuasiveness of an argumentation system that is aimed at preventing privacy violations in OSNs.

Compared to previous research, our approach enables the use of computational argumentative reasoning techniques for approaching and improving the computational persuasion task. Our proposal and results present a significant contribution to the user modelling and personalised computational interaction of argumentative systems. However, there are some limitations in our work. First, the proposed implementation and results of the evaluation are constrained to our domain. We have implemented the APF for the domain of privacy management in OSNs, and our implementation cannot be extrapolated to any other different domain. The same goes for the results. The reported improvement in persuasive performance caused by the use of the APF might differ between different domains and implementations. For example, using different user models or taking a different approach to the implementation of the persuasive function γ may result in significant variations of the perceived persuasiveness of our system by human users. Second, our implementation of the APF has been evaluated using a series of one-shot interactions with the users. Our experiments have not been designed to investigate the definition of persuasive strategies in a dialogue but to estimate persuasive policies in order to persuade user with individual arguments.

7 Conclusion

In this paper, we have proposed argument-based persuasive frameworks. APFs extend the computational argumentative reasoning provided by argumentation frameworks and enable a persuasive interaction with human users. Thus, an argumentation system can computationally approach human argumentative reasoning through an argumentation framework and its semantics and broaden its purposes to persuasive and personalised interaction with human users. In addition to the definition, we have proposed a use case of the APF that is framed within the domain of privacy management in OSNs, and we have provided a complete implementation of the framework in a real situation. We implemented the APF on top of an argumentation framework that is specifically defined for its use in OSNs (i.e. AFOSN), and we modelled our users taking into account their personality and their online behaviour (e.g. number of

friends, comments, or likes). Furthermore, we conducted a persuasive evaluation of our proposal, where we observed that the use of an APF on top of an argumentation framework improves the persuasiveness of the arguments used by the argumentation system during the interaction with human users. We have also observed that the trust placed by human users in an interactive system that provides arguments for behaviour change is really high, meaning that argumentation is a powerful technique for designing trusted and reliable decision support systems. Therefore, the extension of argumentation frameworks for their use in persuasive systems represents a step forward that helps in the convergence between formal computational argumentation and human-computing interaction research.

With all of these findings, we foresee further research at the intersection of the two research areas of computational argumentation and computational persuasion. Specifically, these include analysing different user models, linguistic features, and persuasive functions, in addition to research on the relation between these variables and the application domain. We also find it important to investigate how the APF could be implemented or extended to interact directly with human users in argumentative dialogues.

Acknowledgements This work is partially supported by the Spanish Government projects PID2020-113416RB-I00, the Generalitat Valenciana project CIPROM/2021/077, and TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

A Questionnaires

A.1 Argumentation schemes

See Tables 2 and 3.

Table 2 Arguments used in *Questionnaire A* to represent the five different argumentation schemes evaluated

Scheme	Argument
AFCQ	Making the publication could have bad consequences for your privacy
	Making the publication may result in some of your sensitive data publicly available
	If you share this publication, the revealed information will be completely available for the Social Network users
AFPP	Most of your friends would not publish this content
	People usually do not share publications like this one
	It is not a popular practice among other users to share publications like this one
AFPO	Everyone knows that publishing this is a mistake
	It is widely known that publications of this type will endanger your personal information
	There is no doubt that sharing a publication like this one can be dangerous for your privacy
AFEO	Experts in social network privacy management believe that making a publication of this type could be dangerous
	Based on previous privacy studies, experts do not recommend to share this publication
	With this publication you may be endangering your privacy, as indicated by a privacy expert
AFWT	A user of the social network who has made similar publications considers that it can be dangerous
	Another user has experienced privacy issues when sharing similar publications
	Based on a similar user activity, this publication may endanger your privacy

Table 3 Arguments used in *Questionnaire B* to represent the four different argument types evaluated

Type	Argument (You should not make this publication because...)
Privacy	You should select a more restrictive privacy policy
Trust	Some of the tagged people might get upset
Risk	It could be read by strangers
Content	You are revealing your location. (Location)
	You are giving out personal medical information. (Medical)
	Others may think you consume alcohol/drugs. (Alcohol/Drugs)
	You are revealing sensitive personal information. (Personal)
	You are revealing family sensitive information. (Family/Association)
	You might offend other users. (Offensive)

A.2 Argument types

OSN interaction data

See Table 4.

Table 4 Items used in *Questionnaire C* to measure OSN interaction data of our participants

How often do you do the following activities on your social networks?	Possible Answers
Add as much users as I can	Never, occasionally, sometimes, usually, and always
Make publications	
Like other users' posts	
Comment on other users' posts	
Disclose my posts publicly (all users)	
Disclose my posts just with Friends/Followers	
Disclose my posts just with specific users (or groups)	
Make private publications (Only accesible for me)	
Upload pictures	
Delete my posts because of regrets	
Write long texts	
I am connected to my social network profiles... (per day)	Less than 2 h, between 2 and 4 h, between 4 and 6 h, and more than 6 h

Argument database

- **Argument from Consequences:**
  - *Privacy:* Posting this message may breach the privacy preferences you have set for this network.
  - *Risk:* Posting this message may cause it to be read by people you don't know.
  - *Trust:* Posting this message may cause some of the people involved might get upset.
  - *Content—Location:* Posting this message may reveal sensitive data about your location.
  - *Content—Medical:* Posting this message may reveal sensitive data about your medical conditions.
  - *Content—Drug:* Posting this message may reveal sensitive data about the use of drugs.
  - *Content—Personal:* Posting this message may reveal sensitive personal information.
  - *Content—Family:* Posting this message may reveal sensitive data about your relatives.
  - *Content—Offensive:* Posting this message may offend other users.
- *Argument from Popular Practice:*
  - *Privacy:* Most people with your privacy settings would not post this message.
  - *Risk:* Most people would not take the risk of this message reaching the wrong people.
  - *Trust:* Most people would not post a message that may anger other people involved.
  - *Content—Location*: Most people would not post a message that reveals data about their location.
  - *Content—Medical:* Most people would not post a message that reveals data about their medical conditions.
  - *Content—Drug:* Most people would not post a message that reveals information about using drugs.
  - *Content—Personal:* Most people would not post a message that reveals personal information.
  - *Content—Family:* Most people would not post a message that reveals data about their relatives.
  - *Content—Offensive:* Most people would not post a message that may offend other users.
- **Argument from Popular Opinion:**
  - *Privacy:* Most people think that posting this message may breach the privacy preferences you have set for this network.
  - *Risk:* Most people think that posting this message may cause it to be read by inappropriate people.
  - *Trust:* Most people think that this message may anger other people involved.

  - *Content—Location:* Most people think that this message reveals sensitive data about location.
  - *Content—Medical:* Most people think that this message reveals sensitive data about medical conditions.
  - *Content—Drug:* Most people think that this message reveals inappropriate data about using drugs.
  - *Content—Personal:* Most people think that this message reveals sensitive personal information.
  - *Content—Family:* Most people think that this message reveals sensitive data about your relatives.
  - *Content—Offensive:* Most people think that this message may offend other users.

- *Argument from Expert Opinion:*
  - *Privacy:* According to privacy experts, posting this message may breach the privacy preferences you have set for this network.
  - *Risk:* According to privacy experts, posting this message may cause it to be read by inappropriate people.
  - *Trust:* According to privacy experts, this message may anger other people involved.
  - *Content—Location:* According to privacy experts, this message may reveal sensitive data about your location.
  - *Content—Medical:* According to privacy experts, this message may reveal sensitive data about your medical conditions.
  - *Content—Drug:* According to privacy experts, this message may reveal inappropriate data about the use of drugs.
  - *Content—Personal:* According to privacy experts, this message may reveal sensitive personal information.
  - *Content—Family:* According to privacy experts, this message may reveal sensitive data about your relatives.
  - *Content—Offensive:* According to experts, this message may offend other users.

- **Argument from Witness Testimony:**
  - *Privacy:* According to users that have posted similar messages, posting this message may breach the privacy preferences you have set for this network.
  - *Risk:* According to users that have posted similar messages, posting this message may cause it to be read by inappropriate people.
  - *Trust:* According to users that have posted similar messages, this message may anger other people involved.
  - *Content—Location:* According to users that have posted similar messages, this message may reveal sensitive data about your location.
  - *Content—Medical:* According to users that have posted similar messages, this message may reveal sensitive data about your medical conditions.
  - *Content—Drug:* According to users that have posted similar messages, this message may reveal inappropriate data about using drugs.

- *Content—Personal:* According to users that have posted similar messages, this message may reveal sensitive personal information.
- *Content—Family:* According to users that have posted similar messages, this message may reveal sensitive data about your relatives.
- *Content—Offensive:* According to users that have posted similar messages, this message may offend other users.

References

Acquisti, A., Adjerid, I., Balebako, R., Brandimarte, L., Cranor, L.F., Komanduri, S., Leon, P.G., Sadeh, N.M., Schaub, F., Sleeper, M., Wang, Y., Wilson, S.: Nudges for privacy and security: Understanding and assisting users' choices online. ACM Comput. Surv. **50**(3), 44–14441 (2017)

Alahmari, S., Yuan, T., Kudenko, D.: Reinforcement learning for dialogue game based argumentation. In: Proceedings of the 19th Workshop on Computational Models of Natural Argument co-located with the 14th International Conference on Persuasive Technology, CMNA@PERSUASIVE 2019, Limassol, Cyprus, vol. 2346, pp. 29–37 (2019)

Atkinson, K., Baroni, P., Giacomin, M., Hunter, A., Prakken, H., Reed, C., Simari, G.R., Thimm, M., Villata, S.: Towards artificial argumentation. AI Mag. **38**(3), 25–36 (2017)

Baff, R.E., Wachsmuth, H., Khatib, K.A., Stein, B.: Analyzing the persuasive effect of style in news editorial argumentation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, pp. 3154–3160 (2020)

Baroni, P., Caminada, M., Giacomin, M.: An introduction to argumentation semantics. Knowl. Eng. Rev. **26**(4), 365–410 (2011)

Baroni, P., Romano, M., Toni, F., Aurisicchio, M., Bertanza, G.: Automatic evaluation of design alternatives with quantitative argumentation. Argum. Comput. **6**(1), 24–49 (2015)

Bottou, L.: Stochastic gradient descent tricks. Neural Netw.: Tricks Trade—Second Ed. **7700**, 421–436 (2012)

Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)

Chalaguine, L.A., Hunter, A., Potts, H.W.W., Hamilton, F.: Impact of argument type and concerns in argumentation with a chatbot. In: 31st IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2019, Portland, pp. 1557–1562 (2019)

Chalaguine, L.A., Hunter, A.: A persuasive chatbot using a crowd-sourced argument graph and concerns. Comput. Models Argum. **2020**(326), 9–20 (2020)

Ciocarlan, A., Masthoff, J., Oren, N.: Actual persuasiveness: impact of personality, age and gender on message type susceptibility. In: Persuasive Technology: Development of Persuasive and Behavior Change Support Systems—14th International Conference, PERSUASIVE 2019, Limassol, Cyprus, Proceedings, vol. 11433, pp. 283–294 (2019)

Cocarascu, O., Toni, F.: Mining bipolar argumentation frameworks from natural language text. In: Proceedings of the 17th Workshop on Computational Models of Natural Argument co-located with ICAIL 2017, vol. 2048, London, pp. 65–70 (2017)

Cover, T.M., Hart, P.E.: Nearest neighbor pattern classification. IEEE Trans. Inf. Theory **13**(1), 21–27 (1967)

Donadello, I., Hunter, A., Teso, S., Dragoni, M.: Machine learning for utility prediction in argument-based computational persuasion. CoRR arXiv:2112.04953 (2021)

Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A.J., Vapnik, V.: Support vector regression machines. In: Advances in Neural Information Processing Systems 9, NIPS, Denver, 1996, pp. 155–161 (1996)

Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artif. Intell. **77**(2), 321–358 (1995)

Gleize, M., Shnarch, E., Choshen, L., Dankin, L., Moshkowich, G., Aharonov, R., Slonim, N.: Are you convinced? choosing the more convincing evidence with a siamese network. In: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, vol. 1, pp. 967–976 (2019)

Goldberg, L.R., et al.: A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. Personal. Psychol. Europe **7**(1), 7–28 (1999)

Gunawardana, A., Shani, G.: Evaluating recommender systems. Recommender Systems Handbook, pp. 265–308 (2015)
Hadoux, E., Hunter, A., Corrégé, J.: Strategic dialogical argumentation using multi-criteria decision making with application to epistemic and emotional aspects of arguments. In: Foundations of Information and Knowledge Systems—10th International Symposium, FoIKS 2018, Budapest, Proceedings 10833, pp. 207–224 (2018)
Hadoux, E., Hunter, A., Polberg, S.: Strategic argumentation dialogues for persuasion: framework and experiments based on modelling the beliefs and concerns of the persuadee. CoRR arXiv:2101.11870 (2021)
Hadoux, E., Hunter, A.: Comfort or safety? gathering and using the concerns of a participant for better persuasion. Argum. Comput. **10**(2), 113–147 (2019)
Hunter, A., Chalaguine, L.A., Czernuszenko, T., Hadoux, E., Polberg, S.: Towards computational persuasion via natural language argumentation dialogues. In: KI 2019: Advances in Artificial Intelligence—42nd German Conference on AI, Kassel, Germany, Proceedings, vol. 11793, pp. 18–33 (2019)
Hunter, A.: Towards a framework for computational persuasion with applications in behaviour change. Argument Comput. **9**(1), 15–40 (2018)
Khatib, K.A., Trautner, L., Wachsmuth, H., Hou, Y., Stein, B.: Employing argumentation knowledge graphs for neural argument generation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, vol. 1: Long Papers, Virtual Event, pp. 4744–4754 (2021)
Khatib, K.A., Völske, M., Syed, S., Kolyada, N., Stein, B.: Exploiting personal characteristics of debaters for predicting persuasiveness. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, pp. 7067–7072 (2020)
Khatib, K.A., Wachsmuth, H., Hagen, M., Stein, B.: Patterns of argumentation strategies across topics. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, pp. 1351–1357 (2017)
Kökciyan, N., Yaglikci, N., Yolum, P.: An argumentation approach for resolving privacy disputes in online social networks. ACM Trans. Internet Techn. **17**(3), 27–12722 (2017)
Lawrence, J., Reed, C.: Argument mining: a survey. Comput. Linguist. **45**(4), 765–818 (2019)
McBurney, P., Parsons, S.: Games that agents play: A formal framework for dialogues between autonomous agents. J. Log. Lang. Inf. **11**(3), 315–334 (2002)
Monteserin, A., Amandi, A.: A reinforcement learning approach to improve the argument selection effectiveness in argumentation-based negotiation. Expert Syst. Appl. **40**(6), 2182–2188 (2013)
Mosca, F., Such, J.M.: An explainable assistant for multiuser privacy. Auton. Agents Multi Agent Syst. **36**(1), 10 (2022)
Rosenfeld, A., Kraus, S.: Strategical argumentative agent for human persuasion. In: ECAI 2016 22nd European Conference on Artificial Intelligence, 29 August-2 September 2016, The Hague, The Netherlands—Including Prestigious Applications of Artificial Intelligence (PAIS 2016), vol. 285, pp. 320–328 (2016)
Rothmann, S., Coetzer, E.P.: The big five personality dimensions and job performance. SA J. Ind. Psychol. **29**(1), 68–74 (2003)
Ruiz-Dolz, R., Alemany, J., Heras, S., García-Fornes, A.: Automatic generation of explanations to prevent privacy violations. In: Proceedings of the 2nd EXplainable AI in Law Workshop (XAILA 2019) co-located with 32nd International Conference on Legal Knowledge and Information Systems (JURIX 2019), Madrid, Spain, vol. 2681 (2019)
Ruiz-Dolz, R., Alemany, J., Heras, S., García-Fornes, A.: On the prevention of privacy threats: how can we persuade our social network users? CoRR. arXiv:2104.10004 (2021)
Ruiz-Dolz, R., Heras, S., Alemany, J., García-Fornes, A.: Towards an argumentation system for assisting users with privacy management in online social networks. In: Proceedings of the 19th Workshop on Computational Models of Natural Argument co-located with the 14th International Conference on Persuasive Technology, CMNA@PERSUASIVE 2019, Limassol, Cyprus, vol. 2346, pp. 17–28 (2019)
Ruiz-Dolz, R., Heras, S., García-Fornes, A.: Automatic debate evaluation with argumentation semantics and natural language argument graph networks. CoRR arXiv:2203.14647 (2022)
Ruiz-Dolz, R., Taverner, J., Heras, S., García-Fornes, A., Botti, V.: A qualitative analysis of the persuasive properties of argumentation schemes. In: Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2022, Barcelona, In press (2022)

Ruiz-Dolz, R.: Towards an artificial argumentation system. In: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI, pp. 5206–5207 (2020)
Somasundaran, S., Ruppenhofer, J., Wiebe, J.: Detecting arguing and sentiment in meetings. In: Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue, SIGdial 2007, Antwerp, pp. 26–34 (2007)
Sutton, R.S., Barto, A.G.: Reinforcement learning: an introduction. IEEE Trans. Neural Netw. **9**(5), 1054 (1998)
Thomas, R.J., Masthoff, J., Oren, N.: Can I influence you? development of a scale to measure perceived persuasiveness and two studies showing the use of the scale. Front. Artif. Intell. **2**, 24 (2019)
Vapnik, V.: The support vector method of function estimation. Nonlinear Model. (1998). https://doi.org/10.1007/978-1-4615-5703-6
Walton, D.: Argumentation theory: a very short introduction. In: Argumentation in Artificial Intelligence, pp 1–22 (2009)
Walton, D., Reed, C., Macagno, F.: Argumentation Schemes. Cambridge University Press (2008)
Wiebe, J., Riloff, E.: Creating subjective and objective sentence classifiers from unannotated texts. In: Computational Linguistics and Intelligent Text Processing, 6th International Conference, CICLing 2005, Mexico City, vol. 3406, pp. 486–497 (2005)
Xia, M., Zhu, Q., Wang, X., Nie, F., Qu, H., Ma, X.: Persua: A visual interactive system to enhance the persuasiveness of arguments in online discussion. CoRR arXiv:2204.07741 (2022)

Ramon Ruiz-Dolz is a Postdoctoral Researcher with the Centre for Argument Technology (ARG-tech), University of Dundee, Dundee, United Kingdom, and is currently doing research in computational argumentation and natural language processing. His research interests are focused on argument mining, computational argumentation and persuasion technologies. He received recently the Ph.D. in Computer Science from the Universitat Politècnica de València in 2023.

Joaquin Taverner received the Ph.D. degree in Computer Science (Artificial Intelligence) from the Universitat Politècnica de València in 2022. He is a pos-doctoral researcher in Computer Science at VRAIN, (UPV), Spain. His main research interests include multi-agent systems, computational argumentation, and affective computing.

Stella M. Heras Barberá is a Researcher with the Valencian Research Institute for Artificial Intelligence (VRAIN), Valencia, Spain and an Associate Professor with the Department of Languages and Computer Systems, Polytechnic University of Valencia (UPV), Valencia, Spain. Her research area is focused on the development of artificial intelligence systems (computational argumentation, persuasion technologies, educational recommender systems). She received the Ph.D. degree in computer science (extraordinary prize Cum Laude) from the Polytechnic University of Valencia (UPV), Valencia, Spain.

Ana García-Fornes is a Full Professor with the Department of Information Systems and Computation, Polytechnic University of Valencia (UPV), Valencia, Spain and a Researcher with the Valencian Research Institute for Artificial Intelligence (VRAIN). Her research interests focus on real-time systems, multiagent systems, agreement technologies, and privacy in social media. She received the Ph.D. degree in computer science from the UPV.