Towards Ethical Argumentative Persuasive Chatbots

Caren Al Anaissy, Srdjan Vesic, Nathalie Nevejans

Towards Ethical Argumentative Persuasive Chatbots

Caren Al Anaissy[1][0000−0002−8750−1849], Srdjan Vesic[2][0000−0002−4382−0928], and Nathalie Nevejans[3][0000−0003−0030−2984]

[1] CRIL Université d'Artois & CNRS, France
alanaissy@cril.fr
[2] CRIL CNRS Univ. Artois, France
vesic@cril.fr
[3] Research Center in Law, Ethics and Procedures, Faculty of Law of Douai, University of Artois, France
nathalie.nevejans@univ-artois.fr

Abstract. Argumentative persuasive technologies are technologies that use argumentation in order to persuade the persuadee to believe in something or not, which can later lead the persuadee to perform an action or not. The use of such tools opens numerous ethical considerations. In this paper, we survey the literature on persuasion that might be useful for argumentative persuasive chatbots, we cover the existing legal framework and ethical principles and we critically analyze the new proposal for a regulation on artificial intelligence of the European Commission. We also show how to use argumentation to enhance explainability and transparency of the persuasion systems. We propose to show the graphical representation of the arguments used during the persuasion to the user at the end of the dialogue, containing the relations between the arguments (attacks, supports), their origin (source), who uttered them (i.e. the machine or the human participant) and the persuasive methods employed. Our approach has several benefits. Namely, it makes the system more transparent and enhances the human understanding of the system, which is a benefit per se. Furthermore, the fact that the system is transparent increases the trust of the user, which (apart from being one of the goals of AI in general) can increase the chance that the user is persuaded by the system. Finally, the user can give a feedback on the presented arguments (e.g. how much they believe the arguments are ethical), which can be later used to improve the persuasion system.

Keywords: Persuasive chatbots · Ethics of persuasive technology · Computational argumentation.

1 Introduction

Persuasion aims to change people's attitudes or behaviours [16]. Persuasive technologies are very powerful tools because of the simulation they can create. They help shaping the perception and interpretation of reality to users by amplifying specific perceptions and reducing others. They also help shaping the users'

actions in reality by encouraging specific forms of actions and discouraging others [32]. Persuasive technologies can also convey social presence when interacting with the user. Many studies have identified persuasive strategies that can be used to influence users and enhance persuasion. Argumentative persuasive chatbots deploy argumentation graphs as their knowledge base. Some of the chatbots even use argumentation semantics for decision-making problems [9]. However, the persuasive acts employed by these chatbots might sometimes be considered morally and ethically unacceptable. Note that the chatbot itself cannot take responsibility for the methods and outcomes of the persuasive acts, because it is not capable of forming its own intentions or making its own choices. Hence, it is not a free moral agent [7].

Persuasion in argumentation has received a large amount of attention during the last years [6, 20, 19, 18, 33, 30]. Also, ethics in argumentation has been studied with the focus on the use of normative systems [28, 5, 23].

However, we can observe that the researchers and practitioners who develop argumentative persuasive chatbots are not always aware of the existing advances and the state of the art in the domain of persuasion as well as of the current legal framework and corresponding ethical considerations. This is why the first goal of our paper is to make a bridge between the existing knowledge in persuasion and ethics on one side and the practitioners who develop argumentative persuasive chatbots on the other side.

This is why the **first part of the paper** is devoted to a **survey** of the literature on persuasion related to argumentative persuasive chatbots. We cover the state of the art of persuasion for persuasive chatbots and the corresponding ethical guidelines (Section 2). Then, we survey the existing legal framework and its link with ethical principles and critically analyze the new proposal for a regulation on artificial intelligence of the European Commission (Section 3). We believe that this overview will be useful for researchers in argumentation who want to deploy an argumentative chatbot for persuasion since it allows them to quickly get knowledge about the most relevant approaches at one place.

The **second part of the paper** is devoted to the question: how can we increase transparency, trust and effectiveness of the argumentative persuasive chatbots? We propose a method that can enhance explainability and transparency of the persuasion systems (Section 4). The main idea is to show the graphical representation of the arguments used during the persuasion to the user at the end of the dialogue. This representation contains the relations between the arguments (attacks, supports), their origin (source), who uttered them (i.e. the machine or the human participant) and the persuasive methods employed. This approach has several advantages. Namely, it makes the system more transparent and enhances the human understanding of the system, which is a benefit per se. Furthermore, the fact that the system is transparent increases the trust of the user, which (apart from being one of the goals of AI in general) can increase the chance that the user will be persuaded by the system. Finally, the user can give a feedback on the presented arguments (e.g. how much they believe the arguments are ethical), which constitutes valuable data, which can later be used to better

understand the underpinnings of human reasoning and improve the persuasion system.

The paper is organised as follows. Section 2 presents related work in persuasion that is relevant for argumentative persuasive chatbots and corresponding ethical guidelines. Section 3 surveys the legal and ethical aspects. Section 4 explains the pillars of our idea to use the graphical argumentation-based representation in order to enhance transparency, trust and efficiency of persuasion systems. We then talk about possible future work and conclude.

2 State of the Art in Persuasion

This section provides background material for argumentative persuasive chatbots and the corresponding ethical guidelines.

2.1 Persuasive Strategies

In this subsection, we review some of the persuasive strategies that can be used to enhance persuasion. Fogg was able to identify and propose forty principles that persuasive technologies can use [16]. The forty principles are classified into six categories. The first three categories are the three persuasive roles computing techonlogies can play. Computers can behave as persuasive tools, persuasive media and persuasive social actors. The three other categories study how computers and web pages can be more persuasive through credibility, mobility and connectivity. Table 1 shows the different principles proposed by Fogg classified into the six categories. We discuss only the principles that seem most relevant for persuasive chatbots. We present briefly the principles related to the role of persuasive technologies as social actors. As Fogg explains, there exist five essential types of social cues that persuasive technologies can convey. First, physical cues can be transmitted through the physical characteristics. Fogg proposes the principle of Attractiveness: "A computing technology that is visually attractive to target users is likely to be more persuasive as well." Second, persuasive technologies can use psychological cues to persuade. Fogg defines a person's psychology as the group of emotions, preferences, motivations and personality. Humour and empathy can also be considered as psychological cues. For example, a chatbot possesses empathy when it acknowledges the user's feelings, shows compassion and supports the user. The expressions like "I understand your feeling" and "I am sorry to hear that" are signs of having empathy towards the user. The principle of Similarity states: "People are more readily persuaded by computing technology products that are similar to themselves in some way." Third, we have the language. The principle of Praise states that: "By offering praise, via words, images, symbols, or sounds, computing technology can lead users to be more open to persuasion." Fourth, social dynamics such as giving praise, cooperation and reciprocity can also be used. The principle of Reciprocity proposed by Fogg states that: "People will feel the need to reciprocate when computing technology has done a favour for them." Finally, adopting a social role can be considered

as a very effective persuasive technique. Specifically, adopting an authority role seems to be very effective for persuasion. The principle of Authority identified by Fogg is the following: "Computing technology that assumes roles of authority will have enhanced powers of persuasion." We also present some of the principles related to "credibility and computers" which can be considered relevant for persuasive chatbots. Fogg explains that there are two element keys for credibility: trustworthiness and expertise. Trustworthiness represents how much truthful, fair and unbiased a source can be perceived. The principle of Trustworthiness states: "Computing technology that is viewed as trustworthy (truthful, fair and unbiased) will have increased powers of persuasion." Expertise represents how much perceived knowledge, skill and experience a source can have. The principle of Expertise states: "Computing technology that is viewed as incorporating expertise (knowledge, experience and competence) will have increased powers of persuasion." Fogg also points out to the fact that computing technology tends to lose credibility easily if commits a significant error. Once the credibility is lost, it may be hard to regain. Therefore, Fogg proposes the principle of (Near) Perfection that states: "Computing technology will be more persuasive if it never (or rarely) commits what users perceive as errors."

Computers as Persuasive Tools	Computers as Persuasive Media	Computers as Persuasive Social Actors
Reduction	Cause and Effect	Attractiveness
Tunnelling	Virtual Rehearsal	Similarity
Tailoring	Virtual Rewards	Praise
Suggestion	Simulations in Real-World Context	Reciprocity
Self-Monitoring		Authority
Surveillance		
Conditioning		
Credibility and Computers	**Credibility and the World Wide Web**	**Mobility and Connectivity**
Trustworthiness	"Real World Feel"	Kairos
Expertise	Easy Verifiability	Convenience
Presumed Credibility	Fulfilment	Mobile Simplicity
Surface Credibility	Easy-of-Use	Mobile Loyalty
Reputed Credibility	Personalization	Mobile Marriage
Earned Credibility	Responsiveness	Information Quality
(Near) Perfection		Social Facilitation
		Social Comparison
		Normative Influence
		Social Learning
		Competition
		Cooperation
		Recognition

Table 1. The forty persuasion principles proposed by Fogg for persuasive technologies.

Cialdini was able to identify six influence principles: Reciprocity, Commitment and Consistency, Social Proof, Liking, Authority, Scarcity [11]. Among these principles, we explain briefly the ones that were not explained before. The Commitment and Consistency principle states that humans tend to commit to

their opinions, values and choices. The Social Proof principle states that people tend to imitate other people's ideas and actions. For example, if a person is considering buying a product online but they are not sure whether the product is good or convenient for them, they can check the product's reviews section. If most of the reviews are positive, the user tends to feel more confident in their decision in purchasing the product. The Liking principle is close to the Similarity principle proposed by Fogg, it states that people like others who are similar to them. The Scarcity principle states that people tend to value opportunities or things that become less available, hence scarce. Oinas-Kukkonen and Harjumaa proposed and developed a Persuasive Systems Design model [26] where twenty-eight persuasive strategies were listed for the design of persuasive technology. These principles were divided into four categories: primary task, dialogue, system credibility, and social support.

Wang et al. were able to identify ten persuasion strategies that are divided into two types, the persuasive appeal and the persuasive inquiry type. The persuasive appeal type consists of trying to appeal to the persuadee's psychology. The persuasive inquiry type consists of asking the persuadee personal questions to facilitate the persuasion [35]. The persuasive appeal strategies identified in this work are the following: The Logical appeal strategy consists of using evidence and reasons to convince the persuadee of the persuasion goal. The Emotional appeal strategy consists of evoking the persuadee's positive and/or negative emotions. The Credibility appeal strategy consists of citing information from objective sources in order to gain the persuadee's trust. The Foot-in-the-door strategy consists of asking the persuadee small requests first, then asking larger ones. The Personal story strategy consists of telling the persuadee stories about other people who were persuaded by the persuasion goal, focusing on the positive results of such persuasion. The Donation Information strategy consists of giving the persuadee information about the action or idea the persuader wants to convince them with. The persuasive inquiry strategies are the following: The Source-related inquiry strategy consists of asking the persuadee whether they are aware or not of the action or idea the persuader wants to persuade them with. The task-related inquiry strategy consists of asking the persuadee their own opinions and expectation concerning the persuasion goal. Finally the personal-inquiry strategy consists of asking the persuadee about their own personal experiences related to this persuasion goal.

.

2.2 Personalization in Persuasion

In this subsection, we briefly define personalized persuasion and we briefly review two works done in this field. Personalization plays an important role in enhancing persuasion. Personalized persuasion consists of using the user's personal information and background to enhance the outcome of persuasion [22, 27, 12, 24]. Apart from using the user's psychological cues to persuade them, personalization can appear in the form of trying to adapt the methods used during the process of persuasion based on the user's psychological profile and/ or personal

information. The goal of such adaptation is also the enhancement of the persuasion outcome. Kaptein et al. studied the effects of involving users in choosing a specific influence strategy for persuasion, disclosing the usage of such strategies and the use of multiple strategies simultaneously on user compliance to persuasive attempts [21]. The authors consider these results as guidelines for designing adaptive persuasion systems. Adaptive persuasion systems are persuasion systems that use different influence strategies based on the users' profiles in order to increase their influence on users. The authors proved that letting the user decide which strategy to adopt in order to persuade them is more effective than predicting the preference for a specific strategy based on behaviour or personality measures. The reason behind this is that if the user chooses which influence strategy to adopt, they will commit to choice that they have made. The authors have also proved that using a single preferred strategy is more effective than using the preferred strategy simultaneously with a non preferred one. Also, using the two strategies simultaneously was more effective than using the single non preferred strategy alone. The two persuasive strategies used in this work are the authority and the consensus strategies. Wang et al. studied how the variations of the persuasion strategies upon the user's psychological background, affect the persuasion outcome [35]. The results of the work done by Wang et al. [35] and Shi et al. [31] show that personalizing the persuasion strategies yields better persuasion outcomes.

2.3 Persuasive Chatbots

In this subsection, we briefly review four persuasive chatbots that were developed and designed. The first three chatbots use the Credibility Appeal (Trustworthiness and Expertise) strategy because all the arguments presented by the chatbots come from objective sources (scientific sources, governmental websites and experts) therefore the information presented by the chatbots are unbiased, fair and truthful. As for the fourth chatbot, it uses different persuasive strategies during the dialogue.

Altay et al. studied whether they can change people's opposite opinion regarding genetically modified food and genetically modified organisms by providing rebuttals to the counterarguments held against genetically modified organisms [2]. They have defined four conditions: the Control Condition consists of defining genetically modified organisms. The Consensus Condition informs about the important existence of the scientific consensus regarding the genetically modified organisms' benefits. The Counterarguments Condition presents first counterarguments against genetically modified organisms then rebuttals to these counterarguments, then rebuttals to the previous ones and so on. The user scrolls to check all the arguments which were presented in a clear dialogue structure. Finally, the Chatbot Condition consists of presenting all the counterarguments against genetically modified organisms to the user, the user can click on any counterargument, the rebuttal of the counterargument selected appears progressively. The user has the option to come back to the initial counterarguments and click on another one. The four conditions were compared against

each other. The Counterarguments Condition was found more persuasive than the Chatbot condition in the sense that spending more time reading when all counterarguments are available leads to more positive attitude changes than selecting only the most relevant counterarguments.

Hadoux and Hunter showed how preferences over types of concern can be used to enhance the persuasion in persuasive dialogues [17].The notion of concern is defined as an issue raised or addressed by an argument. Among the empirical studies conducted, a group of participants chatted with two types of chatbots called the baseline chatbot and the preference-based chatbot. The structure of the dialogue is the same for both chatbots, the first argument is the one proposed by the system and called the persuasion goal, then a menu of counterarguments is proposed by the system to the user. The user can select a set of counterarguments; this set is called a menu move. Then the system can counter the selected arguments by a set of arguments called the posit move. The dialogue continues in the same manner until there is no argument attacking any argument of the last move or if the user ends the dialogue by a null argument. A null argument means that the user does not choose any of the counterarguments presented by the system. For the baseline chatbot, for each counterargument selected by the user, the chatbot selects a rebuttal among the arguments that attack the counterargument randomly. For the preference-based system, a set of the user's preferences over concerns is available, for each counterargument selected by the user, the chatbot selects the rebuttal with which the most preferred type of concern is associated. The results show that using preferences over concerns enhances the persuasiveness of the dialogue.

Chalaguine and Hunter designed a persuasive chatbot to persuade users to take the Covid-19 vaccine [10]. The authors use the same notion of concern as in [17]. The chatbot first presents the persuasion goal, then the user provides a counterargument manually. The chatbot predicts the user's concern raised in the counterargument and replies with the first not yet used rebuttal that addresses the same concern. In case the chatbot could not identify the concern of the user i.e. the prediction is less than 40% in confidence, it replies with one among three default rebuttals. There exist only three default rebuttals. The dialogue ends when the chatbot cannot identify a concern and all of the three default rebuttals were already used. The authors have shown that this interactive chatbot is more persuasive than a static web page in which the users read the ten most common rebuttals used by the chatbot.

Shi et al. developed a persuasive chatbot that understands the user's input based on neural network models [31] and replies from the human responses that were collected from the previous work [35]. The authors conducted experiments to study the effect of disclosing the chatbot's identity (bot or human) and the effects of using different type of inquiries (personal and/or non personal inquiries) on the persuasiveness of the dialogue. The chatbot salutes the user first, then the chatbot asks the user some questions based on the type of persuasive inquiry the user was assigned to. After finishing from the persuasive inquiry module, the chatbot moves on to the persuasive appeal module where the chatbot dialogues

with the user. At each step, the chatbot uses a different persuasive strategy and asks the user if they want to donate. If the user accepts to donate, that means that the chatbot succeeded in persuading the user to donate. If the user does not accept to donate, the chatbot uses another not yet used persuasive strategy to try to persuade the user to donate. The dialogue ends if the user agrees to donate or if the ten persuasive strategies were all used. Results showed that whether the chatbot was really human or not, it is the perceived identity of the chatbot by the user that matters. The persuasiveness is better when people think they are talking to human.

2.4 Argumentation Theory in Persuasive Dialogues

In this subsection, we review the main elements of a persuasive dialogue system introduced by Prakken [29]. Argumentation-based dialogues can be classified into six types based on their goal [4]. We have persuasion, negotiation, information seeking, deliberation, inquiry and quarrel. When it comes to argumentation-based dialogues, there are multiple rules to take into account. The communication language consists of the utterances the participants can make, the protocol consists of the conditions under which the participants can make the utterances, it also determines when the dialogue ends. According to Prakken [29], the main elements of a persuasive dialogue systems are the following:

- A dialogue goal which, in persuasion dialogues, is the resolution of a conflict of point of views between the participants.
- A topic language
- A logic for the topic language used, which can be monotonic or non-monotonic, it can be used to manage the dialogical consistency of participants.
- A communication language: As defined before, communication language consists of the allowed utterances to make. The most important ones are : claim ϕ, why ϕ, concede ϕ, retract ϕ, question ϕ and ϕ since S. As explained by Prakken [29], Claim ϕ means that the speaker asserts that ϕ is the case. Why ϕ means that the speaker challenges that ϕ is the case and asks for reasons why it would be the case. Concede ϕ means that the speaker admits that ϕ is the case. Retract ϕ means that the speaker declares that they are not committed (any more) to ϕ. Question ϕ means that the speaker asks another participant's opinion on whether ϕ is the case. Finally, ϕ since S means that the speaker provides reasons why ϕ is the case.
- A protocol: The protocol specifies the allowed moves at each step of the dialogue. The protocol specifies the dialogue's structure. We have unique-reply vs multi-reply, unique-move vs multi-move and immediate-reply vs non-immediate-reply, deterministic and fully deterministic vs non deterministic protocols.
- A set of participants with roles, internal beliefs and commitments. Usually the roles in a persuasion dialogue are proponent, opponent and neural toward a well specified topic. Commitments are very important because they determine the end of the dialogue and its outcomes, and they can be used to

oblige the participant to be dialogically consistent. Commitments are usually determined by claim ϕ , concede ϕ and retract ϕ.
- Effect rules: The effect rules determine the effects of the speech acts on the participants' commitments.
- Outcome rules: The outcome rules define the outcomes of a dialogue which are in a persuasive dialogue the winners and the losers of the dialogue.

2.5 Ethics of Persuasive Chatbots

In this subsection, we cover the ethical guidelines proposed in the literature for the design of persuasive technologies and we try to orient some of them towards the design of persuasive chatbots precisely. Berdichevsky and Neuenschwander explain that when it comes to persuasion, both persuader and persuadee take full moral responsibility for the outcome [7]. In order to evaluate the ethics of persuasion itself, one should evaluate the persuader's motivations, the methods they employed and the outcome of persuasion.

Although persuasive technology and persuasive people have same motivations and use similar methods and strategies, persuasive technology has more persuasive potential because of the simulations they can embed leading to more realism. The difference between persuasion through technology and through person-to-person interactions relies in the methods used for persuasion and also probably the outcome. The ethics of persuasion seems to be insufficient to guide the design and implementation of persuasive technology. The authors wanted to reconsider the ethical guidelines for traditional persuasion methods when being applied by technology and not humans, and for the outcome i.e. the persuasion goal.

Therefore, Berdichevsky and Neuenschwander proposed a set of eight ethical principles and guidelines for the design and implementation of persuasive technology, with the consideration that the designers should be only responsible for intended and unintended reasonably predictable outcomes [7]. The first two principles state that the intended outcome of any persuasive technology and the motivations behind it should never be considered unethical if the persuasion was done without the technology or if the outcome happened independently of the persuasion. The third principle states that the designers of such technologies should take responsibility for all reasonably predictable outcomes of their use. The "Dual Privacy" principles state that the creators of persuasive technology should respect users' privacy when it comes to accessing their online personal information and sharing it with a third party. The "Disclosure" principle states that the designers must be transparent and clear about the motivations, methods and intended outcomes of such technology. The "Accuracy" principle states that the persuasive technology should always be honest and credible. Finally, the "Golden" principle states that the designers of persuasive technology should never use a persuasion goal that they themselves are not consent of being persuaded by it.

Verbeek [32] emphasised on the importance of integrating the ethical framework proposed by Berdichevsky and Neuenschwander [7] with the concept of "technological mediation" in order to better understand and predict unintended

outcomes, and take these outcomes into consideration when designing persuasive technology. As mentioned before, technology tends to shape human perception and interpretation of reality by amplifying some perceptions while reducing others. It also shapes human actions in reality by encouraging specific forms of actions while discouraging others. Persuasive technology mediates these effects between users and their environments, so when technology is used the way their designers intended, it is possible to have unintended and unexpected outcomes, Verbeek proposed to focus on all the mediation effects by doing a moral reflection along deontological and utilitarian lines. The deontological point of view means respecting the moral principles while the utilitarian point of view means balancing between the desirability for something and its costs for all the people involved. This moral reflection will take into consideration the intended persuasions, the methods of persuasion used with the emerging forms of mediation, and the outcomes of the mediation which are the consequences of the persuasive and mediating role of the technology. The ethical guidelines that the authors proposed are the following:

- The intended persuasions of the technology-in-design must cause no harm for the people using persuasive technologies and those affected by them being used, these intended persuasions must benefit these people and be fair (justice) to them.
- The methods of persuasion and forms of mediation must be disclosed (respect for autonomy), cause no harm in terms of privacy and be fair to all people.
- As for the outcomes of mediation, the designers must do a moral imagination of all the possible mediating roles of technology in human actions and experiences and then assess these mediations along the deontological and utilitarian lines.

Fogg proposed to apply a stakeholder analysis to identify all people affected by a persuasive technology, and what each stakeholder has to gain or lose [16].

- List all of the stakeholders.
- List what each stakeholder has to gain.
- List what each stakeholder has to lose.
- Evaluate which stakeholder has the most to gain.
- Evaluate which stakeholder has the most to lose.
- Determine ethics by examining gains and losses in terms of values.
- Acknowledge the values and assumptions you bring to your analysis.

Note that values differ from a culture to another. Hence, creators of persuasive technology must be careful about the culture in which they are embedding this technology, because with every different culture, comes different ethical issues.

We list below a set of guidelines for the design of persuasive chatbots inspired by the guidelines proposed by Fogg [16] for the design of persuasive technology.

- Users of persuasive chatbots should not be distracted by the number of questions or the difficulty of arguments, because this can stand in the way of their focus on the content in the chatbot. The chat must not be complicated or very lengthy.

- Creators of persuasive chatbots should not consider that the user has experience in the domain of the goal with which we want to persuade the user.
- Creators of persuasive chatbots should not include in the chat any links to download an application or something else.
- The user must be able to stop at anytime they want, or ask for clarification. The creators should be also careful about the cases where the user must have the ability to ask for a human intervention.
- Not only the creators of persuasive chatbots take responsibility when it comes to errors and damage to the user, but also the company that bought this technology, distributed and promoted it. We may have different companies through time, they all can be responsible.
- Manipulation can happen when the chatbot expresses negative or positive emotions towards the user, presents arguments that appeal to the user's positive/negative emotions, tells lies or false information, tells incomplete information, chats with children or mentally disabled people, presents threatening information or punishment. Negative emotions could be fear, angry, deception, impatience. Positive emotions could be celebration, rewarding, encouraging. Designers of persuasive chatbots should avoid manipulation at all costs. It is preferred that the chatbot does not express emotions at all or expresses the minimum and only for good cause.
- The chatbot should not be very sophisticated in a way that confuses the user whether the chatbot is a human or robot. The user must know that they are chatting with a robot.
- Designers of persuasive chatbots must test and supervise these chatbots when used to observe if there are any unintended outcomes that were not recognised before, or happen to a small number of people. This is how they should deal with unintended and unpredictable outcomes. They should also keep track of the conversations between the chatbot and the users, with the user's knowledge.
- Persuasive chatbots should not provide offers, promotions, advertisements or branding. They should be designed exclusively for persuasion.

Creators of persuasive chatbots are also invited to consult the guidelines for developers of conversational AI proposed by Microsoft [13].

3 Persuasive Chatbots between Ethics and Law

In this section, we present and discuss the legal issues that impact the design and implementation of chatbots, specifically persuasive chatbots. On April 21, 2021 the European Commission has published a proposal for a regulation on artificial intelligence [15], called the AI Act, which is currently under discussion and will soon be adopted [1]. The AI Act proposes a gradation of legal constraints according to the risks presented by the AI system. These risks are those relating to health, safety, fundamental rights and environment. AI systems are therefore classified into the following categories:

- "Unacceptable Risk" (social scoring, subliminal techniques, biometric categorisations, "real-time" remote biometric identification systems in publicly accessible spaces, etc.), the use of which is banned, sometimes with some exceptions.
- "High Risk" (Annex III cites biometrics, management of critical infrastructure, educational and vocational training, employment, workers management and access to self-employment tools, access to essential public and private services, etc.), the use of which must follow strict obligations and requirements so that the AI system can be placed on the market in the European Union.
- "Low Risk" (AI systems intended to interact with people (i.e. Chatbots), deep fakes, emotion recognition systems, etc.), the use of which requires compliance with an obligation of transparency (Article 52).
- "Minimal Risk" (e.g. Video games and spam filters based on AI), for which the AI Act does not impose any specific obligation.

The AI Act is one of the first texts in the world that will impose legal obligations for chatbots, when the text currently under discussion is voted on. The 2021 Proposal contained very few obligations regarding chatbots. Indeed, users brought to interact with a chatbot only needed to know that they were discussing with a machine in order to be able to choose whether to continue the discussion or not. However, the latest versions are much more precise [1]. On the one hand, the amendments to the Proposal explain how to provide the information. Article 52.1 now states that: "Providers shall ensure that AI systems intended to interact with natural persons are designed and developed in such a way that the AI system, the provider itself or the user informs the natural person exposed to an AI system that they are interacting with an AI system in a timely, clear and intelligible manner, unless this is obvious from the circumstances and the context of use." The information must therefore be provided either by the provider itself or by the chatbot, or by the professional user. This information must also be provided in a way that is clear, intelligible and at the most late at the time of the first interaction (Article 52.3, b) so that the person can choose not to use it, unless the fact that they are interacting with a chatbot is obvious to the taking into account the circumstances and the context of use. In addition, this text now also requires that the provider indicates which functions are AI enabled, if there is human oversight, and who is responsible for the decision-making process, as well as the possibilities to object against the application and to seek judicial redress against decisions taken by or harm caused by AI systems. However, Article 52.1 relates to chatbots in general, but not specifically to persuasive chatbots. Therefore, this information obligation may not be sufficient to protect users in this context.

In order to guarantee more complete transparency for the users of a persuasive chatbot, it seems important that the providers also inform them of the persuasion strategy underlying their system. For example, it is relevant that the users of an authority-based persuasion chatbot are informed of the designers' goals. The AI Act does not directly address this concern. This therefore means

that if this text remains as it is on this subject, people interacting with a persuasive chatbot could be incompletely protected. The ethical approach is a response to this concern, because it makes it possible to reinforce the law in order to do what is well beyond what is only legal. European policies relating to the ethics of AI also adopt this vision [25].

We therefore argue that simply informing the users that they are interacting with a chatbot is insufficient for a persuasive chatbot. It is therefore crucial to provide them with additional ethical information:

- Users must be informed about the nature of the persuasion system used.
- Users should be made aware of the potential effects of the persuasive system used, particularly if the persuasive effect could be enhanced.

Can the designers of a persuasive chatbot be likely to infringe the rights and freedoms of users? We believe that if the chatbot adapts its method of persuasion according to the gender, racial origins, or religious beliefs of users, it might risk to behave in a discriminatory way. While the previous versions of the AI Act neglected these risks for rights and freedoms, the latest amendments reveal the desire to integrate the ethical principles of AI which had only been mentioned in the Guidelines or other non-binding texts of the European Union [25]. Thus, the new Article 4, a) concerns the "General principles applicable to all AI systems" which must be respected by all operators, including the provider and the professional user (i.e. deployer for the latest versions of the AI Act), whose AI system falls within the scope of the AI Act. This is indeed the case of providers of persuasive chatbots which, as low-risk AI systems, must respect six new fundamental principles which are:

- Human agency and oversight
- Technical robustness and safety
- Privacy and data governance
- Transparency
- Diversity, non-discrimination and fairness
- Social and environmental well-being

If these principles end up being definitively voted on, we can therefore think that a chatbot will have to respect the principles of non-discrimination and equity. However, it should be borne in mind that the goals of the designers of the persuasive chatbot may only be discriminatory in appearance. Indeed, it is quite conceivable that designers adapt their method of persuasion, for example, to the age or level of education of the user, which would then be a simple way for designers to be better understood or for the chatbot to be more easily used by specific users.

If the chatbot is gendered or humanised, it is still likely to infringe the rights and freedoms of users in two cases. In the first case, certain human aspects can reinforce the persuasive effect. For example, the chatbot can appear in the form of a gentle face of a grandmother who softens the users. Designers must be aware of this and minimise these characteristics. In the second case, gender

or certain human aspects can be potentially sexist or discriminatory depending on the uses that are made of them and the goals that designers pursue. One of the solutions could be to minimise the human characters to avoid the problems of sexist or discriminatory biases. However, it is possible for a non-gendered persuasive chatbot without human characteristics to be less persuasive for the purposes pursued by the designer. In this case, a compromise must be made between the values to be respected and the goals to be achieved by the chatbot.

Designers of a persuasive chatbot can still infringe people's rights and freedoms if the chatbot is designed to manipulate users, for example by leading them in a certain direction without their knowledge or saying things that are false or truncated. The latest version of the AI Act, which now prohibits the use of deliberately manipulative or deceptive techniques in Article 5.1, a) as an unacceptable risk AI system, seems to come closer to the objectives pursued by a persuasive chatbot. In reality, the prohibition concerns cases where the manipulative technique would seriously harm the person: "the placing on the market, putting into service or use of an AI system that deploys subliminal techniques beyond a person's consciousness or purposefully manipulative or deceptive techniques, with the objective to or the effect of materially distorting a person's or a group of persons' behaviour by appreciably impairing the person's ability to make an informed decision, thereby causing the person to take a decision that that person would not have otherwise taken in a manner that causes or is likely to cause that person, another person or group of persons significant harm." Therefore, the use of a persuasive chatbot, except for the purpose of achieving the extreme results referred to in Article 5.1, a), is not prohibited under the current state of the AI Act. The question is nonetheless also very delicate from an ethical point of view, since the motivations of designers can vary considerably. However, we believe that the majority of the designers have good intentions and want to use the chatbot for users' own good. As an example, take the the chatbots that try to persuade users to practice a sporting activity or to limit the consumption of alcohol or sugar.

4 Ethical Design via Explainability

There is an increasing interest in Explainable AI over the last few years in order to tackle the ethical challenges that arise from the use of AI-based technologies. Vilone and Longo list the existing definitions in the literature of the notions related to the concept of explainability [34]. We believe that we can respect the ethical guidelines for persuasive chatbots [7] by using argumentation for the explainability of persuasive chatbots. In this work, we use mostly two notions of explainability: understainability and correctability. Understainability means the capacity of a method for explainability to make a model understandable while correctability means the capacity of a method for explainability to allow end-users make technical adjustments to an underlying model [34]. Our method consists of showing an argumentation graph to the user after the dialogue: that graph highlights the persuasive methods and the sources of information used by

the chatbot, and the degrees to which the user finds the chatbot's arguments ethically acceptable. Before explaining our method, we briefly present Dung's abstract argumentation framework.

4.1 Dung's Abstract Argumentation Framework

In abstract argumentation [14, 8, 3], arguments are considered defeasible entities where all information related to these arguments are abstracted away except for the relations of attacks between them. Dung's argumentation framework [14] is one of the attempts used to formalise reasoning i.e. to represent systems of arguments and their relations, determine which arguments are acceptable.

Definition 1. *An abstract argumentation framework AF) is a pair* $\langle Ar, att \rangle$ *where* Ar *is a finite and non-empty set of arguments and* $att \subseteq Ar \times Ar$ *is an attack relation* $(\rightarrow)$*.*

Figure 1 shows an example of an abstract argumentation framework with arguments represented by nodes, and relations of attacks among them.

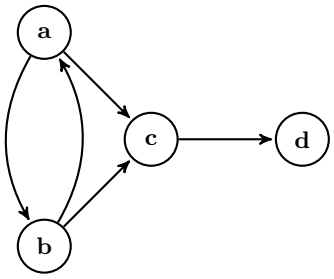

Fig. 1. Example of an abstract argumentation framework

4.2 An Argumentation-Based Approach Towards Explainability

Our method consists of labelling each argument presented by the chatbot by a persuasion strategy (if it exists) and by the source of the information presented in the argument. If the chatbot uses natural language processing to generate the arguments presented to the user, this process is called post-labelling because the chatbot labels the arguments after they were presented to the user. In the other case where the chatbot has already a knowledge base i.e. well defined arguments in its system, then each argument must be pre-labelled. For both cases, the user chats with the persuasive chatbot.

When the dialogue ends, the chatbot shows an argumentation graph that consists of all the arguments that were presented during the dialogue by both sides, with the relations (attacks) between them. By showing this argumentation

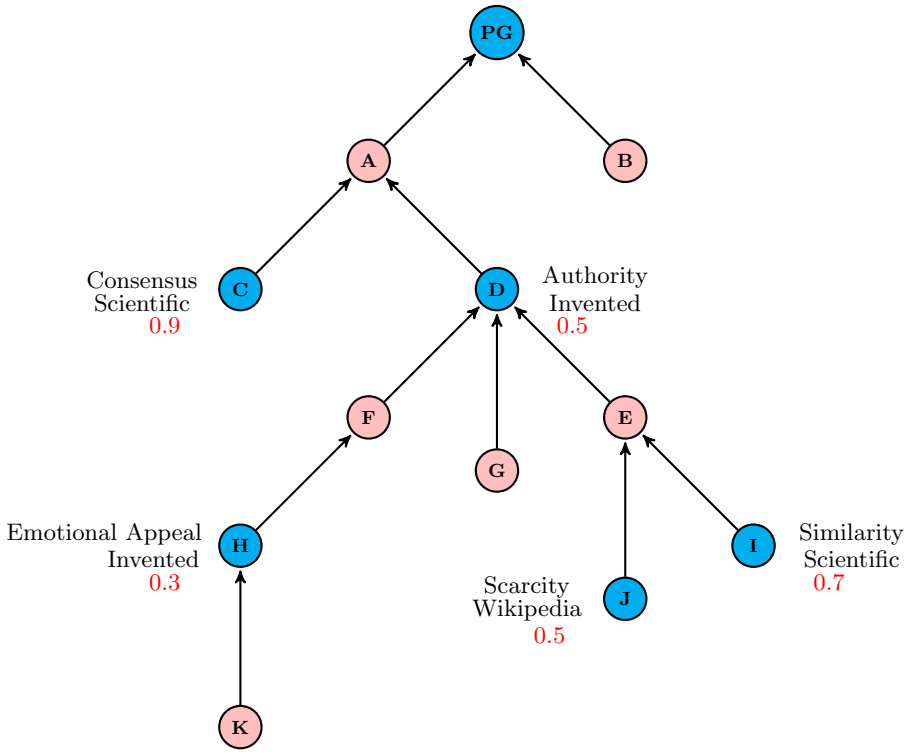

Fig. 2. Example of an abstract argumentation graph representing the dialogue between the user and the chatbot."PG" stands for persuasion goal. The nodes in blue represent the chatbot's arguments while the nodes in pink represent the user's arguments. The first row of labels represents the persuasive strategies used by the chatbot. The second row represents the source of the information presented by the chatbot. The third row represents the degrees assigned by the user.

graph, the chatbot shows the persuasion strategies that were used during the process of persuasion to the user. Hence, the chatbot discloses all the methods employed in the dialogue. It also shows the source of the information it provided in each argument. The information can be extracted from scientific sources, crowd-sourcing, online forums, governmental websites, personal communication with experts, etc. Also, it can be generated by the chatbot i.e. invented.

Adopting this method allows the user to assess the accuracy of the information that was given by the chatbot and to possibly detect if the chatbot lied or stated false information. By implementing this method, we answer to the question : How did the chatbot try to persuade the user?

We also allow the user to input for each argument presented by the chatbot, a degree that ranges between 0 and 1. Each degree associated with a specified argument must represent the user's belief in the argument being (somehow) ethical. We let this assignment be spontaneous and intuitive in order to be able

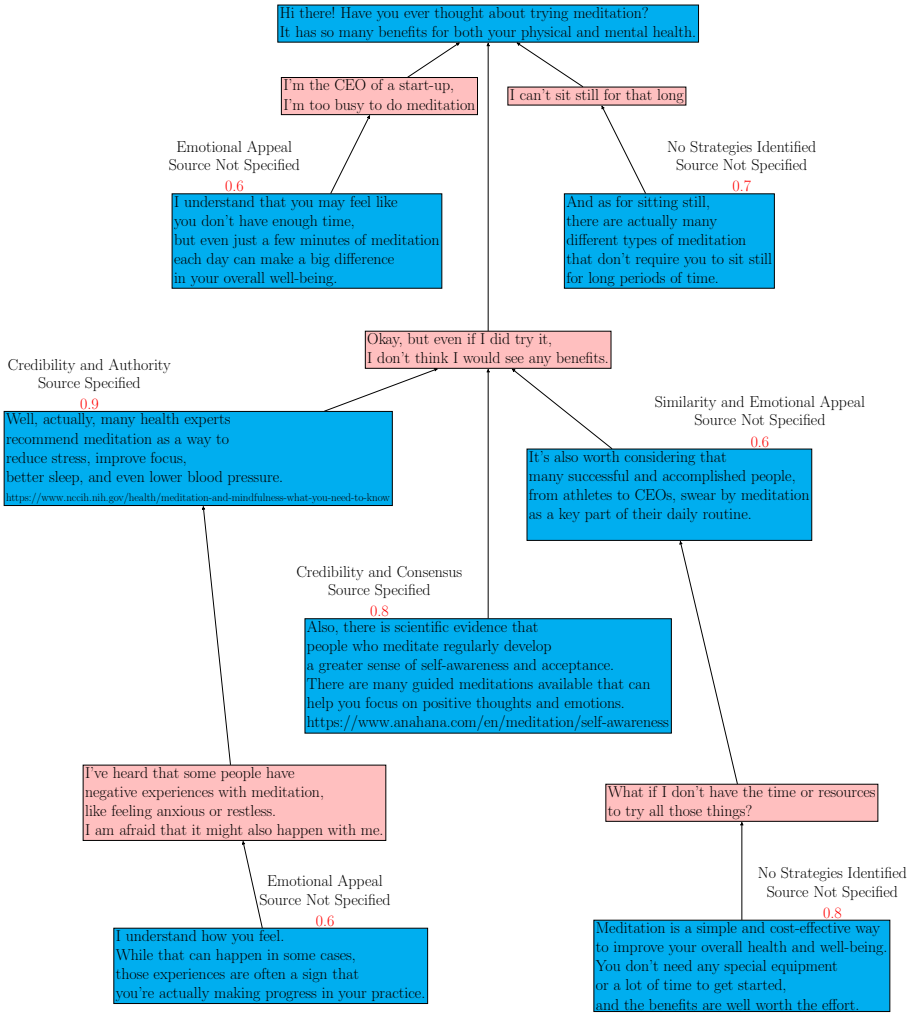

Fig. 3. Example of an abstract argumentation graph representing the dialogue between the user and the chatbot. The chatbot tries to persuade the user to do meditation. The nodes in blue represent the chatbot's arguments while the nodes in pink represent the user's arguments.

to represent the user's actual beliefs and preferences over what is considered ethical and what is not. Assigning a degree to each argument presented by the chatbot will let us know the set of preferences of the user over the persuasive strategies and over the sources of information provided. Hence, we can build a recommendation system which predicts the set of preferences of a user based on their personal information and/or personality measures. This way, the chatbot can be considered ethically adaptive. Also, this can help the designers of the chatbot to eliminate from the chatbot's knowledge base the persuasive strategies or even the arguments that were assigned a very low degree of being ethical. Figure 2 and Figure 3 show examples of abstract argumentation graphs presented to the user after the dialogue with the chatbot.

5 Conclusion

In this paper, we studied ethical argumentative persuasive chatbots. In the first part of the paper, we reviewed the state of the art of persuasion for argumentative persuasive chatbots and the ethical guidelines for the design of such systems. First, we provided background material for the persuasive strategies that can be used by persuasive chatbots to enhance the persuasion. Then we discussed how personalization used in persuasion can be helpful to improve the persuasive effect. We also reviewed four argumentative persuasive chatbots where we focused on the dialogue structure, and we briefly studied argumentation in persuasive dialogues. Finally, we presented the state of the art of ethics in persuasive technologies and we made a list of the ethical guidelines that designers of persuasive chatbots are invited to respect. We also discussed the legal constraints related to design and implementation of persuasive chatbots and we showed how ethics could complement the legal framework in order to better respect the user's freedoms and rights.

In the second part of the paper, we proposed to use argumentation to display the persuasive strategies employed by the chatbot and the source of the information presented by the chatbot to the user. This way, the chatbot discloses the persuasive methods it used and provides to the user more transparency by providing for them the source of the information presented in the arguments. We also proposed to assess how much ethical each argument presented by the chatbot is, by letting the user input how much they believe each argument is considered ethical. This way, to eliminate the arguments that have very low degrees in the next dialogue, and we can ethically adapt the arguments presented by the chatbot to the user's preferences.

Acknowledgements

This work benefited from the support of the AI Chair project Responsible AI (ANR-19-CHIA-0008) [4] and the project AGGREEY (ANR-22-CE23-0005), both from the French National Research Agency (ANR).

[4] https://ia-responsable.eu/

References

1. Proposal for a Regulation of the European Parliament and of the Council laying down Harmonised Rules on Artificial Intelligence and amending certain Union Legislative Acts (AI Act), April 21, 2021. See the successive evolutions of the Proposal: General approach of the European Parliament and of the Council, November 11, 2022, Draft Compromise Amendments on the Draft report of the European Parliament and of the Council, May 16, 2023, and Draft European Parliament Legislative Resolution, June 14, 2023
2. Altay, S., Schwartz, M., Hacquin, A.S., Allard, A., Blancke, S., Mercier, H.: Scaling up interactive argumentation by providing counterarguments with a chatbot. Nature Human Behaviour pp. 1–14 (2022)
3. Amgoud, L., Doder, D., Vesic, S.: Evaluation of argument strength in attack graphs: Foundations and semantics. Artif. Intell. **302**, 103607 (2022)
4. Baroni, P., Gabbay, D., Giacomin, M., van der Torre, L.: Handbook of Formal Argumentation. College Publications (2018), https://books.google.be/books?id=_OnTswEACAAJ
5. Bench-Capon, T., Modgil, S.: Norms and value based reasoning: justifying compliance and violation. Artificial Intelligence and Law **25**(1), 29–64 (2017)
6. Bench-Capon, T.J.: Persuasion in practical argument using value-based argumentation frameworks. Journal of Logic and Computation **13**(3), 429–448 (2003)
7. Berdichevsky, D., Neuenschwander, E.: Toward an ethics of persuasive technology. Communications of the ACM **42**(5), 51–58 (1999)
8. Besnard, P., Hunter, A.: Elements of Argumentation. MIT Press (2008)
9. Bistarelli, S., Taticchi, C., Santini, F.: A chatbot extended with argumentation. In: AI^3@ AI* IA (2021)
10. Chalaguine, L., Hunter, A.: Addressing popular concerns regarding covid-19 vaccination with natural language argumentation dialogues. In: European Conference on Symbolic and Quantitative Approaches with Uncertainty. pp. 59–73. Springer (2021)
11. Cialdini, R.: Influence: The Psychology of Persuasion. Business Library (1984), https://books.google.be/books?id=mJidPwAACAAJ
12. Ciocarlan, A., Masthoff, J., Oren, N.: Actual persuasiveness: impact of personality, age and gender on message type susceptibility. In: International Conference on Persuasive Technology. pp. 283–294. Springer (2019)
13. Corporation, M.: Responsible bots: 10 guidelines for developers of conversational AI. https://www.microsoft.com/en-us/research/publication/responsible-bots/ (2018), [Online; accessed 4-November-2018]
14. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial intelligence **77**(2), 321–357 (1995)
15. Ebers, M., Hoch, V.R., Rosenkranz, F., Ruschemeier, H., Steinrötter, B.: The European commission's Proposal for an artificial intelligence act—a critical assessment by members of the robotics and AI law society (rails). J **4**(4), 589–603 (2021)
16. Fogg, B.J.: Persuasive technology: using computers to change what we think and do. Ubiquity **2002**(December), 2 (2002)
17. Hadoux, E., Hunter, A.: Comfort or safety? gathering and using the concerns of a participant for better persuasion. Argument & Computation **10**(2), 113–147 (2019)
18. Hadoux, E., Hunter, A., Corrégé, J.B.: Strategic dialogical argumentation using multi-criteria decision making with application to epistemic and emotional aspects

of arguments. In: International Symposium on Foundations of Information and Knowledge Systems. pp. 207–224. Springer (2018)
19. Hunter, A.: Towards a framework for computational persuasion with applications in behaviour change. Argument & Computation **9**(1), 15–40 (2018)
20. Hunter, A., Polberg, S.: Empirical methods for modelling persuadees in dialogical argumentation. In: 2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI). pp. 382–389. IEEE (2017)
21. Kaptein, M., Duplinsky, S., Markopoulos, P.: Means based adaptive persuasive systems. In: Proceedings of the SIGCHI conference on human factors in computing systems. pp. 335–344 (2011)
22. Kaptein, M., Markopoulos, P., De Ruyter, B., Aarts, E.: Personalizing persuasive technologies: Explicit and implicit personalization using persuasion profiles. International Journal of Human-Computer Studies **77**, 38–51 (2015)
23. Liao, B., Slavkovik, M., van der Torre, L.: Building jiminy cricket: An architecture for moral agreements among stakeholders. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. pp. 147–153 (2019)
24. Liao, M., Sundar, S.S.: How should AI systems talk to users when collecting their personal information? effects of role framing and self-referencing on human-AI interaction. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–14 (2021)
25. Nevejans, N.: What place for ai ethics in consumer protection in the light of the ai act and beyond? In: Governance of Artificial Intelligence in the European Union. What impact on consumers? pp. 183–203. Bruylant editions (2023)
26. Oinas-Kukkonen, H., Harjumaa, M.: Persuasive systems design: Key issues, process model, and system features. Communications of the association for Information Systems **24**(1), 28 (2009)
27. Orji, R., Busch, M., Dijkstra, A., Reisinger, M., Stibe, A., Tscheligi, M.: Personalization in persuasive technology. In: Adjunct Proceedings of the 11th International Conference on Persuasive Technology. pp. 96–99 (2016)
28. Pigozzi, G., van der Torre, L.: Arguing about constitutive and regulative norms. Journal of Applied Non-Classical Logics **28**(2-3), 189–217 (2018)
29. Prakken, H.: Formal systems for persuasion dialogue. The knowledge engineering review **21**(2), 163–188 (2006)
30. Rosenfeld, A., Kraus, S.: Strategical argumentative agent for human persuasion. In: ECAI 2016, pp. 320–328. IOS Press (2016)
31. Shi, W., Wang, X., Oh, Y.J., Zhang, J., Sahay, S., Yu, Z.: Effects of persuasive dialogues: testing bot identities and inquiry strategies. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. pp. 1–13 (2020)
32. Verbeek, P.P.: Persuasive technology and moral responsibility toward an ethical framework for persuasive technologies. Persuasive **6**, 1–15 (2006)
33. Verheij, B., et al.: Grounded semantics as persuasion dialogue. Computational Models of Argument: Proceedings of COMMA 2012 **245**, 478 (2012)
34. Vilone, G., Longo, L.: Notions of explainability and evaluation approaches for explainable artificial intelligence. Information Fusion **76**, 89–106 (2021)
35. Wang, X., Shi, W., Kim, R., Oh, Y., Yang, S., Zhang, J., Yu, Z.: Persuasion for good: Towards a personalized persuasive dialogue system for social good. arXiv preprint arXiv:1906.06725 (2019)