Argument & Computation 8 (2017) 61–87
DOI 10.3233/AAC-170015
IOS Press

Emotions and personality traits in argumentation: An empirical evaluation[1]

Serena Villata [a,*], Elena Cabrio [a], Imène Jraidi [b], Sahbi Benlamine [b], Maher Chaouachi [b], Claude Frasson [b] and Fabien Gandon [c]
[a] *Université Côte d'Azur, CNRS, Inria, I3S, France*
[b] *University of Montreal, Canada*
[c] *Université Côte d'Azur, Inria, CNRS, I3S, France*

Abstract. Argumentation is a mechanism to support different forms of reasoning such as decision making and persuasion and always cast under the light of critical thinking. In the latest years, several computational approaches to argumentation have been proposed to detect conflicting information, take the best decision with respect to the available knowledge, and update our own beliefs when new information arrives. The common point of all these approaches is that they assume a purely rational behavior of the involved actors, be them humans or artificial agents. However, this is not the case as humans are proved to behave differently, mixing rational and emotional attitudes to guide their actions. Some works have claimed that there exists a strong connection between the argumentation process and the emotions felt by people involved in such process. We advocate a complementary, descriptive and experimental method, based on the collection of emotional data about the way human reasoners handle emotions during debate interactions. Across different debates, people's argumentation in plain English is correlated with the emotions automatically detected from the participants, their engagement in the debate, and the mental workload required to debate. Results show several correlations among emotions, engagement and mental workload with respect to the argumentation elements. For instance, when two opposite opinions are conflicting, this is reflected in a negative way on the debaters' emotions. Beside their theoretical value for validating and inspiring computational argumentation theory, these results have applied value for developing artificial agents meant to argue with human users or to assist users in the management of debates.

Keywords: Emotions, engagement index, abstract bipolar argumentation

1. Introduction

Understanding how humans reason and take decisions in debates and discussions is a key issue in cognitive science and a challenge for social applications. Moreover, with the growing importance of the Web, this issue is complicated by the fact that in such a hybrid space heterogeneous actors, both human and artificial, interact. As a typical example, Wikipedia is managed by users and bots who constantly contribute, agree, disagree, debate and update the content of the encyclopedia. In this context, several dimensions of the interaction affect the reasoning and decision making process, i.e., the arguments that are proposed online, the emotions of those who propose such arguments as well as the emotions of those reading these arguments, the social relationships among the involved actors, the writing of their messages, etc. This underlines the need for multidisciplinary approaches and research for Web applications in general and for detecting and managing the emotional state of a user in particular to allow artificial

[1] This paper is the extended version of the paper titled "Emotions in Argumentation: an Empirical Evaluation" published in the Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI-2015).

[*] Corresponding author. E-mail: villata@i3s.unice.fr.

and human actors to adapt their reactions to others' emotional states. It is also a useful indicator for community managers, moderators and editors to help them in handling the communities and the content they produce.

In this paper, we argue that in order to apply argumentation to scenarios like e-democracy and online debate systems, designers must take both the argumentation and the emotions into account. In order to efficiently manage and interact with such a hybrid society, we need to improve our means to understand and link the different dimensions of the exchanges (e.g. social interactions, textual content of the messages, dialogical structures of the interactions, emotional states of the participants). Beyond the challenges individually raised by each dimension, a key problem is to link these dimensions and their analysis together with the aim to detect, for instance, a debate turning into a flame war, a content reaching an agreement, a good or bad emotion spreading in a community.

In this paper, we aim to answer the research question: *What is the connection between the argumentation and the emotions in online debate interactions?* Such question breaks down into sub-questions:

- How are the arguments and their relations correlated with the polarity of the detected facial emotions?
- Is the relation between the kind and the amount of arguments proposed in a debate correlated with the mental engagement and workload detected for each participant in the debate?
- How do personality traits and opinions affect participants' emotions during the debate?

To answer these questions, we propose an empirical evaluation of the connection between argumentation, personality traits, and emotions in online debate interactions. This paper describes an experiment with human participants which investigates the correspondences between the arguments and their relations put forward during a debate, the emotions detected by emotions recognition systems in the debaters, and the personality traits of the debaters. We designed an experiment where 12 debates were addressed by 4 participants each. Participants were asked to discuss about 12 topics in total proposed by moderators, e.g., "Religion does more harm than good" and "Cannabis should be legalized". Participants argue in plain English proposing arguments, that are in positive or negative relation with the arguments proposed by the other participants and by moderators. During these debates, participants are equipped with emotion detection devices, recording their emotions. Moreover, each participant filled in a questionnaire for Big Five personality traits [26]. We hypothesize that mental engagement and emotions are correlated to the argumentation holding in the debates, namely to the number of arguments that are proposed, and the kind of relations connecting them (i.e., support or attack). Moreover, we hypothesize that personality traits of debaters and debaters' opinions regarding the discussed topics modulate their emotional experiences during the debates.

A key point in our work is that, up to our knowledge, no user experiment has been carried out yet to determine what is the connection between the argumentation addressed during a debate, the emotions emerging in the participants, as well as their personality traits. An important result of the work reported here is the development of a publicly available dataset, capturing several debate situations, annotated with their argumentation structure and the emotional states automatically detected.

It is worth highlighting that bipolar abstract argumentation is used in our experimental setting to pair the arguments, connect them with the appropriate relation (either support or attack), and combine them in bipolar argumentation graphs. This structure allows for further reasoning activities over the data, where for instance the acceptability of the arguments depends on the mental engagement associated to their conception by their proposer, or a ranking is established over the acceptable arguments depending on the emotions (mostly positive, mostly negative, neutral) they generated in the audience. The definition

and evaluation of these reasoning processes are not in the scope of the present paper, and we left them for future research. It is worth clarifying also that, in this paper, we are not interested in verifying explanation and reasoning theories proposed in cognitive psychology like, among others, those of Lombrozo and colleagues [48] about explanations in category learning, and Keil and colleagues [27] about explanatory reasoning through an abductive theory. We rely on bipolar argumentation for representing the debates, and to foster the application of reasoning techniques.

The paper is organized as follows. In Section 2 we describe the main insights of the two components of our framework, namely bipolar argumentation theory and emotion detection systems, then in Section 3, we describe the experimental protocol and the questionnaires we proposed to the debaters during the experiment, and our research hypotheses. In Section 4, we provide a detailed analysis of the experimental results, and we compare this work with the relevant literature in Section 5. Conclusions end the paper, and Appendix 1 presents the Big Five personality traits questionnaire.

2. The framework

In this section, we present the two main components involved in our experimental framework: *i)* computational argumentation theory, and more precisely *bipolar argumentation theory*, i.e., the formalism used to analyze and represent the argumentation elements from the debates, and *ii)* the methodologies and tools used to detect the degrees of attention, engagement, and workload of each participant involved in the debate, as well as her facial emotions.

2.1. Argumentation theory

Argumentation is a very fertile research area in Artificial Intelligence, defined as the process of creating arguments for and against competing claims [39]. What distinguishes argumentation-based discussions from other approaches is that opinions have to be supported by the arguments that justify, or oppose, them. This permits a greater flexibility than in other decision making and communication schemes since, for instance, it makes it possible to persuade other persons to change their view of a claim by identifying information or knowledge that is not being considered, or by introducing a new relevant factor in the middle of a negotiation, or to resolve an impasse.

Computational argumentation is the process by which arguments are constructed and handled. Thus argumentation means that arguments are compared, evaluated in some respect and judged in order to establish whether any of them are warranted. Roughly, each argument can be defined as a set of assumptions that, together with a conclusion, is obtained by a reasoning process. Argumentation as an exchange of pieces of information and reasoning about them involves groups of actors, human or artificial. We can assume that each argument has a proponent, the person who puts forward the argument, and an audience, the persons who receive the argument.

A highly influential framework for studying argumentation-based reasoning was introduced by [14]. A Dung-like argumentation framework (AF) is defined as a pair $\langle A, att \rangle$ where A is a set of elements called *arguments* and *att* is a defeat relation between arguments. This approach is called *abstract* as it focuses on the defeat relation between arguments, leaving aside their origin or their internal structure. In this abstract definition of argumentation, an argumentation framework can be represented as a directed graph in which vertices are arguments and directed arcs characterize defeat among arguments. In order to define a more expressive framework, able to cope with the complexity of real argumentation, the extension of Dung's abstract framework with a support relation has been advocated in [7], with the

introduction of so called *bipolar argumentation frameworks*. Bipolarity refers to the presence of two independent kinds of interactions between the arguments which have a diametrically opposed nature. Thus, a bipolar argumentation framework is defined as a pair $\langle A, att, sup \rangle$ where A is the set of arguments, *att* is the (negative) defeat (i.e., attack) relation between arguments, and *sup* is the (positive) support relation between arguments.

In order to analyze from the argumentation point of view the debates in which the participants to our experiment have been involved, we rely on abstract bipolar argumentation. In this way, we do not need distinguish the internal structure of the arguments (i.e., premises, conclusion), but we consider each argument proposed by the participants in the debate as a unique element, then analyzing the relation (positive or negative) it has with the other pieces of information put forward in the debate. The following example extracted from one of the debates addressed in our experiment shows how the arguments are connected to each other through a defeat or a support relation. Consider the following three arguments proposed by the participants of the debate about "Religion does more harm than good". We have that the issue of the debate is also our first argument whose proponent is the debate moderator, then the other two arguments are proposed by two different participants:

Argument 1: *Religion does more harm than good.*
Argument 2: *During all the existence of the human being, religion makes a lot of issues. It makes more hurts than cures.*
Argument 3: *I think for people, religion is a refuge against the horrors of the world.*

Given such a kind of debate, we have that three arguments are proposed (namely Argument 1, Argument 2 and Argument 3) whose relations with each other are as follows: (Argument 2) supports (Argument 1), (Argument 3) attacks (Argument 1), and (Argument 3) attacks (Argument 2).

Note that in this paper we are not interested in applying natural language processing techniques to detect automatically the relations among the arguments. On the contrary, we have manually built our data set of argumentation and emotions from the data retrieved during the debates of our experiment. Experiments with natural language processing approaches will be part of future work.

2.2. Emotion detection

Human emotion analysis during traditional face-to-face or computer-mediated interaction has always been a challenging and attractive task mainly because of how emotions are closely associated to human behavior and experience. Several theories state that emotions serve as an adaptive function to our behavior, e.g., [19,25,31,42]. Following these theories, the appraisal of an experience and the intention to act to maintain, adjust or change a condition related to this experience is impacted by emotions. During a debate, emotional reactions provoked by others' arguments could be, for example, a trigger for developing attacking or supporting arguments.

Emotion recognition methods can be categorized into three groups, each of them defining one level of how a usual emotional response is displayed, namely, *experiential*, *behavioral* and *physiological* [22]. For example if an individual is annoyed by someone else's argument, the subjective experience could be the anger; the behavioral response will be displayed through a higher voice tone (during a face-to-face conversation) or an angry facial expression; and the physiological response will be activated by an increasing level of heart rate. Usually, the experiential methods use subjective self-report instruments (such as surveys and questionnaires) to determine the emotion relative to a specific event. The behavioral methods are based on external observable clues detected from the individual's behavior (such as

gestures, body movements, facial expression, voice tone and pitch, etc.) that can indicate the type of the emotion. The physiological methods rely on physical sensors (such as EDA, EEG, heart rate, respiration rate, temperature, etc.) to measure specific physiological shifts and patterns that can be related to specific emotions. In a computer-mediated context, the use of self-report surveys to recognize the individuals' emotions could be inconvenient. If these surveys are administrated at frequent intervals during the task, it could be disrupting for the task performance. However, non-frequent administration intervals of the survey could result into an undetailed and ambiguous assessment of the different emotions experienced during the task. Therefore, growing research interest has arisen towards using and combining behavioral and physiological methods for emotion recognition [28,29,37]. These methods allow automatic, objective and reasonably precise emotional recognition level.

In our study, we selected a behavioral method to extract the emotional manifestations. We used a set of webcams (one for each participant in the discussion) whose recordings have been analyzed with the FaceReader software[2] to detect a set of discrete emotions from facial expressions. Furthermore, we also recoded the EEG data from each participant in order to extract more complex information about their internal cognitive state. This cognitive information was aligned and analyzed jointly with the emotional information to have a global overview of the debate experience for each participant.

Detecting emotions from facial expressions. The emotional detection from facial expression is one of the most commonly and predominantly used methods [1,24,47]. In fact, facial expressions of basic emotions are widely believed to be naturally and universally expressed and recognized. In this study, we used the FaceReader software (version 6.0) to automatically extract the emotional reactions from the frame-by-frame videos recoded by the webcam. The FaceReader software launched by Noldus Information technology is able to recognize six basic emotions, namely, *happy*, *sad*, *angry*, *surprised*, *scared* and *disgusted* with an accuracy reaching 87%. The detection process is performed by extracting and classifying in real-time 500 key points in facial muscles of the target face. These key points are provided as input to a neural network trained on a dataset of 10000 manually annotated images corresponding to these six basic emotions. In addition to the probability of the presence of these six emotions, the software output also contains the probability of the neutral state as well as the *valence*, and the *arousal* of the emotional state. Information about the emotional valance defines the nature of the emotion and is ranging from -1 to $+1$. A positive valence value refers to pleasant emotion, whereas a negative valence value characterizes unpleasant emotions. The information about the arousal of the emotion defines its intensity and is also ranging from -1 to $+1$. A high arousal value indicates a high emotional intensity and a low value the opposite.

In this study, at each second in the debate, a dominant emotion is computed for each one of the four participants. This dominant emotion corresponds to the emotion (among the six detected by FaceReader) with the highest probability. Moreover, information about the valence and arousal of this emotion as well as their class (pleasant or unpleasant for the valence, high or low for the arousal) was also considered.

Emotiv EPOC EEG headset. In order to record physiological data of the participants during the debate sessions, we used the 4 Emotiv Epoc EEG headsets (one for each participant). This device contains 14 electrodes spatially organized according to International 10–20 system.[3] The pads of each electrode

[2] http://www.noldus.com/human-behavior-research/products/facereader

[3] International 10–20 system is an internationally recognized method to describe and apply the location of scalp electrodes in the context of an EEG test or experiment.

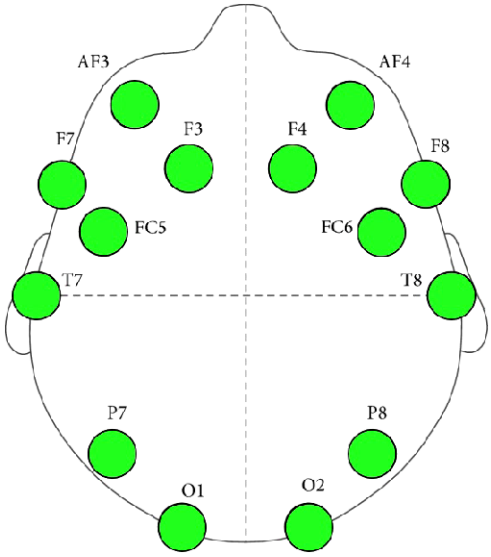

Fig. 1. Emotiv Headset sensors/data channels placement.

were moistened with a saline solution (contact lens cleaning solution) in order to enhance the quality of the signal. Figure 1 depicts the recorded sites, namely: AF3, AF4, F3, F4, F7, F8, FC5, FC6, P7, P8, T7, T8, O1, and O2. The reference of this EEG setup is represented by two other electrodes located behind the user's ears. The generated data are in (μV) with a 128 Hz sampling rate. The signal frequencies are between 0.2 and 60 Hz. An artifact rejection procedure based on the signal amplitude was performed in order to reduce the impact of blinking and body movement effect [17,18]. More precisely, if the amplitude of any 1-s EEG in any site exceeds in 25% of the data point a predefined threshold, the segment is rejected.

Extracting the engagement index. The term mental engagement refers to the level of attention and alertness during a task. The engagement index used in our study is based on the findings of [36] and [18]. In their study, it was found that the user's performance improved when an EEG index is used as a criterion for switching between manual and automated piloting mode. This index is computed from three EEG frequency bands: α(8–12 Hz), β(12–22 Hz) and θ(4–8 Hz):

$$\text{eng} = \frac{\theta}{\alpha + \beta}.$$

This index is computed each second from the EEG signal. To smooth the values of this index and reduce its fluctuation, we used a moving average on a 40-second mobile window. More precisely, the value of the index at time t corresponds to the total average of the ratios calculated on a period of 40 seconds preceding t. The extraction of the frequency bands (namely α, β and θ) was performed by multiplying every second of the EEG signal by a Hamming window (to reduce the spectral leakage) and applying a Fast Fourier Transform (FFT). As the Emotiv headset measures 14 regions at the same time, we used a combined value of α, β and θ frequency bands by summing their values over all the measured regions. To examine participants' engagement, we extract their minimum, average and maximum values during the debate, and we use such values to identify the range of engagement (High, Medium, Low) of every participant. Since its development by Pope and his colleagues, this engagement index has become

a very important and popular technique for real time or offline tracking and analysis of individuals' engagement in several laboratory studies. In the educational sittings for example, this engagement index was used for monitoring learners' engagement and adapting learning activities according to their level of mental engagement by [44] and [11]. In robotics, this index was also used to leverage the interaction between a robot and a user by providing the robot real-time information about the user's engagement while the robot is speaking to him by [43]. The robot was successfully able to detect when the user is not anymore engaged in listening to him and tried to regain his attention by employing verbal and nonverbal techniques. This engagement index was also selected as a criterion for adapting a game's difficulty according to the player's level of engagement, showed promising results [9].

Extracting the workload index. The term workload (or cognitive load in a learning context) refers to a measurable quantity of information processing demands placed on an individual by a task [35]. The mental workload is generally related to the working memory and could be viewed as a mental effort produced to process the quantity of information involved in the task.

Unlike the engagement index which is directly extracted from the EEG data, the EEG workload index was based on pre-trained predictive model [10]. This model was trained using a set of EEG data collected from a training phase during which a group of seventeen participants performed a set of brain training exercises. This training phase involved three different types of cognitive exercises, namely: digit span, reverse digit span and mental computation. The objective of these training exercises was to induce different levels of mental workload while collecting the learner's EEG data. The manipulation of the induced workload level was done by varying the difficulty level of the exercises: by increasing the number of the digits in the sequence to be recalled for digit span and reverse digit span, and the number of digits to be added or subtracted for the mental computation exercises – we refer to [10] for more details on the procedure. After performing each difficulty level, the participants were asked to report their workload level using the subjective scale of NASA Task load index (NASA_TLX) [23]. Once this training phase was completed, the collected EEG raw data were cut into 1-second segments and multiplied by a Hamming window. A FFT was applied to transform each EEG segment into a spectral frequency and generate a set of 40 bins of 1 Hz ranging from 4 to 43 Hz (EEG pre-treated vectors). The dimensionality of the data was then reduced using a Principal Component Analysis (PCA) to 25 components (the score vectors). Next, a Gaussian Process Regression (GPR) algorithm with an exponential squared kernel and a Gaussian noise [40] was run in order to train a mental workload predictive model (the EEG workload index) from the normalized score vectors. Normalization was done by subtracting the mean and dividing by the standard deviation of all vectors. In order to reduce the training time of the predictive model, a faster version of GPR the local Gaussian Process Regression algorithm was used [15]. The evaluation of this model showed a correlation with the participants' subjective scores NASA_TLX reaching 82% (mean correlation 72%). This same approach was used within an intelligent tutoring system called (MENTOR) fully controlled by this workload index to automatically select the most adapted learning activity for the learner. The experimental results showed positive impact of using such index on learners' performance and satisfaction.

The reader may question about the reliability of such kind of neural metrics. Actually, many contributions have tackled the issue of predicting human behavior from neural metrics, e.g., [3,16,30], by collecting EEG data to detect cognitive interest, emotional engagement and decision making of consumers towards communication messages or advertisements in order to optimize them.

3. Experimental setting

This section details the experimental session we set up to analyze the relation between the emotions and the argumentation process. More precisely, we detail the protocol we have defined to guide the experimental setting, and the resulting datasets we have manually annotated in order to combine the arguments proposed in the debates with the detected emotions. Finally, we specify the hypotheses we aim at verifying in this experiment.

3.1. Protocol

The general goal of the experimental session is to investigate the relation (if any) holding between the emotions detected in the participants during a debate session and the argumentation flow of the debate itself. The idea is to associate the arguments and the relations among them to the participants' mental engagement and workload detected via the EEG headset, and the facial emotions identified via the Face Emotion Recognition tool.

More precisely, starting from an issue to be discussed provided by the moderators, e.g., *We have to ban animal testing*, the aim of the experiment is to collect the arguments proposed by the participants on the topic, as well as the relations among them (i.e., support or attack), and to associate such arguments/relations to the mental engagement and workload states and to the facial emotions expressed by the participants. During a post-processing phase on the collected data, we synchronize the arguments put forward by the different participants at time t with the emotional indexes we retrieved. Finally, we build the resulting bipolar argumentation graph of each debate, such that the resulting argumentation graphs are labelled with the source who has proposed each argument, and the emotional state of each participant at the time of the introduction of the argument in the discussion.

The first point to clarify in this experimental setting is the terminology. In this experiment, an *argument* is each single piece of text that is proposed by the participants in the debate. Typically, arguments have the goal to promote the opinion of the debater in the debate. Thus, an *opinion* in our setting represents the overall opinion of the debater about the issue to be debated. The opinion is promoted in the debate through arguments, that will support (if the opinions converge) or attack (otherwise) the arguments proposed in the debate by the other participants.

The experiment involves two kinds of persons:

- *Participant*: she is expected to provide her own opinion about the issue of the debate proposed by the moderators, and to argue with the other participants in order to make them understand the goodness of her viewpoint (in case of initial disagreement).[4]
- *Moderator*: she is expected to propose the initial issue to be discussed to the participants. In case of lack of active exchanges among the participants, the moderator is in charge of proposing pro and con arguments (with respect to the main issue) to reactivate the discussion.

The experimental setting of each debate is conceived as follows: there are 4 participants involved in each discussion group, and 2 moderators. Each participant is placed far from the other participants, even if they are in the same room, while moderators are placed in another room. Moderators interact with the participants uniquely through the debate platform, and the same holds for the interactions among participants. The language used for debating is English.

[4]Note that, in this experimental scenario, we do not evaluate the connection between the emotions and persuasive argumentation. This analysis is out of the scope of this paper and it is left for future research.

In order to provide an easy-to-use debate platform to the participants, without requiring from them any background knowledge, we decide to rely on a simple IRC network[5] as debate platform. The debate is anonymous and participants are visible to each others with their nicknames, e.g., *participant1*, while the moderators are visualized as *moderator1* and *moderator2*. Each participant has been provided with 1 laptop device equipped with Internet access and a camera used to detect facial emotions. Moreover, each participant has been equipped with an EEG headset to detect the engagement and workload indexes. Moderators were equipped with a laptop only.

The procedure we follow for each debate is:

- Participants are firstly equipped with the EEG headset and the good connection of the headset is verified;
- Participants are familiarized with the debate platform;
- The debate starts – Participants take part into two debates each, about two different topics for a maximum of about 20 minutes for each debate:
  - The moderator(s) provides the debaters with the topic to be discussed;
  - The moderator(s) asks each participant to provide a general statement about his/her opinion on the topic;
  - Participants expose their opinion to the others;
  - Participants are asked to comment on the opinions expressed by the other participants;
  - If needed (no active debate among the participants), the moderator(s) posts an argument and asks for comments from the participants;

The variables measured in this experimental setting are the following: *(i)* engagement and workload indexes (measurement tool: EEG headset), and *(ii)* facial emotions, i.e., Neutral, Happy, Sad, Angry, Surprised, Scared and Disgusted (measurement tool: FaceReader).

The post-processing phase of the experimental setting involves the following steps:

- manual annotation of the support and attack relations holding between the arguments proposed in each discussion, following the methodology described in Section Dataset;
- manual annotation of the opinion of the participants at the beginning and at the end of the debates they participated in, and synchronization with the debriefing questionnaire data;
- synchronization of the argumentation (i.e., the arguments/relations proposed at time instant t) with the emotional indexes retrieved at time t using the EEG headset and FaceReader.

Participants. The experiment was distributed over 6 sessions of 4 participants each; the first session was discarded due to a technical problem while collecting data. We had a total of 20 participants (7 women, 13 men), whose age range was from 22 to 35 years. All of them were students in a North American university, and all of them had good computer skills. Since not all of them were native English speakers, the use of the Google translate service was allowed. They have all signed an ethical agreement before proceeding to the experiment.

Participants have been asked to complete a short questionnaire about their viewpoints on the discussed topics. Thus, after each debate session, a debriefing phase was addressed. The questionnaire contained the following questions:[6]

- What was your starting opinion about the discussed topic before entering into the debate?

[5] http://webchat.freenode.net/

[6] Such material is available at http://bit.ly/DebriefingData.

- What is your final opinion about the discussed topic after the debate?
- If you changed your mind, why (i.e., which was the argument(s) that has made you change your mind)?

These questions allowed us to *know* what is the opinion of the participants about the specific topics they debated about, without the need to infer it from the arguments they propose in the debates. The answers participants provided to these questions have been then used to correlate their opinions with the emotions they felt during the debates.

Finally, participants have been asked to fill in a questionnaire for Big Five personality traits. More precisely, participants filled in a questionnaire of 50 items of the kind:

- I get stressed out easily;
- I don't like to draw attention on myself;
- I spend time reflecting on things;
- ...;

where the possible values range over a typical five-level Likert scale: *Totally Disagree*, *Disagree*, *Neutral*, *Agree*, *Totally Agree*. The complete Big Five personality traits questionnaire is reported in Appendix. Such information allowed us to extract the *OCEAN* personality dimensions, i.e.:

O Openness, Originality, Open-mindedness
C Conscientiousness, Control, Constraint
E Extraversion, Energy, Enthusiasm
A Agreeableness, Altruism, Affection
N Neuroticism, Negative Affectivity, Nervousness

These dimensions have been analyzed with respect to their correlation with the detected emotions of participants during the debates. More details about this analysis are provided in the Results Section.

3.2. Dataset

In this section, we describe the dataset of textual arguments we have created from the debates among the participants. The dataset is composed of four main layers: *(i)* the basic annotation of the arguments proposed in each debate (i.e. the annotation in xml of the debate flow downloaded from the debate platform); *(ii)* the annotation of the relations of support and attack among the arguments; *(iii)* starting from the basic annotation of the arguments, the annotation of each argument with the emotions felt by each participant involved in the debate; and *(iv)* starting from the basic annotation, the opinion of each participant about the debated topic at the beginning, in the middle and at the end of debate is extracted and annotated with its polarity. In the reminder of this section, we describe the annotation process of the four layers and the resulting inter-annotator agreement to ensure the reliability of the produced linguistic resource.

The *basic* dataset is composed of 598 different arguments proposed by the participants in 12 different debates. The debated issues and the number of arguments for each debate are reported in Table 1. We selected the topics of the debates among the set of popular discussions addressed in online debate platforms like iDebate[7] and DebateGraph.[8]

[7] http://idebate.org/
[8] www.debategraph.org/

The annotation (in `xml`) of this dataset is as follows: we have assigned to each debate a unique numerical id, and for each argument proposed in the debate we assign an id and we annotate who was the participant putting this argument on the table, and in which time interval the argument has been proposed. An example of basic annotation is provided below:

```
<debate id="1" title="Ban_Animal_Testing">
<argument id="1" debate_id="1" participant="mod"
  time-from="19:26" time-to="19:27">Welcome to
  the first debate! The topic of the first debate
  is that animal testing should be banned.</argu-
  ment>

<argument id="3" debate_id="1" participant="2"
  time-from="20:06" time-to="20:06">If we don't
  use animals in these tests, what could we use?
  </argument>
</debate>
```

The second level of our dataset consists in the annotation of arguments pairs with the relation holding between them, i.e., support or attack. To create the dataset, for each debate of our experiment we apply the following procedure, validated in [4]:

(1) the main issue (i.e., the issue of the debate proposed by the moderator) is considered as the starting argument;
(2) each opinion is extracted and considered as an argument;
(3) since *attack* and *support* are binary relations, the arguments are coupled with:
   – the starting argument, or
   – other arguments in the same discussion to which the most recent argument refers (e.g., when an argument proposed by a certain user supports or attacks an argument previously expressed by another user);
(4) the resulting pairs of arguments are then tagged with the appropriate relation, i.e., *attack* or *support*.

To show a step-by-step application of the procedure, let us consider the debated issue *Ban Animal Testing*. At step 1, we consider the issue of the debate proposed by the moderator as the starting argument (a):

(a) *The topic of the first debate is that animal testing should be banned.*

Then, at step 2, we extract all the users opinions concerning this issue (both pro and con), e.g., (b), (c) and (d):

(b) *I don't think the animal testing should be banned, but researchers should reduce the pain to the animal.*

(c) *I totally agree with that.*

(d) *I think that using animals for different kind of experience is the only way to test the accuracy of the method or drugs. I cannot see any difference between using animals for this kind of purpose and eating their meat.*

(e) *Animals are not able to express the result of the medical treatment but humans can.*

At step 3a we couple the arguments (b) and (d) with the starting issue since they are directly linked with it, and at step 3b we couple argument (c) with argument (b), and argument (e) with argument (d) since they follow one another in the discussion. At step 4, the resulting pairs of arguments are then tagged by one annotator with the appropriate relation, i.e.: **(b)** *attacks* **(a)**, **(d)** *attacks* **(a)**, **(c)** *supports*

(b) and **(e)** *attacks* **(d)**. For the purpose of validating our hypotheses, we decided to not annotate the supports/attacks between arguments proposed by the same participant (e.g., situations where participants are contradicting themselves). Note that this does not mean that we assume that such situations do not arise: no restriction was imposed to the participants of the debates, so situations where a participant attacked/supported her own arguments are represented in our dataset. We just decided to not annotate such cases in the dataset of argument pairs, as it was not necessary for verifying our assumptions.

To assess the validity of the annotation task and the reliability of the obtained dataset, the same annotation task has been independently carried out also by a second annotator, so as to compute inter-annotator agreement. It has been calculated on a sample of 100 argument pairs (randomly extracted). The complete percentage agreement on the full sample amounts to 91%. The statistical measure usually used in NLP to calculate the inter-rater agreement for categorical items is Cohen's kappa coefficient [5], that is generally thought to be a more robust measure than simple percent agreement calculation since κ takes into account the agreement occurring by chance. More specifically, Cohen's kappa measures the agreement between two raters who each classifies N items into C mutually exclusive categories. The equation for κ is:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \tag{1}$$

where $\Pr(a)$ is the relative observed agreement among raters, and $\Pr(e)$ is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly saying each category. If the raters are in complete agreement then $\kappa = 1$. If there is no agreement among the raters other than what would be expected by chance (as defined by $\Pr(e)$), $\kappa = 0$. For NLP tasks, the inter-annotator agreement is considered as significant when $\kappa > 0.6$. Applying such formula to our data, the inter-annotator agreement results in $\kappa = 0.82$. As a rule of thumb, this is a satisfactory agreement, therefore we consider these annotated datasets as reliable (i.e., our *goldstandard* dataset where arguments are associated to participants' emotions detected by EEG/FaceReader) to be exploited during the experimental phase.

Table 1 reports on the number of arguments and pairs we extracted applying the methodology described before to all the mentioned topics. In total, our dataset contains 598 different arguments and 263 argument pairs (127 expressing the *support* relation among the involved arguments, and 136 expressing the *attack* relation among the involved arguments).

The dataset resulting from these three layers of annotation adds to all previously annotated information the player characteristics (gender, age and personality type), FaceReader data (dominant emotion, Valence (pleasant/unpleased) and Arousal (activated/inactivated)), and EEG data (Mental Engagement levels).[9] A correlation matrix has been generated to identify the correlations between arguments and emotions in the debates, and a data analysis is performed to determine the proportions of emotions for all participants. We consider the obtained dataset as the reference dataset to carry out our empirical study.

An example, from the debate about the topic "Religion does more harm than good" where arguments are annotated with emotions (i.e., the third layer of the annotation of the textual arguments we retrieved), is as follows:

```
<argument id="30" debate_id="4" participant="4"
```

[9]The datasets of textual arguments are available at http://project.inria.fr/seempad/datasets/.

Table 1
The dataset of argument pairs resulting from the experiment

Topic	Dataset			
	#arg	#pair	#att	#sup
BAN ANIMAL TESTING	49	28	18	10
GO NUCLEAR	40	24	15	9
HOUSEWIVES SHOULD BE PAID	42	18	11	7
RELIGION DOES MORE HARM THAN GOOD	46	23	11	12
ADVERTISING IS HARMFUL	71	16	6	10
BULLIES ARE LEGALLY RESPONSIBLE	71	12	3	9
DISTRIBUTE CONDOMS IN SCHOOLS	68	27	11	16
ENCOURAGE FEWER PEOPLE TO GO TO THE UNIVERSITY	55	14	7	7
FEAR GOVERNMENT POWER OVER INTERNET	41	32	18	14
BAN PARTIAL BIRTH ABORTIONS	41	26	15	11
USE RACIAL PROFILING FOR AIRPORT SECURITY	31	10	1	9
CANNABIS SHOULD BE LEGALIZED	43	33	20	13
TOTAL	**598**	**263**	**136**	**127**

```
time-from="20:43" time-to="20:43"
emotion_p1="neutral" emotion_p2="neutral"
emotion_p3="neutral" emotion_p4="neutral">
Indeed but there exist some advocates of the devil
like Bernard Levi who is decomposing arabic
countries. </argument>

<argument id="31" debate_id="4" participant="1"
time-from="20:43" time-to="20:43"
emotion_p1="angry" emotion_p2="neutral"
emotion_p3="angry" emotion_p4="disgusted">
I don't totally agree with you Participant2:
science and religion don't explain each other,
they tend to explain the world but in two
different ways.</argument>

<argument id="32" debate_id="4" participant="3"
time-from="20:44" time-to="20:44"
emotion_p1="angry" emotion_p2="happy"
emotion_p3="surprised" emotion_p4="angry">
Participant4: for recent wars ok but what
about wars happened 3 or 4 centuries ago?
</argument>
```

Finally, the fourth annotation task starts from the basic one, and it selects for each participant one argument at the beginning of the debate, one argument in the middle of the discussion, and one argument at the end of the debate. These arguments are then annotated with their *polarity* with respect to the issue of the debate: *negative*, *positive*, or *undecided*. The negative polarity is assigned to an argument when the opinion expressed in such argument is against the debated topic, while the positive polarity label is assigned when the argument expresses a viewpoint that is in favor of the debated issue. The undecided polarity is assigned when the argument does not express a precise opinion in favor or against the debated topic. Selected arguments are evaluated as the most representative arguments proposed by each participants to convey her own opinion, in the three distinct moments of the debate. The rationale

behind this annotation is that it allows to easily detect when a participant has changed her mind with respect to the debated topic. An example is provided below from the debate "Ban partial birth abortions", where Participant4 starts the debate being undecided and then turns to be positive about banning partial birth abortions in the middle and at the end of the debate:

```
<argument id="5" participant="4" time-from="20:36"
time-to="20:36" polarity="undecided">Description's
gruesome but does the fetus fully lives at that
point and therefore, conscious of something ? Hard
to answer. If yes, I might have an hesitation to
accept it. If not, the woman is probably mature
enough to judge.</argument>

<argument id="24" participant="4" time-from="20:46"
time-to="20:46" polarity="positive">In the animal
world, malformed or sick babies are systematically
abandoned.</argument>

<argument id="38" participant="4" time-from="20:52"
time-to="20:52" polarity="positive">Abortion is
legal and it doesn't matter much when and how.
It's an individual choice for whatever reason
it might be.</argument>
```

3.3. Hypotheses

The experiment we have carried out aims at verifying the link between the emotions detected on the participants of the debate, and the arguments and their relations proposed in the debate. Our hypotheses therefore revolve around the assumption that the participants' emotions arise out of the arguments they propose in the debate:

H1: The argumentation process in a debate requires high mental engagement and generates negative emotions when the interlocutor's arguments are attacked.
H2: The number of arguments and attacks proposed by the debaters are correlated with negative emotions.
H3: The number of expressed arguments is connected to the degree of mental engagement and social interactions.
H4: The personality of the participants modulates their emotional experiences during the debates.
H5: The debaters' opinions regarding the discussed topics have an impact on their emotions.

4. Results

In order to verify the above mentioned hypotheses, we first computed the mean percentage of appearance of each basic emotion across the 20 participants. Results show (with 95% of confidence interval) that the most frequent emotion expressed by participants was *anger*, with a mean appearance frequency ranging from 8.15% to 15.6% of the times. The second most frequent emotion was another negative emotion, namely *disgust*, which was present 7.52% to 14.8% of the times. The overall appearance frequency of other emotions was very low. For example, the frequency of appearance of happiness was below 1%. Even if this result might be surprising at a first glance, this trend can be justified by a phenomenon called *negativity effect* [41]. This means that negative emotions have generally more impact on

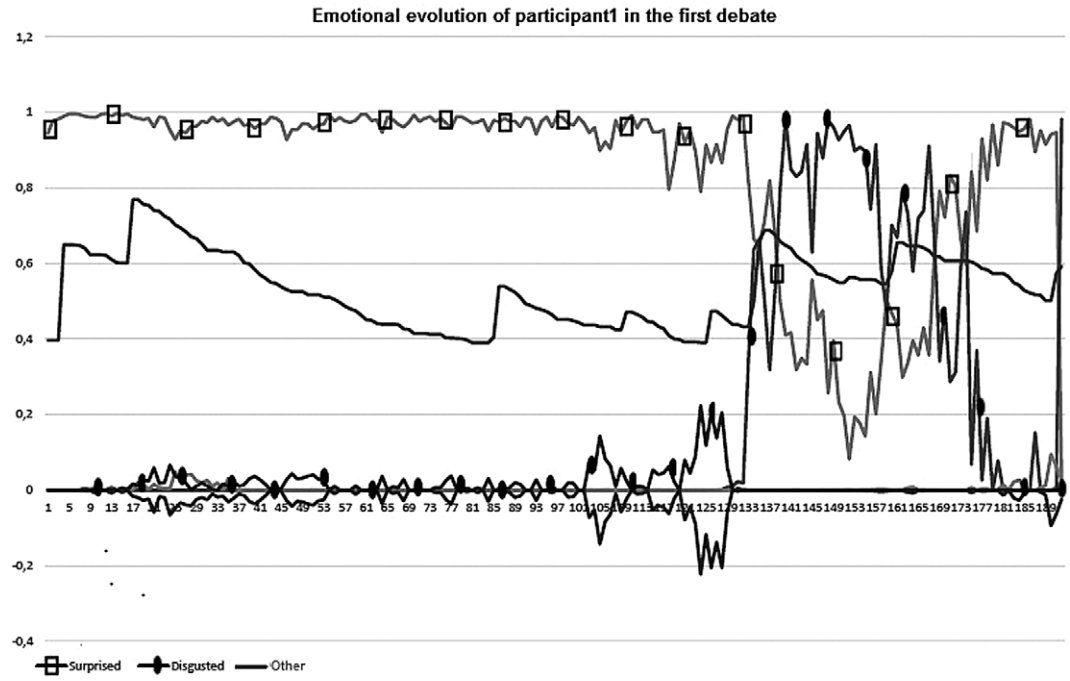

Fig. 2. Emotional evolution of Participant 1 in Debate 1 (lines with squares and circles represent, respectively, the *surprise* and *disgust* emotions).

a person's behavior and cognition than positive ones. So, negative emotions like anger and disgust have a tendency to last in time more than positive emotions like happiness.

With regard to the mental engagement, participants show in general a high level of attention and vigilance in 70.2% to 87.7% of the times. This high level of engagement is also correlated with appearance of anger ($r = 0.306$), where r refers to the Pearson product-moment correlation coefficient. This coefficient is a standard measure of the linear correlation between two variables X and Y, giving a value between $[1, -1]$, where 1 is a total positive correlation, 0 means no correlation, and -1 is a total negative correlation. This trend confirms that, in such context, participants may be thwarted by the other participant arguments or attacks, thus the level of engagement tends to be high as more attention is allowed to evaluate the other arguments or to formulate rebutting or offensive arguments. Thus, our experiments confirm behavioral trends as expected by the first hypothesis.

Figure 2 shows an evolution of the first participant's emotions at the beginning of the first debate. The most significant lines of emotions are surprise and disgust (respectively, the line with squares and the line with circles). The participant is initially surprised by the discussion (and so mentally engaged) and then, after the debate starts, this surprise switches suddenly into disgust, due to the impact of the rejection of one of her arguments; the bottom line with circles grows and replaces the surprise as the participant is now actively engaged in an opposed argument (thus confirming our hypothesis 2). Finally, the participant is calming down. In this line, Fig. 3 highlights that we have a strong correlation ($r = 0.83$) in Session 2 showing that the number of attacks provided in the debate increased linearly with the manifestation of more disgust emotion.

In the second part of our study, we were interested in analyzing how emotions correlate with the number of attacks, supports and arguments. We have generated a correlation matrix to identify the existent

	NB ARG	ATTACK	SUPPORT
Pleasant	0,0962	0,1328	-0,0332
Unpleasant	-0,0962	-0,1328	0,0332
High ENG	-0,0718	**-0,6705**	0,2459
LowENG	-0,2448	0,2115	-0,1063
Neutral	0,0378	0,6173	-0,1138
Disgusted	-0,0580	**-0,4367**	-0,3621
Scared	0,1396	-0,0952	**0,5755**
Angry	-0,1018	-0,4386	0,0582

Fig. 3. Correlation table for Session 2 (debated topics: *Advertising is harmful* and *Bullies are legally responsible*.

	NB ARG	ATTACK	SUPPORT
Pleasant	**0,7067**	-0,3383	-0,3800
Unpleasant	-0,7067	0,3383	0,3800
High ENG	**-0,6903**	-0,3699	-0,1117
LowENG	-0,1705	**0,5337**	-0,0615
Neutral	0,8887	-0,0895	-0,3739
Disgusted	0,1017	**0,8379**	**0,5227**
Scared	0,2606	-0,4132	**-0,7107**
Angry	**-0,7384**	-0,5072	-0,0937

Fig. 4. Correlation table for Session 3 (debated topics: *Distribute condoms at schools* and *Encourage fewer people to go to the university*).

correlations between arguments and emotions in debates. Main results show that the number of arguments tends to decrease linearly with manifestations of sadness ($r = -0.25$). So when the participants start to feel unpleasant emotions, such as sadness, the number of arguments decreases, showing a less positive social behavior[10] and a tendency to retreat into herself. This negative correlation between the number of arguments and sadness even reaches very high level in certain debates (i.e. a mean correlation $r = -0.70$ is registered in the two debates of the second session). Another negative linear relationship is registered with regard to the number of attacks and the anger expressed by the participant ($r = 0.22$). Participants who tend to attack the others in the debate are less angry than those whose number of attacks is smaller. Figure 4 shows the correlation table for Session 3. The analysis of the results we obtained shows the occurrence of strong correlations between emotions and attacks/media/number of arguments in some discussions, but not in others. This is an interesting index to investigate in future work.

Figure 5 shows the most significant correlations we detected. For instance, the number of supports provided in the debate increased linearly with the manifestation of high levels of mental engagement ($r = 0.31$). This trend is more pronounced when the debate does not trigger controversies and conflicts between the participants. For example, in the debate *Encourage fewer people to go to university*, all the participants shared the same opinion (against the main issue as formulated by the moderator) and engaged to support each other's arguments. The correlation between the number of supports and the engagement was very high ($r = 0.80$) in this debate. The number of attacks is more related to low engagement. The moderator can provide more supporting arguments to balance participants' engagement, and if the attacks are increasing, that means participants tend to disengage. The experiments show that

[10]By positive social behavior, we mean that a participant aims at sharing her arguments with the other participants. This attitude is mitigated if unpleasant emotions start to be felt by the participant.

	NB ARG	ATTACK	SUPPORT
Pleasant	**0,1534**	0,0134	-0,0493
Unpleasant	-0,1534	-0,0134	0,0493
High ENG	-0,0246	-0,0437	**0,3185**
LowENG	**0,2054**	0,1147	0,1592
Neutral	0,0505	0,1221	-0,2542
Disgusted	-0,0177	-0,0240	**0,2996**
Scared	-0,0278	0,0297	-0,2358
Angry	0,0344	**-0,2206**	0,0782

Fig. 5. General correlation table of the results.

participants maintaining high levels of vigilance are the most participative in the debate and resulted in a more positive social behavior (thus confirming our hypothesis 3).

Our next objective was to check whether participants' emotions during the debates were modulated by their personality. In other words, we wanted to see whether there was any impact of the debaters' personality on their emotional responses in terms of facial expressions, valence, engagement and cognitive load indexes. The Big Five inventory data [26] were considered for this analysis. Participants were classified according to each of the five OCEAN personality dimensions, namely *openness* (imaginative vs. more pragmatic participants), *conscientiousness* (conscientious vs. non conscientious), *extroversion* (extroverted vs. introverted), *agreeableness* (compassionate toward others vs. more antagonistic), and *neuroticism* (anxious vs. emotionally stable). Multivariate analyses of variance (MANOVA) were run with the OCEAN personality traits as fixed factors and the debaters' emotions as dependent variables. These included the following combined measures: (1) the proportions of occurrences of the six facial expressions, i.e. happiness, anger, fear, sadness, disgust, and surprise, (2) the proportions of negative and positive emotional valence, (3) the proportions of high, medium and low levels of mental engagement, and (4) the proportions of high, medium and low workload. In total, 20 MANOVA were conducted, crossing the 5 personality traits with the 4 dependent variables.

Three statistically reliable MANOVA were found, showing significant relationships between the debaters' personality traits and emotional responses:[11]

- Extroversion and facial expressions ($F(6, 33) = 2.574$, $p < 0.05$, Pillai's Trace $= 0.319$). In particular, extroverted participants showed significantly more frequently expressions of surprise than the introverted participants ($M = 6.70\%$, SE $= 1.10\%$ vs. $M = 1.70\%$, SE $= 1.30\%$; $F(1, 38) = 8.385$, $p < 0.008$). This can be explained by the fact that introverted people tend to hide their emotions as compared to extroverted people.
- Conscientiousness and workload ($F(3, 36) = 5.200$, $p < 0.05$, Pillai's Trace $= 0.302$). More precisely, participants with a non conscientious temperament had significantly more occurrences of low levels of workload as compared to the other participants ($M = 29.6\%$, SE $= 3.8\%$ vs. $M = 16.0\%$, SE $= 3.8\%$; $F(1, 38) = 6.525$, $p < 0.016$). They seemed to experience on average less cognitive load during the discussions.
- Neuroticism and mental engagement ($F(3, 36) = 3.518$, $p < 0.05$, Pillai's Trace $= 0.227$). In particular, participants with an anxious temperament had on average significantly fewer proportions of high engagement levels during the debates as compared to the other participants ($M = 18.0\%$, SE $= 2.4\%$ vs. $M = 28.5\%$, SE $= 3.0\%$; $F(1, 38) = 7.423$, $p < 0.016$). This can be seen as

[11] A Bonferroni correction (0.05 divided by the number of dependent variables) has been applied within the MANOVA follow-up analyses to account for multiple ANOVAs being run.

Table 2
Number of opinions before and after the debates (in bold the number of debaters who kept the same opinion)

Starting/Final	No-opinion	For	Against	*Total*
No-opinion	**2**	5	0	7
For	0	**12**	1	13
Against	0	1	**19**	20
Total	2	18	20	40

follows: anxious people tend to be easily stressed. They are thus likely to have trouble concentrating, and hence have difficulties being mentally engaged, as opposed to less emotionally vulnerable people.

To summarize, these results validate our fourth hypothesis: the personality has an important impact on the debaters' emotional responses. Inner emotions (brain activity) seem to be modulated by the neuroticism and the conscientiousness temperament traits. Outer emotions (facial expressions) were modulated by the extroversion traits. Neuroticism and conscientiousness have both a negative impact on the debaters' brain indexes, with respectively, a reduced mental engagement index and an increased cognitive load. For facial expressions, we have particularly found that the emotion of surprise was more frequent among the debaters with an extroverted temperament. This is an important aspect considering that this expression was the least observed during our experiments. Indeed when analyzing the debaters' emotions with FaceReader, we observed that the expression of surprise was hardly dominant (compared with neutral) during the discussions. The dominance of the other facial expressions, namely anger, fear, sadness and disgust does not seem to be influenced by the participants' personality.

Our next concern was to investigate if there were any differences in terms of emotional experiences (facial expressions, emotional valence, mental engagement and workload) during the debates according to the participants' opinions on the discussed topics. These opinions were given during the debriefing as previously mentioned. Each participant was either for or against the topic of the debate: *for* means that the participant agreed with the subject of the debate (e.g. for distributing condoms to students), and *against* means that the participant disagreed with the addressed topic (e.g. against distributing condoms in schools). A participant could also have no particular opinion (*no-opinion*) regarding the topic of the debate if he was neither for nor against. Moreover, each participant was asked to give a starting opinion, before the discussion, and a final opinion, after the debate. The goal was to assess the impact, if any, of changes in opinions on the debaters' emotions. Table 2 enumerates participants' initial and final opinions.

As for the previous hypothesis, distinct MANOVA were performed to analyze the proportions of occurrence of *(1)* facial expressions: happy, sad, angry, surprised, scared and disgusted; *(2)* valences of emotions: positive and negative; *(3)* engagement levels: high, medium and low; and *(4)* cognitive load levels: high, medium and low. First, we wanted to check whether there were any significant differences in terms of emotions between the participants who kept the same opinion during the debates ($N = 33$) and the participants who changed their opinion ($N = 7$). Then, for those who kept the same opinion, we wanted to compare the emotional responses between the participants who were for and the participants who were against.[12]

[12]The two participants who did not have an opinion throughout the debate were discarded since they have reported they could not follow the discussions because of their lack of understanding of English.

No statistically reliable effect was found in any of the performed analyses (p = n.s.) suggesting that there were no significant differences in terms of facial expressions, valence, engagement and workload, neither between the debaters who kept the same opinions and the debaters who changed their opinion, nor between those who were for and those who were against the discussed topics throughout the debates.

In addition to these analyses, the debaters' starting and final opinions were studied independently. That is, we checked whether either the former or the latter opinions had (independently of each other) an impact on the emotions expressed during the debates. Again no statistically significant effect was found. This has led us to conclude that neither the initial nor the final opinions had an impact on the debaters' emotional states. To summarize, the emotional experience during the debates does not seem to be related to the opinion of the debaters regarding the addressed topics. Emotions are rather depending on the person's temperament and the dynamics of the debate (i.e. arguments, supports and attacks).

4.1. Examples of correlations on single debates

In this section we provide some examples of correlations among the emotions and the argumentative elements emerging from the single debates. The goal is to provide a more detailed analysis, given the fact that some debates have been more passionate than others because of the personal involvement of the participants in the subject of the debate. It is worth noticing that as the number of instances involved in our debates is 4, these correlations cannot be considered as significative. However, we believe that these examples may show interesting insights to be investigated in our future experiments. The categories of correlations we investigate are the following: engagement vs argumentation; engagement vs emotions; workload vs argumentation; workload vs emotions; pleasant/unpleasant vs argumentation; Big5 vs emotions. Correlation values are comprised between -1 (negative correlation) and $+1$ (positive correlation). In our analysis we consider as strong correlations the values between -0.7 and -1 (strong negative correlation), and between 0.7 and 1 (strong positive correlation). All values in between cannot be considered significant to verify our hypothesis.

Debate: Advertising is harmful. Number of arguments in the debate: 71 (64 from participants and 7 from moderators).

Workload vs emotions: The more the participants feel surprised, the higher the workload is high ($r = 0.80$) meaning that the mental load increases if the participants feel surprised about the arguments proposed in the ongoing debate. Moreover, the more the participants feel neutral, the lower the workload ($r = 0.82$) meaning that the mental load decreases if the participants feel neutral with respect to the ongoing debate, i.e., they are not interested in the topic of the debate as well as in the the other participants' arguments.

Big5 vs emotions: Participants with a high degree of *agreeableness* are more inclined to be surprised ($r = 0.90$), while participants with a high degree of *conscientiousness* tend to get sad or angry if the debate is not going in the desired direction (correlations $r = 0.82$ and $r = 78$, respectively), since they are inclined to do their duty well and thoroughly.

Debate: Students are legally responsible for bulling. Number of arguments in the debate: 71 (66 from participants and 5 from moderators).

Engagement vs emotions: On the one side, we have a strong correlation between the high engagement and the emotion *happy* ($r = 0.94$), meaning that when the participants of this debate are experiencing such positive emotion they become more passionate (and therefore engaged) in the

discussed topic. On the other side, we have also a strong correlation between the low engagement and the emotion *disgusted* ($r = 0.82$), meaning that when participants experience such negative emotion, then they become less interested in the ongoing debate.

Pleasant/Unpleasant vs argumentation: Strong correlation between positive valence (pleasant) and the number of arguments proposed in the debate ($r = 0.74$), meaning that when participants propose more arguments in the debate, they are more interested in the debated topic and therefore there is a higher degree of pleasantness in the air.

Big5 vs emotions: Participants with a high degree of *extroversion* or of *agreeableness* are more inclined to externalize that they fell surprised about the ongoing debate (correlations $r = 0.96$ and $r = 89$, respectively). Moreover, participants with a high degree of *neuroticism* are more inclined to be disgusted about the arguments proposed in the debate ($r = 0.82$).

Debate: Distribute condoms in schools. Number of arguments in the debate: 68 (63 from participants and 5 from moderators).

Engagement vs emotions: Strong correlation between the high engagement and the emotion *angry* ($r = 0.96$), meaning that when the participants are experiencing such negative emotion they become more passionate (and therefore engaged) in the discussed topic.

Workload vs argumentation: Strong correlation between a high degree of workload and the number of supports among the arguments ($r = 0.86$), meaning that when the number of supports increases then the participants of this debate are required with a higher mental load to understand how the debate is going on.

Workload vs emotions: The more the participants feel disgusted, the higher the workload is ($r = 0.92$) meaning that the mental load increases if the participants feel disgusted about the arguments proposed in the ongoing debate, as they need to construct in their minds new arguments to defeat the ones proposed by the other participants that make them feel disgusted.

Big5 vs emotions: Participants with a high degree of *extroversion* are more inclined to externalize that they fell surprised about the ongoing debate ($r = 0.77$). Moreover, participants with a high degree of *neuroticism* are more inclined to be angry about the arguments proposed in the debate ($r = 0.75$).

Debate: We should fear the power of government over the Internet. Number of arguments in the debate: 41 (37 from participants and 4 from moderators).

Engagement vs emotions: Strong correlation between the high engagement and the emotion *disgusted* ($r = 0.93$), meaning that when the participants are experiencing such negative emotion they become more passionate (and therefore engaged) in the discussed topic.

Workload vs argumentation: Strong correlation between a low degree of workload and the number of supports among the arguments ($r = 0.86$), meaning that when the number of supports increases participants require a lower mental load to understand how the debate is going on.

Workload vs emotions: The more the participants feel *angry*, the lower the workload is ($r = 0.93$). This means that those participants that become angry due to the arguments that are proposed in the ongoing debate tend to use less mental resources to propose new, possibly effective, arguments.

Pleasant/Unpleasant vs argumentation: Strong correlation between negative valence (unpleasant) and the number of attacks between arguments ($r = 0.70$), meaning that when participants disagree attacking each others there is an higher degree of unpleasantness in the air.

Big5 vs emotions: Participants with a high degree of *extroversion* are more inclined to externalize that they feel happy about the ongoing debate ($r = 0.99$). Moreover, participants with a high degree of *neuroticism* are more inclined to be disgusted about the arguments proposed in the debate ($r = 0.90$).

Note that this study is dealing with correlation between the nature of the arguments and their relations (support/attack) and the participants' emotions, and we are not claiming to have found a direct causal relation between the arguments and such users' emotions – which is out of the scope of this current study. It is however an interesting direction for further work with larger sample size using [21]'s causality test.

4.2. Discussion

We have learnt several lessons from the realized experiment. First, the different debates and participants have confirmed the correctness of our hypotheses. Debates constitute the underlying framework for generating emotions which evolve with the argumentation flow. The difference of opinions is the starting point of the rise and successive transformation of specific emotions. However, so far we have not taken into account the initial emotional state of the participants (i.e. before starting the discussion on the topic), that can influence the participants reactions during the debate. We will have to consider this in further studies.

Facial emotion recognition and EEG measures allowed us to identify not only the type of emotion generated, but also the intensity and the evolution of emotions. Associated with the workload index, this also allowed us to detect how the participant is engaged in the discussion and so, how he holds on to his arguments. Being strongly convinced by an opinion provokes the birth of a mental energy strong enough to increase the workload and develop a justification. Contradictions with the flow of arguments generate anger which evolves progressively into disgust if the participant's arguments apparently cannot convince the opponents. In the classification of emotion, disgust (which is close in terms of emotion) is a normal evolution of anger and appears when the participant feels a dual feeling for two reasons: 1) he is not pleased with himself for not having convinced the opponent (internal feeling), and 2) he has a very low opinion of the opponent (external feeling). We highlighted the evolution of this emotion in several debates showing the important consequence of the argumentation by provoking internal evolution of emotions. This can be explained by the impact of a contradiction on an in depth conviction. The more a participant is convinced of the merits of his position, the more he will be subject to a strong emotion.

The three dimensions of our evaluation framework (emotion recognition, engagement and workload) allowed us to assess more precisely the impact of argumentation on emotional response throughout the debate. Workload decreases when participants feel angry, which means that they reduce their ability to use or construct new arguments. When this emotion evolves to disgust, the workload increases which means that they have to reconsider their own set of arguments either for an update or a new construction. High engagement provokes the rise of strong positive or negative emotions while, on the other side, we have confirmed that disgusted participants become less engaged in the ongoing debate. Finally, we have considered the influence of personality to the type of generated emotion. Participants with a high degree of neuroticism are converging to be angry or disgusted, which are close emotions. Participants with a high degree of agreeableness or extroversion are more open to feel surprise.

Note that the goal of this experiment is not to learn how to best intervene to improve online discourse but to study what are the insights that online cognitive agents and bots need to implement to address dialogues with humans. A cognitive agent, in order to behave like humans in debates, has to feel emotions and generate them in the other agents (being them humans or artificial) that interact with it. This

extensive study is the first but essential step towards a better comprehension of the relation between human emotions and argumentation. As a shorter term objective, the aim of this contribution is to guide the definition of the next argumentation frameworks such that not only objective elements are taken into account but also cognitive ones.

From this experiment, we learnt that argumentation in online debates cannot be considered as a standalone process, as it discloses many emotional aspects, e.g., when users are more engaged in a discussion more arguments are proposed, and the most engaging discussions are correlated with negative emotions like anger and disgust. Moreover, a strong correlation exists among personality traits and the emotions felt by participants during online argumentation, e.g., the dominance of emotions like anger, fear, sadness and disgust does not seem to be influenced by the participants' personality where emotions of happiness and surprise were more frequent among the debaters with an extroverted temperament.

5. Related work

A first analysis of the experimental results presented in this paper has been proposed in [2]. However, several aspects of the collected data were neglected in that work. In this extended version, the following issues have been tackled:

- *Personality traits, emotions and argumentation*: in [2], we did not consider in our analysis the Big Five inventory data we collected during the experiments. In this paper, an additional hypothesis is formulated concerning the connection among participants' personality and emotions. The hypothesis has been then validated on the data collected from our experiments.
- *Opinions and emotions*: in [2], we did not consider the opinions of the participants with respect to the debated topics, their possible change during the debate, and the emotions. In this paper, a fifth hypothesis is formulated and then validated on the collected data.
- *Correlations on single debates*: in [2], we considered correlations holding over the whole set of debates, i.e., over the whole set of collected data. However, we realized that some of the debates showed significant correlations that were not present in others, due to the involvement of different participants and to the interest of the participants in the debated topic. In this paper, we have proposed an analysis of some of the more relevant debates held in our experimental sessions. These single debate analysis considers also the workload index, computed at the data collection time but never discussed in the first version of the paper [2].
- *Fourth layer of annotation*: in [2], three layers of annotation have been proposed over the collected textual data. In this paper, we included a fourth annotation layer with the aim to highlight the change in the viewpoint of the participants during the debate.

In the literature, only few works deal with empirical experiments involving human participants to verify assumptions from argumentation theory. Cerutti et al. [8] propose an empirical experiment with humans in the argumentation theory area. However, the goal of this work is different from ours, since emotions are not considered and their aim is to show a correspondence between the acceptability of arguments by human subjects and the acceptability prescribed by the formal theory in argumentation. Rahwan and colleagues [38] study whether the meaning assigned to the notion of *reinstatement* in abstract argumentation theory is perceived in the same way by humans. They propose to the participants of the experiment a number of natural language texts where reinstatement holds, and then ask them to evaluate the arguments. Also in this case, the purpose of the work differs from ours, and emotions are not considered at all.

Emotions are considered, instead, by Nawwab et al. [33] that propose to couple the model of emotions introduced by Ortony and colleagues [34] in an argumentation-based decision making scenario. They show how emotions, e.g., gratitude and displeasure, impact on the practical reasoning mechanisms. A similar work has been proposed by Dalibon et al. [12] where emotions are exploited by agents to produce a line of reasoning according to the evolution of its own emotional state. Finally, Lloyd-Kelly and Wyner [32] propose emotional argumentation schemes to capture forms of reasoning involving emotions. All these works differ from our approach since they do not address an empirical evaluation of their models, and emotions are not detected from humans.

Several works in philosophy and linguistics have studied the link between emotions and natural argumentation, like [6,20,45]. These works analyze the connection of emotions and the different kind of argumentation that can be addressed. The difference with our approach is that they do not verify their theories empirically, on emotions extracted from people involved in an argumentation task. A particularly interesting case is that of the connection between persuasive argumentation and emotions, studied for instance by DeSteno and colleagues [13].

Concerning the empirical study of workload and emotional changes, [46] study pupillary response to detect workload and emotional changes performing an arithmetical task associated with pleasant/unpleasant images. The idea of the empirical study on workload and emotional changes is similar, even if the goal of the experiment is different, as our goal is connected to the argumentative process and not with arithmetical tasks performed by isolated participants.

6. Conclusions

In this paper, we have presented an investigation into the links between the argumentation people use when they debate with each other, the emotions they feel during these debates, and their personality traits. We conducted an experiment aimed at verifying our hypotheses about the correlation between the positive/negative emotions emerging when positive/negative relations among the arguments are put forward in the debate, and the correlation between the personality traits of the debaters and their opinions on the debated topics, and the emotions felt during the debate interactions. The results suggest that there exist trends that can be extracted from emotion analysis. Moreover, we also provide the first annotated dataset and gold standard to compare and analyze emotion detection in an argumentation session.

The take-home message of this paper is twofold: first, high engagement is correlated with negative emotions showing that participants are mentally involved in producing arguments to rebut those which are not in line with their viewpoint, and second, neuroticism and conscientiousness have both a negative impact on the debaters' brain indexes ending up into a reduced mental engagement index and an increased cognitive load. Finally, the surprise emotion is shown by extroverted debaters.

Several lines of research have to be considered as future work. First, we intend to study the link between emotions and persuasive argumentation. This issue has already been tackled in a number of works in the literature [13], but no empirical evaluation has been addressed so far. Second, we aim to study how emotions persistence influences the attitude of the debates: this kind of experiment has to be repeated a number of times in order to verify whether positive/negative emotions before the debate influence new interactions. Third, we plan to add a further step, namely to study how sentiment analysis methods developed in Computational Linguistics are able to automatically detect the polarity of the arguments proposed by the debaters, and how they are correlated with the detected emotions. More precisely, the annotated dataset we published provides a valuable resource to improve the performances of sentiment

analysis systems allowing them to learn about the correlation among the relations among the arguments and the emotions aligned with the arguments. Moreover, we plan to study emotions propagation among the debaters, and to verify whether an emotion can be seen as a predictor of the solidity of an argument, e.g., if I write an argument when I am angry I may make wrong judgments. Finally, argumentation theory has often been proposed as a technique for supporting *critical thinking*, thus studying the relation of these philosophical theories with emotions and personality traits of the actors involved in the argumentation is a further step to investigate.

Acknowledgements

The authors acknowledge support of the SEEMPAD associate team project (http://project.inria.fr/seempad/). The authors from the University of Montreal acknowledge support of the NSERC (National Science and Engineering Research Council of Canada).

Appendix

Big Five questionnaire

1. I am the life of the party
2. I feel little concern for others
3. I am always prepared
4. I get stressed out easily
5. I have a rich vocabulary
6. I don't talk a lot
7. I am interested in people
8. I leave my belongings around
9. I am relaxed most of the time
10. I have difficulty understanding abstract ideas
11. I feel comfortable around people
12. I insult people
13. I pay attention to details
14. I worry about things
15. I have a vivid imagination
16. I keep in the background
17. I sympathize with other's feelings
18. I make a mess of things
19. I seldom feel blue
20. I am not interested in abstract ideas
21. I start conversations
22. I am not interested in other's people problems
23. I get chorves done right away
24. I am easily disturbed
25. I have excellent ideas
26. I have little to say

27. I have a soft heart
28. I often forget to put things back in their proper place
29. I get upset easily
30. I do not have a good imagination
31. I talk to a lot of different people at parties
32. I am not really interested in others
33. I like order
34. I change my mood a lot
35. I am quick to understand things
36. I don't like to draw attention on myself
37. I take time out for others
38. I shirk my duties
39. I have frequent mood swings
40. I use difficult word
41. I don't mind being the center of attention
42. I feel other's emotions
43. I follow a schedule
44. I get irritated easily
45. I spend time reflecting on things
46. I am quiet around strangers
47. I make feel people at ease
48. I am exacting in my work
49. I often feel blue
50. I am full of ideas

References

[1] I. Arroyo, D.G. Cooper, W. Burleson, B.P. Woolf, K. Muldner and R. Christopherson, Emotion sensors go to school, in: *Proceedings of the 2009 Conference on Artificial Intelligence in Education: Building Learning Systems That Care: From Knowledge Representation to Affective Modelling*, IOS Press, Amsterdam, The Netherlands, 2009, pp. 17–24.

[2] S. Benlamine, M. Chaouachi, S. Villata, E. Cabrio, C. Frasson and F. Gandon, Emotions in argumentation: An empirical evaluation, in: *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015*, Buenos Aires, Argentina, July 25–31, 2015, Q. Yang and M. Wooldridge, eds, AAAI Press, 2015, pp. 156–163.

[3] E.T. Berkman and E.B. Falk, *Curr. Dir. Psychol. Sci.* **22** (2013), 45–50. doi:10.1177/0963721412469394.

[4] E. Cabrio and S. Villata, A natural language bipolar argumentation approach to support users in online debate interactions, *Argument & Computation* **4**(3) (2013), 209–230. doi:10.1080/19462166.2013.862303.

[5] J. Carletta, Assessing agreement on classification tasks: The kappa statistic, *Computational Linguistics* **22**(2) (1996), 249–254.

[6] V. Carofiglio and F. de Rosis, Combining logical with emotional reasoning in natural argumentation, in: *9th International Conference on User Modeling. Workshop Proceedings*, C. Conati, E. Hudlicka and C. Lisetti, eds, 2003, pp. 9–15.

[7] C. Cayrol and M.-C. Lagasquie-Schiex, Bipolarity in argumentation graphs: Towards a better understanding, *Int. J. Approx. Reasoning* **54**(7) (2013), 876–899. doi:10.1016/j.ijar.2013.03.001.

[8] F. Cerutti, N. Tintarev and N. Oren, Formal arguments, preferences, and natural language interfaces to humans: An empirical evaluation, in: *ECAI 2014 – 21st European Conference on Artificial Intelligence*, 2014, pp. 207–212.

[9] G. Chanel, C. Rebetez, M. Bétrancourt and T. Pun, Emotion assessment from physiological signals for adaptation of game difficulty, *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on* **41**(6) (2011), 1052–1063.

[10] M. Chaouachi, I. Jraidi and C. Frasson, Modeling mental workload using eeg features for intelligent systems, in: *User Modeling, Adaption and Personalization*, Springer, 2011, pp. 50–61.

[11] M. Chaouachi, I. Jraidi and C. Frasson, Mentor: A physiologically controlled tutoring system, in: *User Modeling, Adaptation and Personalization*, Springer, 2015, pp. 56–67. doi:10.1007/978-3-319-20267-9_5.

[12] S.E.F. Dalibón, D.C. Martinez and G.R. Simari, Emotion-directed argument awareness for autonomous agent reasoning, *Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial* **15**(50) (2012), 30–45.
[13] D. DeSteno, D.T. Wegener, R.E. Petty, D.D. Rucker and J. Braverman, Discrete emotions and persuasion: The role of emotion-induced expectancies, *Journal of Personality and Social Psychology* **86** (2004), 4356.
[14] P.M. Dung, On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games, *Artif. Intell.* **77**(2) (1995), 321–358. doi:10.1016/0004-3702(94)00041-X.
[15] N. Duy, R.P. Jan and S. Matthias, Local Gaussian process regression for real time online model learning, in: *Advances in Neural Information Processing Systems*, D. Koller, D. Schuurmans, Y. Bengio and L. Bottou, eds, Vol. 21, Curran Associates, 2009, pp. 1193–1200.
[16] E.B. Falk, C.N. Cascio and J.C. Coronel, Neural prediction of communication-relevant outcomes, *Communication Methods and Measures* **9** (2015), 30–54. doi:10.1080/19312458.2014.999750.
[17] F. Freeman, P. Mikulka, L. Prinzel and M. Scerbo, Evaluation of an adaptive automation system using three eeg indices with a visual tracking task, *Biological psychology* **50**(1) (1999), 61–76. doi:10.1016/S0301-0511(99)00002-2.
[18] F.G. Freeman, P.J. Mikulka, M.W. Scerbo, L.J. Prinzel and K. Clouatre, Evaluation of a psychophysiologically controlled adaptive automation system, using performance on a tracking task, *Applied Psychophysiology and Biofeedback* **25**(2) (2000), 103–115. doi:10.1023/A:1009566809021.
[19] N.H. Frijda, *The Emotions. Studies in Emotion and Social Interaction*, Cambridge University Press, 1986.
[20] M.A. Gilbert, Emotional argumentation, or, why do argumentation theorists argue with their mates? in: *Proceedings of the Third ISSA Conference on Argumentation*, F.H. van Eemeren, R. Grootendorst, J.A. Blair and C.A. Willard, eds, Vol. II, 1995.
[21] C.W.J. Granger, Investigating causal relations by econometric models and cross-spectral methods, *Econometrica* **37** (1969), 424–438. doi:10.2307/1912791.
[22] J.J. Gross, Emotion regulation: Affective, cognitive, and social consequences, *Psychophysiology* **39** (2002), 281–291. doi:10.1017/S0048577201393198.
[23] S.G. Hart and L.E. Stavenland, Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research, in: *Human Mental Workload*, P.A. Hancock and N. Meshkati, eds, Elsevier, 1988, pp. 139–183, chapter 7. doi:10.1016/S0166-4115(08)62386-9.
[24] M.E. Hoque, D. McDuff and R.W. Picard, Exploring temporal patterns in classifying frustrated and delighted smiles, *T. Affective Computing* **3**(3) (2012), 323–334. doi:10.1109/T-AFFC.2012.11.
[25] C.E. Izard, *The Psychology of Emotions*, Springer Science & Business Media, 1991.
[26] O.P. John and S. Srivastava, The big-five trait taxonomy: History, measurement, and theoretical perspectives, in: *Handbook of Personality: Theory and Research*, Guilford Press, 1999, pp. 102–138.
[27] S.G.B. Johnson, T. Merchant and F. Keil, Argument scope in inductive reasoning: Evidence for an abductive account of induction, in: *Proceedings of the 37th Annual Meeting of the Cognitive Science Society, CogSci 2015*, Pasadena, California, USA, July 22-25, 2015, D.C. Noelle, R. Dale, A.S. Warlaumont, J. Yoshimi, T. Matlock, C.D. Jennings and P.P. Maglio, eds, 2015, cognitivesciencesociety.org.
[28] I. Jraidi, M. Chaouachi and C. Frasson, A dynamic multimodal approach for assessing learner's interaction experience, in: *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ACM, 2013, pp. 271–278.
[29] I. Jraidi and C. Frasson, Student's uncertainty modeling through a multimodal sensor-based approach, *Educational Technology & Society* **16**(1) (2013), 219–230.
[30] T. Julia, M. Garcia-Garcia and M.E. Smith, Consumer neuroscience: A method for optimising marketing communication, *Journal of Cultural Marketing Strategy* **1** (2015), 80–89.
[31] R.S. Lazarus, *Emotion and Adaptation*, Oxford University Press, 1994.
[32] M. Lloyd-Kelly and A. Wyner, Arguing about emotion, in: *Advances in User Modeling – UMAP 2011 Workshops*, 2011, pp. 355–367.
[33] F.S. Nawwab, P.E. Dunne and T.J.M. Bench-Capon, Exploring the role of emotions in rational decision making, in: *Computational Models of Argument: Proceedings of COMMA 2010*, 2010, pp. 367–378.
[34] A. Ortony, G. Clore and A. Collins, *The Cognitive Structure of Emotions*, Cambridge University Press, Cambridge, 1988.
[35] R. Parasuraman and D. Caggiano, Mental workload, *Encyclopedia of the human brain* **3** (2002), 17–27. doi:10.1016/B0-12-227210-2/00206-5.
[36] A.T. Pope, E.H. Bogart and D.S. Bartolome, Biocybernetic system evaluates indices of operator engagement in automated task, *Biological psychology* **40**(1) (1995), 187–195. doi:10.1016/0301-0511(95)05116-3.
[37] A.C. Rafael and D. Sidney, Affect detection: An interdisciplinary review of models, methods, and their applications, *IEEE Transactions on Affective Computing* **1**(1) (2010), 18–37. doi:10.1109/T-AFFC.2010.1.
[38] I. Rahwan, M.I. Madakkatel, J. Bonnefon, R.N. Awan and S. Abdallah, Behavioral experiments for assessing the abstract argumentation semantics of reinstatement, *Cognitive Science* **34**(8) (2010), 1483–1502. doi:10.1111/j.1551-6709.2010.01123.x.
[39] I. Rahwan and G. Simari (eds), *Argumentation in Artificial Intelligence*, Springer, 2009.

[40] C.E. Rasmussen and C.K.I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*, The MIT Press, 2005.
[41] P. Rozin and E.B. Royzman, Negativity bias, negativity dominance, and contagion, *Personality and social psychology review* **5**(4) (2001), 296–320. doi:10.1207/S15327957PSPR0504_2.
[42] K.R. Scherer, What are emotions? And how can they be measured?, *Social science information* **44**(4) (2005), 695–729. doi:10.1177/0539018405058216.
[43] D. Szafir and B. Mutlu, Pay attention!: Designing adaptive agents that monitor and improve user engagement, in: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2012, pp. 11–20.
[44] D. Szafir and B. Mutlu, Artful: Adaptive review technology for flipped learning, in: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2013, pp. 1001–1010. doi:10.1145/2470654.2466128.
[45] D. Walton, *The Place of Emotion in Argument*, Pennsylvania State University Press, University Park, 1992.
[46] W. Wang, Z. Li, Y. Wang and F. Chen, Indexing cognitive workload based on pupillary response under luminance and emotional changes, in: *18th International Conference on Intelligent User Interfaces, IUI, 13*, Santa Monica, CA, USA, March 19-22, 2013, J. Kim, J. Nichols and P.A. Szekely, eds, ACM, 2013, pp. 247–256.
[47] T. Wehrle and S. Kaiser, Emotion and facial expression, in: *Affective Interactions*, Springer, Berlin Heidelberg, 2000, pp. 49–63. doi:10.1007/10720296_5.
[48] J.J. Williams and T. Lombrozo, The role of explanation in discovery and generalization: Evidence from category learning, *Cognitive Science* **34**(5) (2010), 776–806. doi:10.1111/j.1551-6709.2010.01113.x.