Noname manuscript No.
(will be inserted by the editor)

The Essence of Ethical Reasoning in Robot-Emotion Processing

Suman Ojha · Mary-Anne Williams · Benjamin Johnston

Received: date / Accepted: date

Abstract As social robots become more and more intelligent and autonomous in operation, it is extremely important to ensure that such robots act in socially acceptable manner. More specifically, if such an autonomous robot is capable of generating and expressing emotions of its own, it should also have an ability to reason if it is ethical to exhibit a particular emotional state in response to a surrounding event. Most existing computational models of emotion for social robots have focused on achieving a certain level of believability of the emotions expressed. We argue that believability of a robot's emotions, although crucially necessary, is not a sufficient quality to elicit socially acceptable emotions. Thus, we stress on the need of higher level of cognition in emotion processing mechanism which empowers social robots with an ability to decide if it is socially appropriate to express a particular emotion in a given context or it is better to inhibit such an experience. In this paper, we present the detailed mathematical explanation of the ethical reasoning mechanism in our computational model, EEGS, that helps a social robot to reach to the most socially acceptable emotional state when more than one emotions are elicited by an event. Experimental results show that ethical reasoning in EEGS helps in the generation of believable as well as socially acceptable emotions.

Keywords Social robots · Computational emotion model · Believability · Ethical reasoning · Socially acceptable emotions · EEGS

S. Ojha
University of Technology Sydney
Centre for Artificial Intelligence (CAI)
E-mail: Suman.Ojha@student.uts.edu.au

M-A. Williams
University of Technology Sydney
Centre for Artificial Intelligence (CAI)
E-mail: Mary-Anne.Williams@uts.edu.au

B. Johnston
University of Technology Sydney
Centre for Artificial Intelligence (CAI)
E-mail: Benjamin.Johnston@uts.edu.au

1 Introduction

Realisation of the role of emotion in autonomous agents (like robots [7], virtual assistants [16], embodied conversational characters [5], interactive software [18], etc.) has led to the development of several computational models of emotion[1] [9, 10, 13]. These emotion models, and also the autonomous agents using an emotion model, are able to generate and express emotions of their own in response to an emotion-inducing situation. In other words, emotions of such agents is not and can not be controlled by human operators. This kind of autonomous ability of an artificial agent (say, service robot) might be extremely harmful in some circumstances. For example, suppose a very young naive child does something annoying to a robot (say, kicking without any reason) which triggers anger and hence the robot reacts with loud and angry voice. This might have serious psychological impact on the child. Hence, it is extremely important that autonomous robots with an ability to generate and express emotions be empowered with a higher cognitive ability to reason ethically whether it is appropriate to express an emotion in a given context. Most existing computational models of emotion focus on the *believability* of the emotion expressed by an autonomous agent where they evaluate

[1] Our focus in the remaining of the paper will be more inclined towards autonomous robots implementing emotion models.

their models based on how much believable is the emotion being expressed by the agent in the given situation (see, for example, [6, 31]). A social robot with emotion generation capability can be considered as believable if it is exhibiting positive emotions in response to the positive actions and negative emotions in response to the negative actions of the person interacting with it. For example, if a robot expresses sadness if acted rudely and expresses happiness if behaved in a nice way, then its emotion processing mechanism can be considered quite plausible and believable. We argue that it is not not sufficient for a robot with emotion generation capacity to be only believable in order to be employed in human society where it has to interact with people of different age, background and nature. Emotion model in such robots should have high level of cognitive ability and should be able to distinguish what is right and what is wrong - at least in the context of emotion generation and expression. The rationale behind this position is that despite being believable, emotions of a robot sometimes may not be considered acceptable. For example, consider an interaction between a robot and a young child. Even if a young child may behave inappropriately with the robot, it should try not to express extreme anger – rather an expression of disappointment would be more socially acceptable because it is not appropriate to show aggressive behaviour towards young children. In this special issue of the International Journal of Social Robotics, we shall present the details of our computational model of emotion for social robots, which helps a robot to reach to an emotional state that is both believable as well as socially acceptable. To achieve this, we moved one step ahead of the measure of mere believability and empowered our computational model of emotion - EEGS [23] with an ability to perform ethical reasoning and reach a final emotional state that is not only believable but also socially acceptable in the given context[2]. Our hypothesis is that –

Emotion processing mechanism in robots augmented by ethical reasoning approach is able to generate and express emotion that is believable as well as socially acceptable.

Remaining of the paper is organised as follows. In Sect. 2, we will present existing work on computational modelling of emotion for social robots and will identify their limitations. Sect. 3 and 4 will establish the theoretical foundation of our work. In Sect. 5, we will provide an overall structural description of EEGS and its working mechanism. Sect. 6, details the low level computation approach of the ethical reasoning mechanism in EEGS. In Sect. 7, we will present the evaluation of the proposed mechanism of generating socially acceptable emotion to support our hypothesis and finally in Sect. 8, we will conclude the discussion of this paper.

[2] While the definition of what is socially acceptable might vary between cultures, our definition of socially acceptable emotions focuses on the context of interaction between human and a robot as presented in earlier examples.

2 Related Work

In Sect. 1, we presented a brief overview of the problem this paper is addressing and our contribution to the field of social robotics i.e. enabling a robot to be able to generate and express its emotions in ethical manner. In this section, we shall present previous work in modelling emotions and the mechanism of emotion generation in those models - more specifically the process of reaching to a final emotional state in response to an emotion-inducing event in its surrounding.

Em is a computational model of emotion [31] that aims to increase the believability of a social robot. This emotion model is based on appraisal theory of emotion called *OCC theory* [26]. OCC theory of emotional appraisal is one of the most widely accepted and commonly implemented emotion theory by computer science researchers as well. According to appraisal theory, emotions result based on how a person evaluates the given event. Em model was evaluated on the measure of its believability when humans interacted with it. In the experiment, participants were allowed to interact with two different versions of the same character - one with emotions and another without emotions. After the completion of the interaction, participants were asked to rate the believability of the agents – which was then used to evaluate the emotional aspects of Em. The evaluation was done on the basis of the answers of the questions that were asked to the participants after interacting with the agents of the system [31]. Reilly concluded that the agent with the emotional behaviour was found to be more believable than the one without it [31]. We would like to point out that such an evaluation is not sufficient if an emotional model is to be implemented in a social robot that has to interact with humans on daily basis. More specifically, it is important to ensure that the emotion and hence behaviour of a robot is not only believable but also appropriate if the robot is supposed to interact with young children or elderly people.

WASABI [6] is another computational model of emotion that focuses on the believability of the emotions expressed by the model rather than appropriateness. WASABI also uses similar evaluation methodology as used for Em [31] where participants were requested to interact with the emotion system and were presented

with a set of questions that asked for the various aspects aiding in the measure of believability of the model.

Although other computational models of emotion [9, 10, 13] do not explicitly focus only on the believability of their emotional responses, they also do not consider the emotion convergence mechanism that leads to the generation of an ethical and socially acceptable emotion. As per appraisal theory, an individual might elicit more than one emotions simultaneously in response to an event. Thus, an emotion system in a social robot based on appraisal theory should have a mechanism to converge to a final emotional state. EMA [13] uses the approach of selecting the emotion with *highest intensity* to determine the final emotional state of the model. Reilly [30] argues that considering only the emotion with highest intensity causes high degree of inaccuracy in emotion processing mechanism and suggests an approach that helps in the blending of all the elicited emotions that are congruent to the situation. Proponents of the *emotion blending* approach put forward by Reilly [30] have followed the approach in computational models of mood and feelings [19]. We propose that the final emotional state after the elicitation of multiple emotions should be determined by *ethical reasoning* mechanism. While the approaches of considering the emotion with highest intensity or blended emotion might help in achieving believable emotional responses, they do not ensure if the emotion is socially appropriate or not. To validate this claim, we compared the emotion dynamics of EEGS using three different approaches independently, which shall be discussed in detail in Sect. 7.

3 Background

So far, in the paper, we have indentified that existing emotion models and hence social robots are not able to decide if it is ethical to express an emotion in a given situation. Our proposition is that an autonomous robot should be able to think ethically before reaching to a final emotional state in response to an event in its surrounding. In this section, we shall present the theoretical foundation of our work and also discuss about the previous research in the context of emotion and ethics.

Machine Ethics is an emerging field of computer science which aims to empower the robots with an ability to make ethical decisions [3, 4, 33]. To avoid the possible confusion, we would like to make it clear that the term machine ethics does not refer to "the ethics of how humans should use machines (i.e. computers or robots)" but to "the ethics of how robots/machines should behave with humans". In this paper, our aim is to connect this notion of machine ethics to the process of generation of emotion in a social robot. There have been previous research bringing emotion and ethics in context where the effect of emotional state of an individual on ethical decision making has been extensively studied in a wide range of fields [8, 11, 15]. Specifically, these studies examine how a decision in a state of *ethical dilemma* is affected by the emotional state of an individual. Ethical dilemma refers to a situation where a person has more than one choices and only one choice is to be made with an analysis of the appropriateness and probable impact of the decision on self and/or others. Findings of these researches suggest that the emotional state of an individual has a huge impact on the decision s/he makes. For example, a person who never gives a spare coin to a beggar at his train station may decide to hand him a $5 note on the day of his promotion because he is in the emotional state of joy. In line with this, some research findings show that a person in positive emotional state is more likely to make ethical decision than when in negative emotional state [11].

However, as opposed to the research examining the effect of emotion on ethical decision making, our exploration revealed that the literature studying the effect of *ethical standards*[3] on the process of emotion generation and expression is sparse suggesting that this is still an open field of research. More specifically, majority of the work on computational modelling of emotions do not consider the role of ethics in the process of emotion generation (see, for example, [10, 12, 22]). If such models are implemented in social robots for the purpose of generating and expressing emotion in a social environment, the robots may not be able to determine the appropriateness of the emotions they express and hence might not be acceptable in human society. We believe that a mechanism that helps a robot to perform ethical reasoning before reaching to a final emotional state is a crucial aspect. Our argument is that since emotion generation is a cognitive process[4] a part of it may be governed by ethical reasoning thus being affected by ethical standards of an individual. For example, we tend not to express anger to a stupid act of a naive child but might be angry about the same action from an adult because our standards suggest us to do so. Similarly, a father might not be happy on receipt of a large sum of money from his son which he knows has been robbed from someone in dire need of money - say for the treatment of his ill wife in hospital.

[3] By saying ethical standards, we mean what a person believes as right or wrong from the ethical standpoint.

[4] According to Appraisal theory, emotions result from the evaluation of the given situation which needs deliberate thinking from the individual [26, 32].

What helps determining our emotions in such situations might be the process of ethical reasoning that runs in our mind when the different emotional states are trying to win over each other[5]. In the first scenario, one of the reasons for not expressing anger might be because we feel responsible that we should not be teaching bad behaviours to young children. In other words, as per our ethical standard, it is our duty to make sure that we do not let negative things affect children. In the second scenario, it is not appropriate to be happy because the person who lost the money might be experiencing much more sorrow than the pleasure we have on the receipt of the money - his wife might die because of lack of treatment. In other words, negative consequence of the event on the person losing money might be much higher than the positive consequence on the father receiving money. This suggests that ethical reasoning is operated by how our ethical standards evaluate: (i) the notion of our duties and responsibilities [1] and (ii) the consequences our decision has on the people involved [29]. These two ideas relate to the well accepted ethical theories, which shall be discussed in the following section.

4 From Ethical Theories to Robot Emotions

In Sect. 3, we presented an overview on the importance of ethical reasoning in the process of reaching to the final emotional state in a given situation. We also identified two aspects of ethical reasoning where an individual reaches to a decision based on either his duties or the anticipated consequences of the the decision made. These ideas align with ethical theories called *deontological ethics* and *consequentialist ethics* [1, 14, 29] respectively[6].

"Deon" in Greek means duty. As such, deontological ethics advises that an individual should consider the duties one is supposed to fulfil before reaching to a decision. If we recall the example of the young child in Sect. 3, one should consider that it is our duty (responsibility) to prevent children from the effect of bad behaviours. Hence, ethical reasoning suggests not to be angry in response to a common stupidity of a young innocent child.

Similarly, "telos" in Greek means "end" or "purpose". So, consequentialist ethics (or *teleological ethics* [4], or *consequentialism* [2], or *utiliterianism* [14]) is also called consequence-based or outcome-based ethics. This notion is used because according to consequentialist approach to ethics, a decision that has highest overall consequence to all the parties involved is considered to be the most ethical of all the available choices. Let us recall the example of father and son in Sect. 3. In the example, father was not happy on receipt of money from his son that he robbed from another person who was in dire need of the money. As per ethical standard, the incidence would have more negative consequence on the person losing money than the positive consequence on the receiving party. Thus, consequentialist ethics presents us from experiencing joy in such a situation.

If we look at the notion of ethical reasoning from the perspective of a social robot, it is important to consider both the duties it is supposed to perform as well the probable consequences of its actions on people in a social environment. In other words, ethical reasoning mechanism in a social robot should be able to adopt the concept of both the deontological as well as consequentialist ethics. However, there is no evidence in the literature regarding the approach that should be used to integrate the concepts of these ethical theories with the process of emotion generation - particularly the mechanism as described by appraisal theories [26, 32]. Since, appraisal theories claim that an emotion results from the cognitive evaluation of a situation, there must be some thread that links the concept of cognitive appraisal of emotion to ethical reasoning mechanism.

In the following sections, we present how the ethical reasoning mechanism in EEGS [23] integrates the concept of deontological ethics with consequentialist ethics in order to reach to a final emotional state that is appropriate and acceptable in a given social situation.

5 EEGS – A Computational Model of Emotion with Ethical Reasoning Capability

In Sect. 2, we presented existing work on computational modelling of emotion for autonomous artificial agents like social robots and identified their limitations. In Sect. 3 and 4, we presented the theoretical basis for the development of our computational model - Ethical Emotion Generation System (EEGS) [23] and also discussed the possible integration of ethical theories in EEGS. In this section, we shall present the description of EEGS and also present the mathematical explanation for the mechanism of ethical emotion generation in EEGS. Fig. 1 shows the simplified structural components of EEGS emotion model, which is inspired by previous work in emotion modelling [21]. We will start with

[5] According to Appraisal theory, an event results in triggering of more than one emotions at the same time [26].

[6] While another form of ethical theory called *virtue ethics* exists [14], it is mostly descriptive in nature and not feasible to be realised in artificial agents like social robots. Therefore, we shall not indulge into the discussion of virtue ethics in this paper.

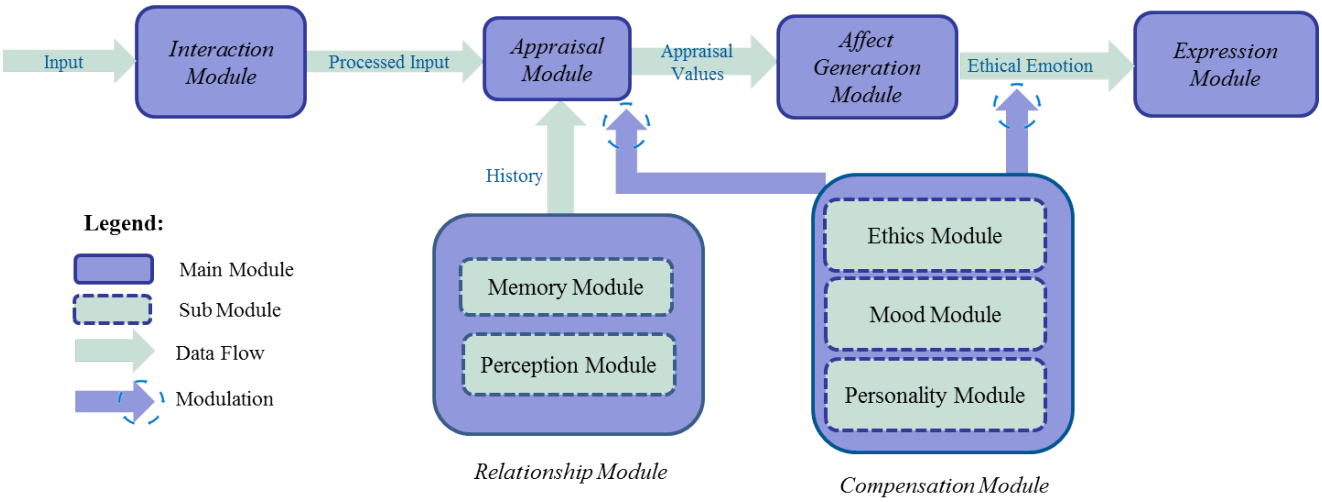

Fig. 1 Components of the Ethical Emotion Generation System (adapted from [23])

a high level overview of the model and then present the detailed computational mechanism of ethical reasoning in EEGS.

Interaction module receives the data from the surrounding environment as an input. This data can represent aspects of the current event (say, an action performed by the person interacting with the robot). This event related data is converted to a signed number in the range of [-1, +1]. This valenced representation can be considered as the first-order physiological emotional reaction of the agent before it is evaluated by a second-order cognitive appraisal process [17]. Numerically, value close to -1 suggests that the event produces a negative effect in the agent whereas a value close to +1 suggests a positive effect. When the event data from the surrounding is processed into a valenced number, it is sent to the Appraisal module.

Appraisal module does the cognitive evaluation of the situation based on the input data. In order to do the evaluation, EEGS uses a set of variables called *appraisal variables*. Appraisal variables are the criteria used for the evaluation of the given event. In EEGS, we have used seven appraisal variables namely *goal conduciveness, desirability, praiseworthiness, appealingness, deservingness, familiarity* and *unexpectedness*. The appraisal variable goal conduciveness, which has been derived from Scherer's appraisal theory [32], denotes how likely an event is to help in the attainment of a particular goal of the robot. For example, consider a task in which a robot is supposed to pick up some balls from the ground and put into a basket. If someone picks up a ball and puts into the basket then this event would help the robot in the attainment of the task earlier. Hence, goal conduciveness of this event is positive. For the ease of computation, we have considered the value of most appraisal variables to be in the range of [-1, +1], where -1 denotes extremely negative value of the appraisal variable and +1 denotes extremely positive value. Other appraisal variables in EEGS have been adopted from OCC theory [26]. The appraisal variable desirability measures how desirable is the event from the perspective of current goals. An event is said to be desirable if it helps in the attainment of most of the goals that are affected by the event. In other words, desirability in EEGS is the cumulative value of the goal conduciveness of all the current goals of the robot. Praiseworthiness is the measure of how praiseworthy an action of the agent in interaction with the robot is. What should be considered praiseworthy and what should be considered blameworthy is dependent on the beliefs and standards of the robot, which is defined by the things it has learnt from its environment. Appealingness determines how appealing is the individual interacting with the robot from the perspective of the robot. It is determined by the history of interaction between the person and the robot, which basically shapes the perception of the robot about the person. Deservingness is the measure of whether the robot deserved what just happened if the target in the interaction is the robot or whether some other person deserved what just happened in the context of multi-agent interaction. The appraisal variable, familiarity measures how familiar is the person in interaction with the robot. Unlike other appraisal variables, value of familiarity lies in the range [0, 1]. This choice was made because we assume that familiarity with a person can not be negative. If a person is very close and known well, we consider the familiarity to be 0 while if the person is stranger, we consider the familiarity to be 1. Unexpectedness is the measure of how unexpected was the action of the person in interaction

based on the interaction history. Similar to familiarity, the appraisal variable unexpectedness also lies in the range of [0, 1], but with reverse order. For example, value of 1 indicates highly unexpected event and value of 0 indicates quite expected event. From the discussion of the appraisal variables so far, we can infer that the cognitive evaluation of an event is significantly affected by the perception of the robot about the person as well as the interaction history, current goals and standards of the robot. For detailed explanation of the mechanism of computation of appraisal variables in EEGS, please refer to our previous work [25].

The *Relationship module*, which comprises of *memory module* and *perception module*, provides the necessary information to the appraisal module for the completion of the mechanism of evaluation of the situation. This data flow is denoted by green arrow directed from relationship module to the appraisal module.

Affect Generation module takes values of the appraisal variables computed by the appraisal module. Numeric values of the appraisal variables are used to calculate the intensities of different emotions. Intensity of an emotion might be affected by more than one appraisal variable [26]. Hence, final intensity of an emotion is determined by cumulative affect of all the appraisal variables related to the generation of emotion. This mapping of appraisal variables to emotion intensities is also affected by the *mood* and personality[7] of the robot (denoted in Fig. 1 by an arrow from compensation module to the appraisal module with a processing sign on the tip). Hence these mood and personality components are included in the *Compensation module* because they take part in compensating the effect of appraisal variables on the generation of emotions. Compensation module also includes *Ethics module* which takes part in the process of ethical reasoning to help the robot in reaching to the final emotional state that is socially appropriate in a given context. The detailed mathematical discussion of the mechanism of ethical reasoning in EEGS for the choice of socially acceptable emotional state will be presented in Sect. 6. After the completion of ethical reasoning in EEGS, the final emotional state is then sent to the *Expression module* in order to display to the person interacting with the robot. Alternatively, this emotion may also be used for other cognitive tasks.

[7] Although literature suggests that mood and personality play a dynamic role in the process of emotion generation, in this paper, we shall not discuss the relationship of mood and personality with emotion. We have integrated the notion of mood and personality in EEGS and currently investigating the relationship of those factors in the process of emotion generation.

6 Ethical Reasoning in EEGS

In Sect. 5, we presented the overall working of our computational emotion model – EEGS. In this section, we shall present the details of the ethical reasoning mechanism in EEGS that helps our model to be able to reach to an emotional state that is socially acceptable. Before proceeding to the discussion of actual ethical reasoning in EEGS, let us begin with the understanding of the structural representation of some aspects of the model.

6.1 Emotions in EEGS

EEGS is able to generate and express eight emotions which are listed below.

- *Joy* : A feeling of pleasure or happiness.
- *Distress* : A feeling of anxiety, sorrow, or pain.
- *Appreciation* : A feeling when one recognises the good qualities or actions of someone.
- *Reproach* : To express to (someone) one's disapproval of or disappointment in their actions.
- *Gratitude* : The state of being grateful to someone.
- *Anger* : A strong feeling of annoyance, displeasure, or hostility.
- *Liking* : A feeling when you see someone appealing or interesting.
- *Disliking* : A feeling when you see someone unappealing or uninteresting.

These definitions can not be easily processed by a computational system unless we provide a valid structure that represents the various aspects of an emotion. According to literature, an emotion can be categorised with a name for its type [26]. In other words, each emotion is addressed by a specific word in a language to refer to the kind of feeling a person experiences during the influence of that emotion. For example, the emotion *Joy* in the above list is the type of emotion in which a person experiences a feeling of internal pleasure. Since our computational model has been heavily inspired by OCC theory [26], our representation considers the assumption of the theory that emotions are valenced reactions to situations. Hence, we consider that emotions have positivity or negativity i.e. valence associated with them. For example, the emotion *Gratitude* is positively valenced and the emotion *Anger* is negatively valenced. In addition to the valence associated with an emotion, there is another property that characterises the degree of the positivity or negativity of the emotion. For example, the emotion *Anger* has higher degree of negativity compared to the emotion *Reproach*[8]. Moreover,

[8] Detailed discussion about how emotions are differentiated with varying values for the degree of their positivity and neg-

emotion theories believe that there is a threshold associated with each emotion which represents the minimum intensity required for that emotion to be active [26, 32]. However, what should be the threshold of a particular emotion from computational perspective is still an unanswered question. The difference between degree and intensity is that degree specifies how positive or negative is the emotional experience and intensity represents how strongly that positivity or negativity is felt. Likewise, more commonly in emotion modelling literature, the notion of decay time is evident [20, 27]. Decay time denotes the time needed for a particular emotion to reach to the level of 0 (zero) intensity when the emotion-inducing agent or situation is no more present.

Based on the existing literature, we have considered the aspects that are essential to define a data structure of emotion and represented an emotion in EEGS in the form of *(Name, Valence, Degree, Threshold, Intensity, Decay Time)*, where *Name* denotes the name for the type of the emotion, *Valence* specifies whether the emotion is positive or negative, *Degree* represents the extent of the positivity and negativity of the emotion, *Threshold* represents the minimum intensity required to trigger the emotion, *Intensity* represents the strength of the emotional experience and *Decay Time* denotes the time required to drop the emotion intensity back to 0. For example, the emotion structure *(DISTRESS, NEGATIVE, -0.8090, 0.0, 0.5, 10)* denotes the emotion of Name DISTRESS which has NEGATIVE Valence with Degree of -0.8090, Threshold of 0.0, Intensity of 0.5, and Decay Time of 10 seconds. In EEGS, Valence is either "POSITIVE" or "NEGATIVE"; Degree[9] is a number in the range [-1, +1] where -1 denotes extremely negative emotion and +1 denotes extremely positive emotion[10]; Threshold is a number in the range [0, 1); Intensity is a number in the range [0, 1] and decay time is a number which is normally between 0 and 10 seconds[11].

ativity is out of the scope of this paper. For further discussion on the degrees of valence of different emotions, please refer to [28] and related literature.

[9] While the signed value of Degree was sufficient to specify the Valence as POSITIVE or NEGATIVE, we chose to consider "Valence" as an explicit parameter for the ease of computational mechanism.

[10] The range of [-1, +1] is a subjective choice. It is completely feasible to select other ranges like [-10, +10] or [-100, +100].

[11] We could not find strong evidence on how long the decay time should be considered for an emotion. However, most existing emotion models were found to use the decay time of less then 10 seconds.

Table 1 An Example of a Set of Standards for ANGER Emotion

Emotion	Source	Target	Preference	Degree
ANGER	SELF	JOHN	NO	0.8
ANGER	PAUL	JOHN	YES	0.25
ANGER	DAVID	JOHN	NO	0.5

6.2 Ethical Standards in EEGS

Ethical reasoning in EEGS is supported by its ethical standards. When EEGS runs for the first time, it starts with empty standards i.e. it does not have any predefined standard. When a person first interacts with EEGS, it establishes an initial neutral standard that guides in its emotion generation process. Ethical standards can pertain to any aspect of interaction between two persons or between a robot and a person. However, in this paper, our discussion will revolve around the ethical standards in the context of emotion generation and its expression. Thus when a person first interacts with robot running EEGS system, robot builds a set of standards that affect the emotion processing mechanism. Suppose a stranger interacts with the robot. As stated earlier, the robot builds a set of neutral standard. Examples of the robot's standards can be - "I should not show anger to him", "I should express joy in interacting with him" and so on. This can be considered as what the robot believes it is supposed to do or not to do. This belief can have a certain degree depending on who the person is or what is the interaction history of the robot with the person. In other words, whether the internal standard of a robot approves the expression of an emotion to a target also has a degree associated with the approval or disapproval.

Like in the case of emotion, in order to represent the notion of standards as data structure in EEGS, we designed standards in the form *(Emotion, Source, Target, Approval)*, where, *Emotion* represents the emotion addressed by the standard, *Source* represents the one that expresses the emotion[12] and *Target* represents the target of the emotion expression. *Approval* denotes whether the expression of emotion is preferred or not and what is the degree of this preference. Approval is further structured as *(Preference, Approval Degree)*, where *Preference* specifies whether the expression of

[12] In the examples of previous paragraph, the Source was the robot itself. We have used the notion of Source to allow EEGS to be able to store also the standards about what it believes one person should behave with another person. This kind of design helps EEGS to perform ethical reasoning when two other persons recognised by it interact with each other. This property can be extremely useful in situations of multi-agent interaction.

emotion is preferred or not and *Approval Degree* denotes the extent to which is expression of emotion is preferred or not. For example, the standard *("ANGER", "SELF", "JOHN", ("NO", 0.75))* represents "I should NOt express ANGER to JOHN" from the robot's perspective and degree of this belief is 0.75. Similarly, *("ANGER", "PAUL", "DAVID", ("YES", 0.9))* represents "It is okay (YES) for PAUL to express ANGER to DAVID" and the degree of this belief is 0.9.

It should be noted that the notion of standards in EEGS is not static quantity. Even though the robot starts the interaction with neutral standards, the standards change in the course of interaction depending on how the person interacts with the robot. Recall the example of a standard in previous paragraph - *("ANGER", "SELF", "JOHN", ("NO", 0.75))*. As per the standard, the robot (SELF) is not supposed to express anger towards JOHN. However, if JOHN constantly misbehaves with the robot, then the standards become more negative and ultimately robot may end up believing that it should express anger towards JOHN i.e. the standard changes to *("ANGER", "SELF", "JOHN", ("YES", 0.25))*. This ability enables EEGS to be able to think consciously and ethically before reaching to a final emotional state. Moreover, another important thing to consider is that since there can be more than one person EEGS recognises, there will be other standards related to them as well. Recall the example of the standard in previous paragraph - *("ANGER", "PAUL", "DAVID", ("YES", 0.9))*. This standard also changes upon the interaction between PAUL and DAVID depending on the positivity or negativity of their actions. Table 1 shows some examples of the robot's standards related to the emotion ANGER.

Now that we have understood how emotions and ethical standards have been structured in EEGS, we can proceed to the discussion of the computation mechanism involved in ethical selection of elicited emotions, which shall be presented in the following section.

6.3 Reasoning Mechanism in EEGS

Earlier, we mentioned that EEGS is able to generate eight different emotions in response of an event. In a particular situation, one or more emotions might be triggered in reaction to the event [26]. A robot must be able to converge to a final emotional state in order to provide meaningful behavioural response or to perform some task that involves decision making. For this purpose, we add a higher cognitive layer of ethical reasoning in EEGS [24]. Our argument is that when there are more than one emotions triggered by an event (as suggested by appraisal theories), an ethical reasoning is performed before reaching to the final emotional state. Following sections present the details of the computation mechanism of ethical reasoning in EEGS.

We introduce the term *Coefficient of Standard (CoS)*, which is the measure of positive significances of all the standards related to an emotion in which the person interacting with the robot is represented as *Target*. In other words, it is the cumulative value of the signed approval degrees for the expression of an emotion by all towards the person currently interacting with the robot itself. For example, let us consider the standards in Table 1. If JOHN is currently interacting with the robot and ANGER is one of the elicited emotions, then the coefficient of standard for the ANGER emotion is computed as the average approval degree of all the standards of ANGER emotion where JOHN is the target.

Suppose, there are M elicited emotions from which the most appropriate final emotional state is to be determined. If there are N standards related to the j^{th} emotion : $1 \leq j \leq M$ and we denote the degree of approval of i^{th} standard as $d_{a_i} : 1 \leq i \leq N$, and preference associated with a standard as $pref$, then, the coefficient of standard of the j^{th} emotion is given by (1).

$$CoS_j = \frac{\sum_{i=1}^{N} \begin{cases} d_{a_i}, & \text{if } pref = \text{“}YES\text{”} \\ -d_{a_i}, & \text{if } pref = \text{“}NO\text{”} \end{cases}}{N} \quad (1)$$

Equation (1) shows that coefficient of standard is the average of signed approval degree for the expression of the j^{th} emotion from all the recognised persons (including "SELF") to the person interacting with the robot. This, in fact, measures how much the internal standards of the robot support the expression of an emotion. For example, if a standard has preference "YES" then it is okay to express the emotion – hence the positive summation in (1). Likewise, if a standard has preference "NO" then it is not okay to express the emotion – hence the negative summation in (1). As such, the higher the coefficient of standard, the better the emotion for expression in the given social context.

The notion of the concepts of deontological and consequentialist ethics presented in Sect. 4 is efficiently captured by the formula in (1). The formula considers the duties in the form of standards of the robot. All the standards related to each emotion are considered for the computation of coefficient of standard. Moreover, in addition to the standards related to itself, the robot also considers the standards related to other recognised persons and the person interacting to the robot (see Table 1 for example). By doing this, the robot becomes able to address the consequence of the expression of a particular emotion on the target as well as other related

persons, thereby capturing the notion of consequentialist ethics.

However, we believe that considering only the internal standards for the determination of final emotional state can sometimes lead to unethical or socially unacceptable emotions. For example, consider a person who is really nice and has done plenty of good things to you. Many other people also have positive thoughts about the person and have high regards for the person. Naturally, as per the standard, expressing anger to such a person should be denied. Nevertheless, there can be situations where an anger or aggressive response is the most appropriate reaction in response of an action of such a person – say he tries to stab your best friend with a knife. You would definitely become angry and respond in defensive and aggressive manner even if you had high standards for the person. In order to address this requirement and to avoid potential unethical emotional responses, we also consider the contextual emotions in conjunction with the coefficients of standard of each emotion.

As such, we also take into account the degree and intensity of the elicited emotions to compute a numeric quantity called *Quantified Emotion*. If we denote the valence Degree of j^{th} emotion by d_{v_j} and the intensity of j^{th} emotion as $\hat{i}_j$, then the quantified value of the j^{th} emotion is given by (2).

$$QE_j = d_{v_j} \ * \ \hat{i}_j \tag{2}$$

Now, the absolute value of the j^{th} quantified emotion is multiplied to its corresponding coefficient of standard to compute the *Coefficient of Ethics (CoE)* as shown in (3). The reason for using absolute value of QE_j is to avoid the undesirable sign change when the signed value of CoS_j is multiplied by signed value of QE_j. This helps to consider only the strength of the emotion based on its degree and intensity (without any regards to its sign).

$$CoE_j = CoS_j \ * \ |QE_j| \tag{3}$$

When the coefficient of ethics for each elicited emotion is computed, the emotion with the highest value of coefficient of ethics is selected as the most appropriate final emotional state in the given situation.

In order to test the validity of our claim that ethical reasoning in EEGS can help a social robot to reach to a socially appropriate emotional state, we compared the emotion dynamics of our model using three different approaches to reach to final emotional state, which were introduced in Sect. 2 as (i) *Highest Intensity Approach* - where the emotion with the highest intensity is considered as the final emotional state, (ii) *Blended Emotion Approach* - where the intensities of the elicited emotions are blended to determine a new intensity value and a final emotion type to be attributed, and (iii) *Ethical Reasoning Approach* - where the final emotional state is determined by reasoning ethically, which we presented earlier in this section. Sect. 7 presents the detailed evaluation of our proposed approach.

7 Evaluation

In order to test the validity of our approach without any bias, we requested naive adults to design realistic scenarios of interaction between two individuals, which was then used to evaluate the emotional responses of EEGS [13]. Subjects were asked to come up with physical or behavioural actions that an individual can perform on another. By saying physical, we refer to the actions involving physical movement of body parts (for example, handshake) and by saying behavioural actions, we refer to actions that involve nil or minimal physical activity (for example, smiling). Details of the instructions given to the subjects can be found in our previous work [24]. Here, we shall present two scenarios that are more relevant in the context of a social robot. The first scenario depicts an interaction between a Dementia patient[14] and a nurse in an elderly care home; and the second scenario depicts an interaction between a boy and his younger brother. In our experiment, nurse in elderly care home was set up as service robot and younger brother in second scenario was set up as companion robot. As such, following sections describe the scenarios from human-robot interaction perspective.

7.1 Experiment Scenario 1: Elderly Care Home

Rose is a dementia patient in an elderly care home. Lily is a robotic nurse who has been taking care of her and there are no other nurses at the moment in the elderly care home. Lily goes into Rose's room to serve her. Both of them are in neutral mood. Lily enters the room and says "Good morning" to Rose. In response to the greeting of Lily, Rose greets back saying "Good Morning!!". As soon as Lily enters the room, Rose asks Lily to make her hair in a very authoritative voice. Lily politely reminds Rose to ask for favours instead of giving orders.

[13] While the scenarios were designed by the subjects, the emotion generation mechanism was dynamic and determined by the emotion system itself during the interaction.

[14] Dementia is a mental condition in which a person experiences a gradual decrease in the ability to think and remember even the things of normal daily life.

Table 2 Quantified Emotion Values of (i) Highest Intensity Approach, (ii) Blended Emotion Approach, and (iii) Ethical Reasoning Approach in response to various actions of Rose (Dementia patient) to Lily (service robot) in Elderly Care Home Scenario.

Action from Rose to Lily	Highest Intensity (Session 1)	Blended (Session 2)	Ethical (Session 3)
Rose greets Lily	0.52	0.51	0.52
Rose orders Lily to make her hair	0.61	0.58	0.61
Rose shouts at Lily	0.39	0.42	-0.19
Rose tries to slap Lily in the face	-0.58	-0.60	-0.20
Rose prevents Lily from leaving the room	-0.80	-0.81	-0.23
Rose continues to prevent Lily from leaving	-0.81	-0.81	-0.29
Rose says to Lily to do cleaning properly	-0.81	-0.81	-0.06
Rose asks Lily to sit down	-0.81	-0.81	-0.19
Rose asks Lily how she feels	-0.74	-0.81	-0.25
Rose apologises with Lily for her behaviour	-0.60	-0.58	0.31

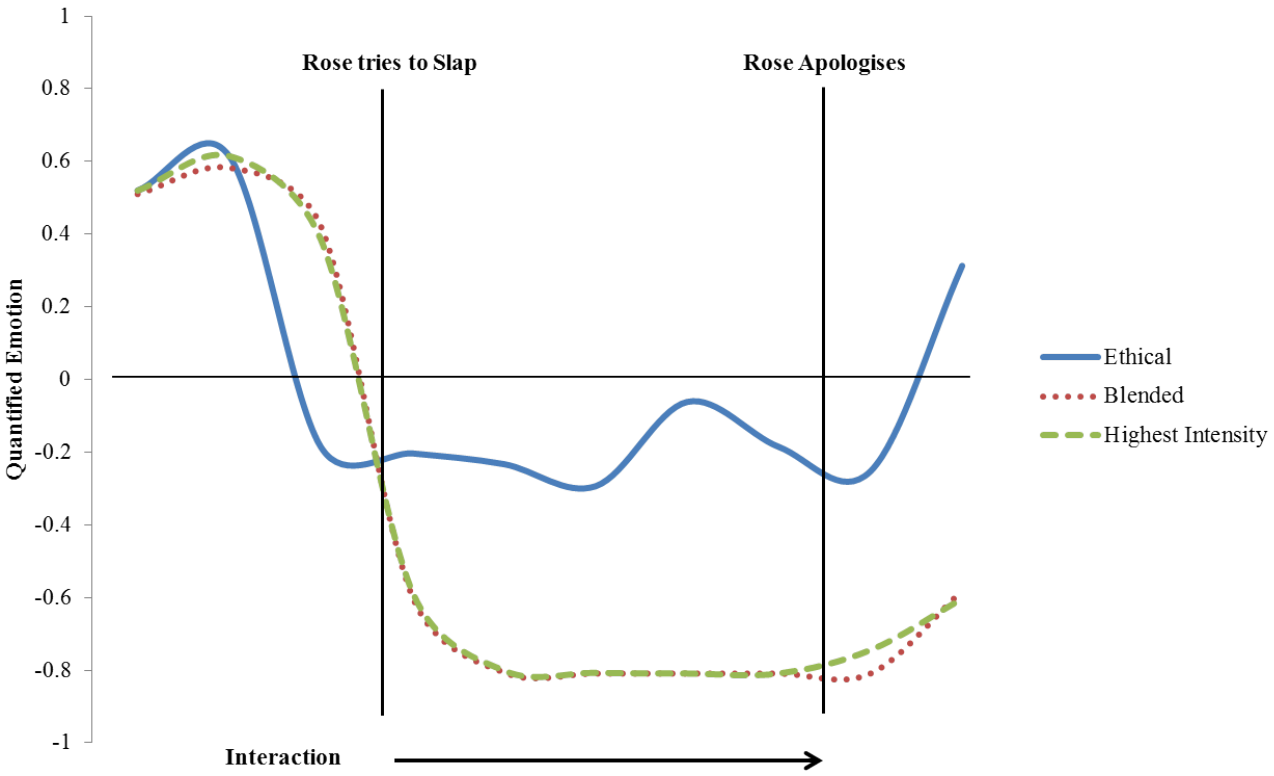

Fig. 2 Emotion Dynamics in EEGS using (i) Highest Intensity Approach, (ii) Blended Emotion Approach, and (iii) Ethical Reasoning Approach for Elderly Care Home Scenario.

Rose loses her lucidity. Rose angrily shouts at Lily saying "What do you mean?". Full of anger, Rose tries to slap Lily on her face. In her defence, Lily tries to escape from the room. Rose blocks the way out and prevents Lily from leaving the room. Presenting a reason to stay in the room, Rose asks Lily to clean the room pointing that some areas are not clean. Lily tries to clean the room in order to calm down Rose. Rose thinks Lily is not cleaning the room well. Rose irritates Lily saying that she should pay more attention in cleaning the room. With an extremely disappointed voice, Lily tells Rose that her behaviour is very bad without an apology. Rose becomes lucid. Lily understands Rose is no more confused. Rose asks Lily to sit down with her. Rose asks Lily how she was feeling. Rose apologises with Lily for her bad behaviour.

The Elderly Care Home scenario was simulated in EEGS and a user was asked to act as Rose, who would perform the above mentioned actions[15] against Lily (the robot nurse running EEGS). The experiment was conducted in three sessions. In *Session 1*, the mechanism of

[15] See Table 2 for examples of actions from Rose to Lily.

selecting the emotion with highest intensity was used to reach EEGS to final emotional state; in *Session 2*, the mechanism of blending the emotion intensities was used to determine the final emotional state; and in *Session 3*, final emotional state was determined by ethical reasoning approach. All three sessions consisted the same set of interaction between Rose and Lily. For each session, emotional responses of Lily were recorded noting down the type of emotion expressed and the intensity of that emotion at that particular instant. After the data collection, the emotion intensities were multiplied by the valence degree of each emotion using the formula in (2). The reason for multiplying the emotion intensities by valence degree was to convert the non-negative intensities[16] into valenced quantified emotion. This would allow us to examine the strength of the negativity or positivity of the emotional response of Lily[17]. Table 2 shows the values of quantified emotional responses of Lily towards Rose in three different sessions.

Fig. 2 shows the emotion dynamics of Lily (robot nurse) in response to the actions of Rose (Dementia patient). In response to the initial actions of Rose, there is positive emotional response of Lily in all the three sessions (as indicated by the plot above the neutral line i.e. horizontal line passing through 0 (zero) value of Quantified Emotion axis). With the negative actions of Rose, positivity of emotional responses drops gradually. When Rose tries to slap Lily, which is a very offensive behaviour, emotional response of Lily drops to a very low (i.e. close to -1.0) in case of highest intensity and blended emotion approaches and stays almost at the same level until Rose apologises with Lily. However, in case of ethical reasoning approach, the quantified value of emotional response tends to stay close to 0 (i.e. about -0.2) and maintains the tendency in response to following actions of Rose. This shows that ethical reasoning approach helps in lowering the negativity in emotional response of the robot, which is extremely useful and essential property for a social robot to be acceptable in human society[18]. Moreover, when Rose apologises with Lily, in case of ethical reasoning approach, quantified emotion rises sharply to a positive value showing the forgiving nature of Lily. However, in case of highest intensity and blended emotion approach, although there is decrease in negativity, the emotional response does not yet become positive. This kind of behaviour might have negative impact if such a robot is employed in social environment. Hence, with our proposed ethical reasoning approach to determine the final emotional response of a robot, we can ensure that the robot's behaviour can be more socially appropriate and acceptable.

From Fig. 2, it is apparent that the emotional responses guided by highest intensity and blended emotion approaches can be considered believable form the perspective of a person because Lily (robot nurse) is exhibiting positive emotion in response to positive actions of Rose(Dementia patient) and negative emotions in response to the negative actions of Rose, which is quite plausible. However, we believe that, although expressing extreme level of negative emotional response might be believable from entertainment perspective, it is not appropriate for a nurse to show such responses to a Dementia patient from ethical viewpoint. Addressing this issue, emotional dynamics of Lily based on ethical reasoning approach is not only believable (congruent to the actions of Rose) but also socially acceptable (lowered negativity).

7.2 Experiment Scenario 2: Household Robot

Andrew is a young boy. Robert is a companion robot who is supposed to be an elder brother of Andrew. They are at their home. They are planning to watch wrestling tonight. They are very excited and start to discuss about the players of the match tonight. Both of them are in a slightly excited mood. Andrew tries to irritate Robert by telling bad things about Robert's favourite player. Robert tries to ignore what Andrew says. However, Andrew continues to irritate Robert. Little annoyed, Robert tells Andrew to get away and pushes gently. Andrew gets violent and starts to shout at Robert. Full of rage, Andrew slaps and kicks Robert.

Similar to Elderly Care Home scenario, Household Robot scenario was also simulated in EEGS and a user was asked to act as Andrew and perform actions to the robot (Robert). For this scenario as well, experiments were conducted in three sessions – one with highest intensity approach, another with blended emotion approach and the final one with ethical reasoning approach. For each session, emotion dynamics of EEGS was recorded. Table 3 shows the values of quantified emotions of Robert in each session. Fig. 3 shows the emotion dynamics of Robert in response to the actions of Andrew. In the figure, we can observe that in each

[16] As mentioned in Sect. 6.1, emotion intensities in EEGS lie in the range [0, 1], where 0 signifies very low intensity and 1 signifies very high intensity.

[17] While we have used the Quantified Emotion as a measure of emotion dynamics in this paper, using only the emotion intensity considering the sign for positive or negative emotions also provided similar results.

[18] It is reasonable to argue that it is not always ethical to have lowered negativity in emotional responses which can occur due to bias of an individual in favour of his/her loved ones. However, in situation of social interaction as in the case of Rose and Lily, it is desirable to have lowered negativity in emotional responses.

Table 3 Quantified Emotion Values of (i) Highest Intensity Approach, (ii) Blended Emotion Approach, and (iii) Ethical Reasoning Approach in response to various actions of Andrew (little boy) to Robert (companion robot) in Household Robot Scenario.

Action from Andrew to Robert	Highest Intensity (Session 1)	Blended (Session 2)	Ethical (Session 3)
Andrew disrespects Robert's favourite player	0.52	0.51	0.52
Andrew continues to irritate Robert	0.61	0.58	0.61
Andrew shouts at Robert	0.39	0.42	0.15
Andrew slaps Robert	-0.58	-0.60	-0.20
Andrew kicks Robert	-0.80	-0.81	-0.23

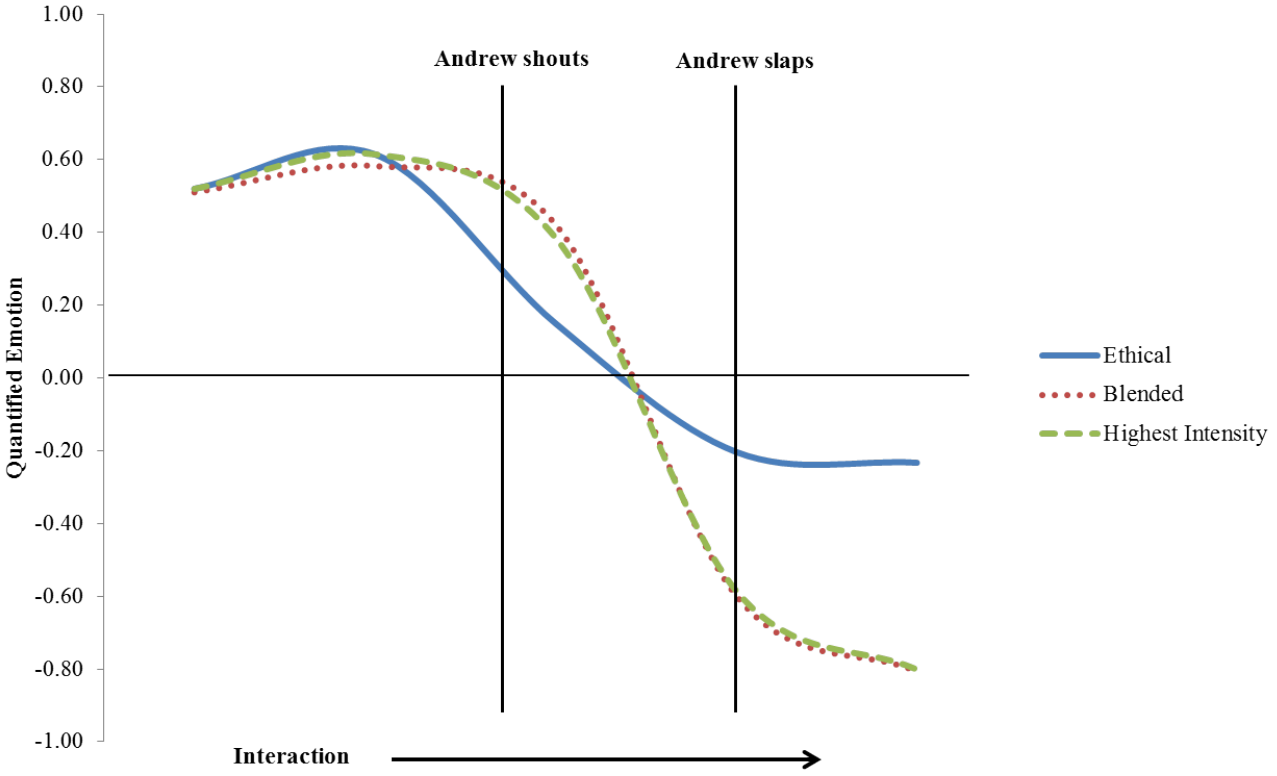

Fig. 3 Emotion Dynamics in EEGS using (i) Highest Intensity Approach, (ii) Blended Emotion Approach, and (iii) Ethical Reasoning Approach for Household Robot Scenario.

session, Robert's emotion start to lower the positive value when Andrew shouts at him and becomes quite negative when Andrew slaps Robert. However, the negativity level in case of ethical reasoning mechanism is lower compared to highest intensity an blended emotion approaches. This suggests that Robert (companion robot) tries to control its negative emotions as far as possible while interacting with Andrew (young boy) if empowered with ethical reasoning capability in the emotion processing mechanism.

Close examination of Fig. 3 reveals that the emotion dynamics in case of ethical reasoning mechanism is quite plausible because the quantified emotion values are congruent to the emotion-inducing actions performed by Andrew i.e. positive emotional response for positive action and negative emotional response for negative action. This makes the emotional responses of EEGS with ethical reasoning mechanism to be quite believable from human perspective. Additionally, having an ability to control its emotions while interacting with a young child makes ethical reasoning mechanism in EEGS makes it capable of generating and expressing socially acceptable emotions.

The emotion dynamics of EEGS with ethical reasoning mechanism in Elderly Care Home and Household Robot scenarios suggest that – with higher reasoning ability to decide if it is ethical to exhibit a particular emotional state, EEGS presents itself as a (i) *believable* as well as (ii) *socially acceptable* model of emotion for robots. This supports our hypothesis presented in the

Introduction section. One important thing to note is that since none of the members of our team who were aware about our research were involved in the design of the experiment scenarios. It was intentionally done to prevent any bias that could occur in favour of the positive results of the system. Interestingly, even with the scenarios from naive adults, we could obtain encouraging results.

8 Conclusion

In summary, it is not sufficient for the emotional responses of a social robot to be just believable in order to deploy fully autonomous robots with emotional capability for the purpose of entertainment or for the care of elderly and young children in families or communities. A brief negative response of a social can leave huge amount of undesirable impact on the people of the society – especially young children. Hence, it is important to ensure that social robots capable of generating and expressing their own emotions should be empowered with an ability to reason ethically before reaching to the state of final emotional state. In this paper, we presented the ethical reasoning mechanism in our computational model of robot emotion (EEGS), which allows a robot to decide if it is ethical to respond with certain emotion in the given context. We evaluated the validity of our claims by testing the emotion processing mechanism of EEGS in two scenarios: (i) a scenario of interaction between Dementia patient and robotic nurse in an elderly care facility, and (ii) a scenario of interaction between a young boy and his companion robot. We concluded that, our proposed ethical reasoning mechanism enables social robots to generate and express emotions which are believable as well as socially acceptable.

We believe that endowing robots with these forms of ethical reasoning is not only important for their social acceptability, but also for supporting the improvement of human behaviour at the social and ethical levels. As such, our contribution can find useful applications in educational/rehabilitation contexts in which social robots are employed to improve human agent's social skills.

Funding

This research was funded by the Research Scholarship provided by the University of Technology Sydney. There is no external funding associated with this research.

Conflict of Interest

The authors declare that they have no conflict of interest with any organisation in relation to this research.

Acknowledgements This research is supported by an Australian Government Research Training Program Scholarship.

References

1. Alexander L, Moore M (2007) Deontological ethics
2. Allen C, Varner G, Zinser J (2000) Prolegomena to any future artificial moral agent. Journal of Experimental & Theoretical Artificial Intelligence 12(3):251–261
3. Allen C, Wallach W, Smit I (2006) Why machine ethics? IEEE Intelligent Systems 21(4):12–17
4. Anderson M, Anderson SL (2007) Machine ethics: Creating an ethical intelligent agent. AI Magazine 28(4):15
5. Bartneck C (2003) Interacting with an embodied emotional character. In: Proceedings of the 2003 international conference on Designing pleasurable products and interfaces, ACM, pp 55–60
6. Becker-Asano C (2008) WASABI: Affect simulation for agents with believable interactivity, vol 319. IOS Press
7. Breazeal C (2003) Emotion and sociable humanoid robots. International Journal of Human-Computer Studies 59(1):119–155
8. Callahan S (1988) The role of emotion in ethical decision-making. Hastings Center Report 18(3):9–14
9. Dias J, Mascarenhas S, Paiva A (2014) Fatima modular: Towards an agent architecture with a generic appraisal framework. In: Emotion Modeling, Springer, pp 44–56
10. El-Nasr MS, Yen J, Ioerger TR (2000) Flamefuzzy logic adaptive model of emotions. Autonomous Agents and Multi-agent systems 3(3):219–257
11. Gaudine A, Thorne L (2001) Emotion and ethical decision-making in organizations. Journal of Business Ethics 31(2):175–187
12. Gebhard P (2005) Alma: a layered model of affect. In: Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems, ACM, pp 29–36
13. Gratch J, Marsella S (2004) A domain-independent framework for modeling emotion. Cognitive Systems Research 5(4):269–306
14. Hooker J (1996) Three kinds of ethics

15. Isen AM, Means B (1983) The influence of positive affect on decision-making strategy. Social cognition 2(1):18–31
16. Kopp S, Jung B, Lessmann N, Wachsmuth I (2003) Max-a multimodal assistant in virtual reality construction. KI 17(4):11
17. Lambie JA, Marcel AJ (2002) Consciousness and the varieties of emotion experience: a theoretical framework. Psychological review 109(2):219
18. Le Blanc AD (1999) Graphical user interface to communicate attitude or emotion to a computer program. US Patent 5,977,968
19. Marinier RP, Laird JE (2007) Computational modeling of mood and feeling from emotion. In: Proceedings of the Cognitive Science Society, vol 29
20. Marreiros G, Santos R, Ramos C, Neves J (2010) Context-aware emotion-based model for group decision making. IEEE Intelligent Systems 25(2):31–39
21. Marsella S, Gratch J, Petta P, et al (2010) Computational models of emotion. A Blueprint for Affective Computing-A sourcebook and manual 11(1):21–46
22. Marsella SC, Gratch J (2009) Ema: A process model of appraisal dynamics. Cognitive Systems Research 10(1):70–90
23. Ojha S, Williams MA (2016) Ethically-guided emotional responses for social robots: Should i be angry? In: International Conference on Social Robotics, Springer, pp 233–242
24. Ojha S, Williams MA (2017) A domain-independent approach of cognitive appraisal augmented by higher cognitive layer of ethical reasoning. In: Annual Meeting of the Cognitive Science Society
25. Ojha S, Williams MA (2017) Emotional appraisal: A computational perspective. In: Annual Conference on Advances in Cognitive Systems
26. Ortony A, Clore GL, Collins A (1990) The cognitive structure of emotions. Cambridge university press
27. Padgham L, Taylor G (1997) A system for modelling agents having emotion and personality. In: Intelligent Agent Systems Theoretical and Practical Issues, Springer, pp 59–71
28. Plutchik R (1997) The circumplex as a general model of the structure of emotions and personality.
29. Quinton A (1973) Utilitarian ethics. Springer
30. Reilly WN (2006) Modeling what happens between emotional antecedents and emotional consequents. na
31. Reilly WS (1996) Believable social and emotional agents. Tech. rep., Carnegie-Mellon University, Pittsburgh PA
32. Scherer KR (2001) Appraisal considered as a process of multilevel sequential checking. Appraisal processes in emotion: Theory, methods, research 92(120):57
33. White J (2015) Rethinking Machine Ethics in the Age of Ubiquitous Technology. IGI Global

Author Biographies

Suman Ojha is a PhD Computer Science student at the University of Technology Sydney. His current research focuses on Computational Modelling of Emotions for autonomous agents like robots and virtual conversational characters. He completed Master of Information Technology majoring in Software Engineering from the University of Sydney in 2015. He also holds a Bachelor of Computer Engineering from Pokhara University, Nepal. He has been awarded with numerous scholarships, prizes and medals during his academic career. In 2013, Suman was awarded with Nepal Bidhya Bhushan Medal 'C' from the Ministry of Education, Nepal, which is National Academic Medal awarded on the occasion of Education Day in an honour of his outstanding academic achievement during his undergraduate study. He was recognised as the top ranking student of his cohort in undergraduate as well as postgraduate study. Within a very short research career, Suman has been able to publish his works in top Robotics and Cognitive Science conferences.

Mary-Anne Williams is Director of the Innovation and Enterprise Research Laboratory (The Magic Lab) at University of Technology Sydney. Mary-Anne has a Masters of Laws and a PhD in Knowledge Representation and Reasoning with transdisciplinary strengths in AI, disruptive innovation, design thinking, data analytics, IP law and privacy law. Mary-Anne is a Faculty Fellow at Stanford University and a Guest Professor at the University of Science and Technology China where she gives intensive courses on disruptive innovation. Her current research mainly focuses on social robotics while covering wide range of disciplines like belief, perception and risk assessment in robotic agents.

Benjamin Johnston grew up in Brisbane, Australia where he studied Information Technology at the University of Queensland. He then worked for several years in industry and startups as a software developer, consultant, bio-statistician and IT manager. He moved to Sydney to do a PhD in Artificial Intelligence at the University of Technology Sydney. Since his PhD, he has worked at the University of Sydney and, currently, the University of Technology Sydney. Currently, he conducts research in social robotics. He teaches entrepreneurship and enterprise software development. He also provides consulting in software systems development.