Rational Agents That Blush

Paolo Turrini[1], John-Jules Ch. Meyer[2], and Cristiano Castelfranchi[3]

[1] University of Siena, Italy
[2] University of Utrecht, The Netherlands
[3] ISTC-CNR, Italy

> "Video meliora proboque; deteriora sequor"
> (Ovidius, Metam. VIII, 18-21)

1 Introduction

A student, supported by his classmates, throws a piece of chalk at the teacher who is writing on the blackboard. The teacher rapidly turns back and promptly catches him in the act. The student blushes and suddenly realizes how bad it was what he did.

What happened to the guy? Why did he decide to throw the piece of chalk? And why, after a few seconds, he would have liked that what he did had never happened? Aim of this work is to provide a formal characterization of those emotions that deal with normative reasoning, such as shame and sense of guilt, to understand their relation with rational action and to ground their formalization on a cognitive science perspective. In order to do this we need to identify *when* agents feel ashamed or guilty and *what* agents do when they feel so. We will also investigate how agents can induce and silence these feelings in themselves, i.e. the analysis of defensive strategies they can employ. We will argue that agents do have control over their emotions and we will analyze some operations they can carry out on them. After presenting a cognitive model of shame and guilt as social emotions, we will provide a formal representation of them and their dynamics in terms of basic notions such as beliefs, goals and violations, following the rational action approach of [26] [25] [13] and the cognitive approach in [16] [3].

Related Work. As witnessed by [6] the study of emotions has recently gained much attention in the fields of artificial intelligence [21] [13], evolutionary computation [22] and multi agent systems [19], due to the encounter between computer science tools and neuro, cognitive and social sciences analyses [5] [18] [9]. Ours is a cognitive perspective: even though we agree that it is important to study emotions from a computational and emergentist point of view, we argue that in

A. Paiva, R. Prada, and R.W. Picard (Eds.): ACII 2007, LNCS 4738, pp. 316–327, 2007.

order to build an anatomy of emotions it is as important to understand them in terms of their interaction with other cognitive ingredients[1].

The most influential cognitive paradigm for studying and constructing cognitive agents with emotions has been that by Ortony, Clore and Collins [18]. Nevertheless, in [18] the characterizations of feelings related to norms are not deeply investigated:

> "In order to feel shame one must have violated a standard one takes to be important, as moral standards are. Such violations are held to be inexcusable. This is not necessary for a person who is feeling guilty.(...) In fact, we do not think that there is a distinct emotion of feeling guilty. Rather, we view feelings of guilt as mixtures of distinct emotions such as shame and regret, perhaps accompanied by certain cognitive states, such as the belief that one was, at least technically, responsible." (p. 142-143)

Many expressions here would need to be explicated further: why are violations only in case of shame held to be inexcusable? What is a mixture of emotions? And a technical responsibility? If we find the distinction between shame and guilt and all the other related feelings as meaningful at all, we need to have clear-cut definitions that relate those feelings to agents' mental states and to precisely understand their functioning.

We will pursue a formal investigation on emotions, as done for instance in [19], but adding a closer look to the formal properties of our notions, that we construct in a well known logical framework such as KARO [26] [25] [13]. From a cognitive point of view we will follow the analysis of [16] that grounds emotional displays like blushing and feelings like loneliness or pride on complex multiagent interaction. We claim that such model overcomes the oversimplifications in [18] while keeping a semiformal approach that eases a proper formal investigation.

2 A Cognitive Model for Shame and Sense of Guilt

Shame and sense of guilt are seen as social emotions. They are social because they: are socially acquired (through values and norms internalization); have social targets (the victim, for instance, in case of sense of guilt) and referents (the

[1] Many criticisms have been moved to cognitivists theories, some of which can be hardly addressed in this context. Nevertheless we would like to point out how several ones are based on what we think is a misconception of the use of formal models of cognition. In the study of normative emotions of [22] it is argued that "logic does not provide an adequate foundation" to the study of human behaviour and "the necessary abandonment of logical models for the explanation and simulation of human social behaviour" is advocated. Even though we share the worries in [22] w.r.t. representing humans as perfect reasoners, we claim that an anti-logical position in modelling interaction is simply wrong: emotions can be studied as mechanisms that act on human cognition. But mechanisms do have a logic. What is more, the recent breakthroughs of logical models in the study of social interaction and information flow [23] have shown that formal semantics can lead to the construction of rigorous models of complex phenomena such as emotions (as in [13]).

holders of a value); achieve social functions (like regulation)[2] Miceli and Castelfranchi [16] emphasize how the perceived causal responsibility (see also [4]), that is the belief of having had the capacity to avoid a damage or a violation, is a crucial notion for distinguishing these feelings: “when ashamed, one sees oneself as incompetent or inadequate with respect to some goal; when guilty, one sees oneself as endowed with negative power.” ([16],295)[3].

Internalization. When agents use values or norms as input for their decisions to comply with what prescribed, we say that these values or norms have been internalized. An internalized norm, thus, drives agents’ behaviour towards obedience. Psychological research has shown that internalization can be induced, by *significant others* expectations [8] and commands [10]; the importance of implementing the role of significant others for judgment formation is therefore already acknowledged as crucial in pedagogical activity and interaction design [11]. In this article the dynamics of internalization will be taken into account w.r.t. the way changing significant others influences the set of world states agents label as good and acceptable, and what the consequences of these are in terms of feeling guilty and ashamed.

Normativity comes from outside. Shame and sense of guilt are associated with violation of normative standards. As emphasized in [3] agents may not fulfill others’ expectations, and, even worse, they can go against others’ values. Agents are continuously judged in this respect. But, following [8], they do not fear such evaluation from all others, but only from a particular subset of them, those ones on which they are dependent, they care about, and whose standards they have the goal to meet: the significant others.

It goes without saying that agents that have not internalized at all a value cannot feel guilty or ashamed w.r.t. that value. Therefore the situation in which only significant others share a norm that is not internalized is a severe conditioning factor for agents’ emotional reaction.

We may observe, due to difference in personality traits, a pride reaction in some agents, while instead others may show feelings of loneliness and exclusion [17]. On the other side, others not sharing a value or a norm for which the agents has perception of negative power does not necessarily cancel its perception of

[2] The paper takes a cognitive approach which is quite different from the believability approach (see for instance [11]). We believe that the appropriate *affective interaction* should be based on mental contents and interpretations, not just on a superficial reaction. Complex agents, such as human beings, react to the ascribed mind. For example, to a shameful face we can respond in many different ways, depending on our interpretation of the grounding mental content. “Yes I think that what you said/did is ridiculous/ugly!”; “Why do you care about their judgment?! They have stupid prejudices”; “This feature of yours is very special and nice; you should be proud of it!”; “You are wrong, they appreciate you the way you are”. etc.

[3] In Maria Miceli’s words shame is the perception of oneself being a dull knife, whereas feeling guilty is the perception of oneself being a sharp knife [17].

wrongdoing. Instead more personal feelings can come up, like sense of guilt and shame towards the self. The interplay is sketched in Table 1.

When agents realize that some significant others do not share fundamental values or norm their being significant is rapidly questioned. We believe that the different opinions that an agent and its significant others have concerning a relevant value can be associated with phenomena like cognitive dissonance and the processes that lead to its resolution [10].

Table 1. Multi Agent Perspective

	Others Don't Share		Others Share	
	Internalized Norm / Value	Not Internalized Norm / Value	Internalized Norm / Value	Not Internalized Norm / Value
Negative Power	Sense of Guilt towards the self	Indifference	Sense of Guilt	Pride / Loneliness
Inadequacy	Shame towards the self	Indifference	Shame	Pride / Loneliness

2.1 Controlling Emotions

The most interesting intuition that the cognitive approach in [16] [3] [2] gives is that human beings do have some control of their emotions. “In particular, people can react to their own emotions defending themselves from the disturbing or blamed ones. They try to repress, deny, and manipulate them.” [2].

We are going to push this intuition further by looking at both *control via belief manipulation* and *control via goal manipulation.* The first type of control will act on an agent belief base in order to eliminate those beliefs that induce in the self the emotion (for instance, justifications), while the second will act with the same purpose on the agent motivational base (for instance, apologies). But how would these strategies work? Elaborating on Miceli and Castelfranchi’s proposal in [16], the different types of reaction can be related to basic psychological types of agents, such as high self esteem (HSE) and low self esteem (LSE) ones. HSE people might try to question the basis on the grounds of which they are blamed or they feel bad; they will react actively by trying to produce justificatory beliefs for their actions. Instead LSE agents will tend to apologise and find excuses, as they are more unlikely to question the values which they are accused to go against. In this sense HSE agents will tend to react with pride to the above mentioned situation, while LSE agents will react with for instance feeling lonely or rejected. We distinguish a typical LSE reaction, “I did not know”, that we classify as an excuse by claiming ignorance of the relevant effects of a dangerous action that has been carried out, from a typical HSE reaction, “It was not that bad”, that is a justification on a presumed violation. It claims that what others may think as wrong is not really so. We argue that these mechanisms can be formally modeled. For lack of space, we leave the formal treatment of offensive moves, precisely analyzed in [14], to future work.

2.2 Feeling Ashamed

Shame is not necessarily a moral emotion, because the needs not to be, e.g., poor, ugly, fat, bald are not moral, although they involve a normative standard. As argued in [2] feeling ashamed always involves a believed negative self evaluation (concerning one's inadequacy) related to somebody whose judgment agents care about. As far as the behaviour is concerned, the goal of an ashamed person is to reduce exposure [3]. This may be translated into various actions, either minimizing the importance of a value (agents that are proud of something), which is an HSE agent's typical reaction, or minimizing their active contribution to a damage by declaring submission and imperfection, which are typical LSE agents' reactions. Emotional displays of shame are admission of imperfection, that instantiate for instance with blushing. Blushing is not a confession of guilt [2] but it still has a precise communicative function [3]. We can be caught doing something that looks bad w.r.t. others (even though we know they are wrong) and yet blush.

2.3 Feeling Guilty

The feeling of guilt is usually linked to the conviction of having actively injured someone or broken some moral imperative or norm [2]. It is associated to the evaluation of having negative power against a given value, i.e. a perception of responsibility. Looking at agent types reactions and beliefs of responsibility, it is intuitively clear how difficult it is to both feel ashamed and guilty at the same time for the very same thing. Either agents believe to be a dull knife or a sharp knife. Nevertheless the complex transitions between the two feelings are common in everyday life, by means of a belief revision [26] concerning responsibility .

As far as reactive behaviour is concerned, one first goal of agents that feel guilty is that of reparation, which triggers the agent to care about the damaged person, and to expiate, to pay in some sense for what has been done. A very interesting property of agents that feel guilty is to regret doing something. But what is regret? It is the desire (very easy to frustrate) of not having done something we actually did. Regret acts as a situation marker [5]: not only do we feel guilty about a situation but we also associate to that situation a further punishment, due to the desire frustration, in order to recognize the preconditions that caused us to feel guilty[4].

3 A Language for Rational Agents with Emotions

Emotional Multiagent Karo. The reconstruction of cognitive agents that feel guilty and ashamed needs some further concepts with respect to those already in [13]. We need to reason about agents' values, and their mismatch with others' perceived ones; operations that act directly on the agent emotional state, i.e. allow changing perception of situations in order not to feel bad; finally,

[4] This is an interesting link with the notion of sadness as formalized in [13].

in order to reason about multiagent interaction we need to extend dynamic deontic logic to the multiagent case. The works in [12] [20] and [7] provide a solid basis for addressing this issue. We will introduce in such a framework a set of violation constants V_i indexed with agents, that will label worlds that are bad for particular agents, as well as different dynamics for different types of agents.

3.1 The Language

Action Expression. The actions that agents perform can have different form. We classically consider a set of atomic actions from which all the others can be obtained compositionally. For easing notation, we will not consider parallel execution and nondeterministic choice for complex actions and events, while action negation (i.e. refraining) will be limited for technical reasons ([12]) to the atomic case. Finally the planning component in emotions requires the treatment of the notion of action composition. The set of action expressions *Act* is the smallest set containing all actions of the following form:

$$\alpha ::= b|skip|\mathbf{kick}(x,\phi)|\mathbf{welcome}(x,\phi)|\mathbf{replace}(x,y,\phi)|\alpha_1;\alpha_2|\alpha_1^n$$

where $b = a|\overline{a}$ is an atomic action or its negation; $\phi \in \Pi_0$, where Π_0 is the set of atomic propositions; $x, y \in Agt$, which is the set of agents; *kick*, *welcome*, *replace* actions will be used for updating evaluations and will be dealt as special actions later on in the paper. The set of events *Evt* has the following grammar:

$$\xi ::= X : \alpha|\xi_1;\xi_2|\xi_1^n$$

To ease reading, we skip the technical construction of action interpretation and we send the reader to [12] for a thorough account of the single agent case, which inspired in turn the multiagent extension contained in [20] and ours.

As soon as the reader encounters expressions of the form $x[\![\xi]\!]_R y$, where x, y are couples model-world, we ask him or her to attach to them the informal reading of "state y is an effect of the execution of event ξ in state x".

For convenience, we sometimes view $[\![\xi]\!]_R$ as a functional relation.

Syntax. Our language is given by the following syntax:

$p(p \in \Pi_0)|V_i(i \in Agt)|\neg\phi|\phi \wedge \psi|\mathbf{Aut}_{i,j}\phi|\mathbf{B}_i\phi|\mathbf{D}_i\phi|\mathbf{P}\phi|[\xi]\phi|\mathbf{A}_i\alpha|\mathbf{Com}_i(\alpha)|$
$DONE_X(\alpha)$[5]

We moreover use the following abbreviations: $\phi \vee \psi := \neg(\neg\phi \wedge \neg\psi)$; $\phi \rightarrow \psi := \neg\phi \vee \psi$; $\phi \leftrightarrow \psi := (\phi \rightarrow \psi) \wedge (\psi \rightarrow \phi)$; $\langle\xi\rangle\phi := \neg[\xi]\neg\phi$; $\mathbf{I}_i(\alpha,\phi) := \mathbf{B}_i(\neg\phi \wedge \mathbf{D}_i\phi \wedge \langle\alpha\rangle\phi \wedge \mathbf{A}_i\alpha)$, the last with the informal reading of "possible intention".

[5] The informal reading of modalities is "i forbids j to be in ϕ", "i believes that ϕ is true", "i desires that ϕ is true", "ϕ was true" (as in [1]) , "after ξ, ϕ becomes true", "i is able to do α", "i is committed to α", "X did α".

Structure. The structures interpreting the language $\mathcal{L}$ are given by a class of E-KAROUS[6] models in which each model M is of the following shape:

$$M =< Agt, W, Act, \sigma, \{B_i|i \in Agt\}, \{D_i|i \in Agt\}, Aut >$$

In which Agt is a finite nonempty set of agents. We take $Agt = Agt_h + Agt_l$. It is partitioned into Agt_h which will represent the HSE agents and Agt_l, the LSE agents. $Agt_h \cap Agt_l = \emptyset$; W is a finite nonempty set of worlds; Act is the set of basic actions; $\sigma : \Pi_0 \cup \{V_i|i \in Agt\} \rightarrow 2^W$ is the augmented valuation function, that assigns each atom to a set of worlds, with the intended meaning that they are those worlds in which the atom is true. $\{B_i|i \in Agt\}$ is an epistemic accessibility relation; Each $B_i \subseteq W \times W$ is composed by couples $\{w, w'\}$ in such a way that the world w' represents an epistemic alternative for agent i at world w. We indicate with $[w]_{B_i}$ the set of epistemic alternatives for agent i at world w. $\{D_i|i \in Agt\}$ is defined as B_i for desired worlds. Aut is a function g such that $g : Agt \times Agt \times (\mathcal{M} \times W) \rightarrow 2^{L_0}$. This function associates to each agent a set of agents that point to a group of propositions. The idea is that such agents indicate which situations are to be considered bad by the agent. As the function ranges over a powerset we can have more agents that point to a proposition. We say that those agents block the proposition. We now label as $Sig_{(i,M,w)} = \{j|\exists\phi s.t.\phi \in Aut(i, j, (\langle M, w\rangle))\}$ the set of significant others for agent i at the situation $\langle M, w\rangle$. This set comprises those agents that block at least a proposition in a given situation for agent i. Moreover we impose an order $<$ that links couples model-world with events, in such a way to induces a tree on transitions that guarantees linear past and branching future[7].

Semantics. The formulas of our language $\mathcal{L}$ are interpreted as follows:

- $M, w \models p$ iff $w \in \sigma(p)$;
- Propositional cases are usual;
- $M, w \models V_i$ iff $w \in \sigma(V_i)$;
- $M, w \models \mathbf{Aut}_{i,j}\phi$ iff $\phi \in Aut(i, j, (\langle M, w\rangle))$;
- $M, w \models \mathbf{B}_i\phi$ iff $M, w' \models \phi$ for all w' s.t. wB_iw';
- $M, w \models \mathbf{D}_i\phi$ iff $M, w' \models \phi$ for all w' s.t. wD_iw';
- $M, w \models \mathbf{P}\phi$ iff $\exists(\langle M', w'\rangle)$ s.t $(\langle M', w'\rangle) < (\langle M, w\rangle)$ and $M', w' \models \phi$;
- $M, w \models [\xi]\phi$ iff $M', w' \models \phi$ for all $(\langle M', w'\rangle)$ s.t. $(\langle M, w\rangle)[\![\xi]\!]_R(\langle M', w'\rangle)$;
- $M, w \models \mathbf{A}_i\alpha$ iff $\alpha \in c(\langle M, w\rangle)(i)$; $M, w \models \mathbf{Com}_i(\alpha)$ iff $\alpha \in Ag(\langle M, w\rangle)(i)$[8]
- $M, w \models \mathbf{Sig}_{j,i}$ iff $j \in Sig_{(i,M,w)}$; $M, w \models DONE_X(\alpha)$ iff $\exists(\langle M', w'\rangle)$ s.t. $(\langle M', w'\rangle) < (\langle M, w\rangle)\, and\, (\langle M', w'\rangle)[\![X : \alpha]\!]_R(\langle M, w\rangle)$.

[6] E-KAROUS is a fancy transformation of KARO to an emotional multiagent shape, resembling moreover the name of a person that did not pay much attention to the suggestions of the friends.

[7] We point to [1] for a formal characterization of such constraint.

[8] The behaviour of Ag and c functions is described in details in [24]. The first updates the agent's agenda concerning commitments, the second returns the set of actions the agent has the internal ability to carry out.

Constraints on the models. We denote with $[\![\phi]\!]_M$ the set A such that $A = \{w|M, w \models \phi\}$. We constrain our models in the following way:

- For all $w \in W, [w]_{B_i} \neq \emptyset$; $w' \in [w]_{B_i} \Rightarrow [w']_{B_i} = [w]_{B_i}$
- $[\![\mathbf{Sig}_{i,j}]\!]_M \subseteq [\![\mathbf{Bel}_j\mathbf{Sig}_{i,j}]\!]_M$; $W \backslash [\![\mathbf{Sig}_{i,j}]\!]_M \subseteq W \backslash [\![\mathbf{Bel}_j\mathbf{Sig}_{i,j}]\!]_M$
- $[\![V_i]\!]_M \subseteq [\![\mathbf{Bel}_i V_i]\!]_M$; $W \backslash [\![V_i]\!]_M \subseteq W \backslash [\![\mathbf{Bel}_i V_i]\!]_M$

Proposition 1. *The following propositions are valid in E-KAROUS models:*

- $\models \boldsymbol{B}_i\top$*;* $\models \neg\boldsymbol{B}_i\phi \rightarrow \boldsymbol{B}_i\neg\boldsymbol{B}_i\phi$*;* $\models \boldsymbol{B}_i\phi \rightarrow \boldsymbol{B}_i\boldsymbol{B}_i\phi$*;*
- $\models \boldsymbol{Sig}_{i,j} \rightarrow \boldsymbol{B}_j\boldsymbol{Sig}_{i,j}$*;* $\models \neg\boldsymbol{Sig}_{i,j} \rightarrow \boldsymbol{B}_j\neg\boldsymbol{Sig}_{i,j}$
- $\models V_i \rightarrow \boldsymbol{B}_i V_i$*;* $\models \neg V_i \rightarrow \boldsymbol{B}_i\neg V_i$

The first three entries are standard for $S4$ models of beliefs [1], forbidding logical inconsistency and allowing positive and negative introspection. The fourth and fifth add positive and negative introspection for significant others. It makes sense to claim that if agents have some agent as significant other then they believe so, and vice versa for those that are not significant others. The last two items state that the positive and negative perception of the own valuations is valid.

3.2 Changing Friends

In [26] non standard actions such as those that induce mind changing are described. In the same fashion we would like to describe those actions that update the authority relations among agents. In particular agents should be able to resolve their cognitive dissonance by eliminating significant others or welcoming new ones.

We describe functions the transition function $[[]]_R$ for actions *welcome*, *kick*, *replace* leaving the treatment of the capability function c ([24]) that tells us when agents have the internal ability to perform these actions, to future work. These are special actions that transform the models in a peculiar way. The first updates the set of relevant others by adding a new agent. Violation states are updated as specified. The second deletes an agent from such set. The third first deletes some agents and after adds new ones to the set.

Definition 1. *For some E-KAROUS model*
$M =< Agt, W, Act, \sigma, \{B_i|i \in Agt\}, \{D_i|i \in Agt\}, Aut >$ *with* $w \in W$ *and* $\phi, \psi \in L_0$ *be given. We define: All* $\langle M', w'\rangle \in [\![i : \boldsymbol{welcome}(\phi, j)]\!]_R(\langle M, w\rangle)$ *are such that:*
$M' =< Agt, W, Act, \sigma', \{B_i|i \in Agt\}, \{D_i|i \in Agt\}, Aut' >$ *with*

- $\models Aut'(i', j', (\langle M', w'\rangle)) = Aut(i', j', (\langle M, w\rangle))$ *if* $i \neq i'$ *or* $j \neq j'$ *or* $w' \neq w$*;* $\models Aut'(i, j, (\langle M', w'\rangle)) = \{\phi\}$ *;* $\models \sigma'(V_i) \cap [\![\phi]\!]_{M'} = \emptyset$*;* $\models \sigma'(\psi) = \sigma(\psi)$ *for* $\psi \neq V_j$

Definition 2. $\langle M', w'\rangle \in [\![i : \boldsymbol{kick}(\phi, j)]\!]_R(\langle M, w\rangle)$ *are such that:*
$M' =< Agt, W, Act, \sigma', \{B_i|i \in Agt\}, \{D_i|i \in Agt\}, Aut' >$ *with*

- $\models Aut'(i', j', (\langle M', w' \rangle)) = Aut(i', j', (\langle M, w \rangle))$ *if* $i \neq i'$ *or* $j \neq j'$ *or* $w' \neq w$; $\models Aut'(i, j, (\langle M', w' \rangle)) = \{\emptyset\}$; $\models \sigma'(V_i) = \sigma(V_i) \backslash \sigma(V_k)$; $\models \sigma'(\psi) = \sigma(\psi)$ *for* $\psi \neq V_j$

Definition 3. $[\![i : \boldsymbol{replace}(\phi, j, k)]\!]_R (\langle M, w \rangle) = [\![(i : \boldsymbol{kick}(\phi, j); \boldsymbol{welcome}(\phi, k)]\!]_R (\langle M, w \rangle)$.

Proposition 2. *The following propositions are valid in E-KAROUS models:*

- $\models \boldsymbol{Sig}_{ik} \leftrightarrow [i : \boldsymbol{welcome}(\phi, j)] \boldsymbol{Sig}_{ik}$;$\models [i : \boldsymbol{welcome}(\phi, j)] \boldsymbol{Sig}_{ij}$
- $\models [i : \boldsymbol{welcome}(\phi, j)](\phi \leftrightarrow \neg V_i)$; $\models \psi \rightarrow [i : \boldsymbol{welcome}(\phi, j)]\psi$
- $\models \boldsymbol{Sig}_{ik} \leftrightarrow [i : \boldsymbol{kick}(\phi, j)] \boldsymbol{Sig}_{ik}$; $\models [i : \boldsymbol{kick}(\phi, j)] \neg \boldsymbol{Sig}_{ij}$
- $\models [i : \boldsymbol{kick}(\phi, j)](\phi \leftrightarrow V_i)$; $\models \psi \rightarrow [i : \boldsymbol{kick}(\phi, j)]\psi$
- $\models \boldsymbol{Sig}_{ij} \rightarrow [i : \boldsymbol{replace}(\phi, j, k)] \boldsymbol{Sig}_{ik}$; $\models [i : \boldsymbol{replace}(\phi, j, k)] \neg \boldsymbol{Sig}_{ij}$
- $\models [i : \boldsymbol{replace}(\phi, j, k)](\phi \leftrightarrow V_i)$; $\models \psi \rightarrow [i : \boldsymbol{replace}(\phi, j, k)]\psi$

The first four items deal with the welcoming operation. The first of them says that welcoming a new agent does not affect the perception of others; the second simply that welcoming causes an agent to be a significant other; the third that the reason of welcome is not seen as bad by the agent; the fourth that the evaluation of the other propositions do not change. The kicking operations behaves dually, while replacing can be obtained by composing the other two.

4 The Dynamics of Guilt and Shame

4.1 Sense of Guilt

An agent feels guilty when it observes that what he actively caused was violation for some of his significant others[9]. This means that the agent has the intention to do π for achieving goal ϕ but he believes the actual world state he actively chose (that he could avoid) satisfies a violation condition for some agent j which happens to be a significant other.

For the heavy use of refraining actions, we limit the treatment to atomic cases. With technical extensions, as suggested in [12], it is possible to address complex actions. Taken $a, b \in Act$, with $a \neq b$,

$$\mathbf{I}_i(\pi, \phi) \wedge \mathbf{Com}_i(\pi) \wedge \mathbf{I}_i(\pi', \phi) \wedge \mathbf{Com}_i(\pi') \wedge$$
$$\wedge\, a \preceq \pi \wedge b \preceq \pi' \wedge \mathbf{Sig}_{j,i} \wedge \mathbf{B}_i(V_j \wedge DONE_i(a) \wedge \mathbf{P}(\mathbf{A}_i(b) \wedge < i : b > \neg V_j)) \rightarrow guilty(i, a, j).$$[10]

[9] Following [15] we can say that an agent i evaluates as morally wrong the formula ϕ iff $M \models \phi \leftrightarrow V_i$ for some model M. We can also say that an action α is wrong for i iff $M \models [X : \alpha]V_i$, for any actor X. Things get more interesting if we move to a local level, in which actions can be evaluated as bad $M, w \models [i : \alpha]V_i$ or they can be always safe $M, w \models [i : \alpha^n]\neg V_i$ (values), or even the only possible cure $M, w \models [i : \alpha^n]\neg V_i \wedge [i : \overline{\alpha}^n]\neg V_i$ or a safe resort at our disposal $M, w \models [i : \overline{\alpha}^n; \alpha]\neg V_i \wedge [i : \overline{\alpha}^n; \overline{\alpha}]\neg V_i$. Of course feeling guilty or ashamed for having challenged a value of other agents can be much more painful and dangerous, but we are not going to go that further with the distinctions, which are in principle possible.

[10] In KARO framework a classical deliberation cycle is assumed [13]. In our case deliberation is a program that updates beliefs, desires, commitments and status of significant others by means of the above defined revision actions.

This means that our agent believed he could in fact avoid the violation state for his significant other and yet pursue his plan π, in that he had a choice w.r.t. to the action to carry out.

Reactions to sense of guilt are influenced by the level of self esteem agents have. We can distinguish two categories, Agt_h, high self esteem agents, which will react providing justifications to their actions and in extreme cases changing the significance they attribute to people. On the other side, Agt_l, low self esteem people will try to find excuses for their actions and to generate reparation goals, that is to perform an action in such a way to avoid further violations. Both agents could also generate the goal of feeling regret, that is to feel bad concerning a past event. With abuse of notation we will write in the object language $i \in Agt_h$ to mean that agent i is a high self esteem agent.

$$\mathbf{I}_i(\pi, \phi) \wedge \mathbf{Com}_i(\pi) \wedge i \in Agt_h \wedge \mathbf{B}_i(V_j) \wedge \\ \wedge\, guilty(i, a, j) \rightarrow [deliberate_i](\neg \mathbf{Sig}_{j,i}) \vee \mathbf{B}_i(\neg V_j) \vee \mathbf{D}_i \neg DONE_i(a)$$

So either i will update authority relations by cancelling j, or he will believe the present state is not violation for j or he will merely wish so (but still believing it, so frustrating his desire). On the other hand...

$$\mathbf{I}_i(\pi, \phi) \wedge \mathbf{Com}_i(\pi) \wedge i \in Agt_l \wedge \mathbf{B}_i(V_j) \wedge guilty(i, a, j) \rightarrow \\ \rightarrow [deliberate_i](\mathbf{I}_i(\pi', \phi) \wedge \mathbf{Com}_i(\pi') \wedge \\ \wedge\, \mathbf{B}_i[i : \pi']\neg V_j) \vee \mathbf{B}_i(\neg \mathbf{P}[i : \overline{a}]\neg V_j)) \vee (\mathbf{D}_i \neg DONE_i(a))$$

The low self esteem agent either will commit to a plan that escapes from violation state, or he will find excuses for his wrongdoing, namely he will generate the belief that what it did was unavoidable, or he will wish that he did not do what he actually did (regret).

4.2 Shame

Shame is the believed lack of a relevant feature, that is the believed incapacity to achieve a value that is important for the agent or for its significant others. If only the first is present we will talk of shame towards the self. We are going to formalize shame by considering an agent that believes it is possible to get over a violation state but that there is no capability for him/her to do so.

$$\mathbf{Sig}_{j,i} \wedge \mathbf{B}_i(\mathbf{A}_{Agt}(a) \wedge \mathbf{A}_{Agt}(\overline{a}) \wedge V_j \wedge [i : a]V_j \wedge [i : \overline{a}]V_j \wedge [Agt \backslash \{i\} : a]\neg V_j) \rightarrow shame(i, a, j)$$

So avoiding V_j is a “norm” or a standard to which i is not able to comply.

What do ashamed agents do? We distinguish LSE reactions and HSE reactions. Similarly with sense of guilt, the first types of reactions will tend to manipulate the belief base in such a way to remove the belief of incapacity, and moreover they will try to repair to their incapacity, which can be done in various way, for instance adopting a goal of the significant other.

$$shame(i, a, j) \wedge i \in Agt_l \rightarrow [deliberate_i]\mathbf{B}_i(\neg[i : \cup Act_i]V_j) \wedge (\mathbf{B}_i(\mathbf{D}_j\phi \rightarrow ([i : \pi]\phi \rightarrow \mathbf{I}_i(\pi, \phi)) \wedge \mathbf{I}_i(\rho, \phi) \wedge \mathbf{Com}_i(\pi) \wedge \mathbf{B}_i[i : \rho] \bigvee_{k \in Agt} \mathbf{B}_k[i : \cup Act_i]V_j$$

The agent openly communicates to others the own incapacity [3], that is, he blushes. The second type of reactions will try to update authority relations, in such a way not to perceive their incapacity as wrong. This is a typical pride reaction.

$$shame(i, a, j) \land i \in Agt_h \rightarrow [deliberate_i]\neg\mathbf{Sig}_{j,i}$$

5 Conclusion and Future Work

In this paper we provided a formal language to describe sense of guilt and shame as social emotions. In order to do this we grounded our work on the cognitive theory of Castelfranchi and Miceli, the psychological theory of significant others by Higgins, the rational action theory in the KARO framework by Meyer and colleagues. The cognitive science perspective has allowed us to build an anatomy of these emotions in terms of basic cognitive ingredients such as Beliefs, Goals and Values. We described formally the operations that allow agents to change their evaluations together with the people they take as references, and we connected these to shame, sense of guilt and their dynamics.

Much work still needs to be done. Apart from what already pointed throughout the paper, we would like to: investigate further the theory of cognitive dissonance and to give a formal characterization of the role of emotions in its resolution; to shed more light on the characterization of emotions by studying the logical models we used to talk about them: could we rewrite the conditions that trigger these emotions without recurring to past reasoning, but only as in [13] reasoning about the resulting conditions after an action execution? Finally we would like to investigate the connection of feeling ashamed and guilty with other feelings like happiness and sadness already formally described in [13] and the agent types defensive and offensive strategies in [16] and [14].

Acknowledgments. The authors thank Maria Miceli, Frank Dignum, Rosaria Conte, Davide Grossi for the inspiring discussions over norms and emotions; a further special thanks to the anonymous reviewers for their precise remarks and to Davide Grossi for the Ovidius quote.

References

1. Blackburn, P., de Rijke, M., Venema, Y.: Modal Logic. Cambridge Tracts in Theoretical Computer Science (2001)
2. Castelfranchi, C.: Cognitive anatomy of shame and guilt: Differences, functions, defensive moves. EABCT, Manchester (2004)
3. Castelfranchi, C., Poggi, I.: Blushing as a discourse: Was darwin wrong? In: Shyness and Embarrassment: Perspectives from Social Psychology, pp. 230–251 (1990)
4. Conte, R., Paolucci, M.: Responsibility for societies of agents. JASSS (2004), `http://jasss.soc.surrey.ac.uk/7/4/3.html`
5. Damasio, A.: Descartes' Error: Emotion, Reason and the Human Brain. NY G.P. Putnam's Sons (1999)

6. de Sousa, R.: Emotion. In: Stanford Encyclopedia of Philosophy (2003)
7. Grossi, D., Royakkers, L.M.M., Dignum, F.: Organizational structure and responsibility. Forthcoming (2007)
8. Higgins, E.T.: Self-discrepancy: A theory relating self and affect. Psychological Review 94, 319–340 (1987)
9. Oatley, K., Jenkins, J.M.: Understanding Emotions. Blackwell, Oxford (1996)
10. Festinger, L., Carlsmith, J.M.: Cognitive consequences of forced compliance. Journal of Abnormal and Social Psychology 58, 203–210 (1959)
11. Marsella, S.C., Johnson, W.L., Labore, C.: Interactive pedagogical drama. In: Sierra, C., Gini, M., Rosenschein, J.S. (eds.) Proceedings of the Fourth International Conference on Autonomous Agents, pp. 301–308. ACM Press, New York, USA (2000)
12. Meyer, J.J.C.: A different approach to deontic logic: Deontic logic viewed as a variant of dynamic logic. Notre Dame J. of Formal Logic 29(1), 109–136 (1988)
13. Meyer, J.J.Ch.: Reasoning about emotional agents. In: Mántaras, R.L.d., Saitta, L. (eds.) Proc.16th European Conf. on Artif. Intell (ECAI 2004), pp. 129–133. IOS Press, Amsterdam (2004)
14. Miceli, M.: How to make someone feel guilty: Strategies for guilt inducement and their goals. Journal for the Theory of Social Behaviour 22, 81–104 (1992)
15. Miceli, M., Castelfranchi, C.: A cognitive approach to values. Journal for the Theory of Social Behaviour 19, 169–194 (1989)
16. Miceli, M., Castelfranchi, C.: How to silence one's conscience: Cognitive defenses against the feeling of guilt. Journal for the Theory of Social Behaviour 28, 287–318 (1998)
17. Maria M.: Personal communication (2006)
18. Ortony, A., Clore, G.L., Collins, A.: The cognitive structure of Emotions. Cambridge University Press, Cambridge (1988)
19. Pitt, J.: Digital blush: towards shame and embarrassment in multi-agent information trading applications. Cognition, Technology and Work 6, 23–36 (2004)
20. Royakkers, L.M.M.: Extending deontic logic for the formalization of legal rules. Kluwer Academic Publishers, Dordrecht (1998)
21. Sloman, A.: Motives, mechanisms and emotions. In: Boden, M. (ed.) The Philosophy of Artificial Intelligence, pp. 231–247 (1990)
22. Staller, A., Petta, P.: Introducing emotions into the computational study of social norms: A first evaluation. Journal of Artificial Societies and Social Simulation, 4(1) (2001), `http://www.soc.surrey.ac.uk/JASSS/4/1/2.html`
23. van Benthem, J.: Where is logic going, and should it? In: Bencivenga, E. (ed.) What is to be Done in Philosophy? (2005)
24. van der Hoek, W., van Linder, B., Meyer, J.-J.C.: A logic of capabilities. In: Matiyasevich, Y.V., Nerode, A. (eds.) LFCS 1994. LNCS, vol. 813, pp. 366–413. Springer, Heidelberg (1994)
25. van der Hoek, W., van Linder, B., Meyer, J.-J.C.: An integrated modal approach to rational agents. In: Proc. of PRR97, Practical Reasoning and Rationality (1997)
26. van Linder, B., van der Hoek, W., Meyer, J.J.C.: Actions that make you change your mind. Knowledge and Belief in Philosophy and Artificial Intelligence, 103–146 (1995)