Shame: when emotion and reasoning are linked

Carole Adam[1] and Dominique Longin[2]

[1] Univ. Grenoble-Alpes (UJF) - LIG - Grenoble, France
carole.adam@imag.fr,
http://membres-lig.imag.fr/cadam/
[2] CNRS, Univ. Toulouse (UPS) - IRIT - Toulouse, France
Dominique.Longin@irit.fr,
http://www.irit.fr/∼Dominique.Longin/

Abstract. Some emotions, described as "basic" in the literature, are almost reflexes. Other emotions are triggered *via pattern matching* mechanisms operating on specific mental states (most often epistemic and motivational) to determine the (in)congruence of these states. Yet other emotions come from more or less complex cognitive mechanisms (and we thus call them complex emotions) such as counterfactual reasoning (*e.g.* guilt or regret), normative judgement (*e.g.* shame or pride), probabilistic evaluations of the world (*e.g.* surprise), *etc.*. In the following, we study and formalise the complex emotion of shame that is of particular importance in social behaviour, and illustrate it on some scenarios.

Keywords: emotions, shame, modal logic

1 Introduction

Elster [1, p. 145] highlights "*an immensely powerful influence*" of social norms on behaviour. In particular, shame touches us in what is most intimate and personal because it has a strong influence on our self-image and the way we believe to be socially perceived [2]. According to Elster, shame is the support of social norms: for instance, if an agent violates a social norm, we can refuse to deal with it, which may make it shameful; and the more it costs us to refuse to deal with it, the most important its shame will be [1, p. 146]. In other words, shame influences our social behaviour. It is thus an emotion of the greatest importance, but paradoxically very little studied in computer science. The goal of this paper is to propose a fine-grained formalization of shame, allowing individual agents to adopt an appropriate behaviour in particular circumstances. Possible applications include entertainment (*e.g.* role-playing games) or education (*e.g.* serious games, tutoring systems). For instance, if a pedagogical agent detects that its student is ashamed of speaking English because he does not feel confident in his ability, it could decide to set up strategies to reassure him. Reasoning about the user's shame can be used in anticipation to decide (not) to perform a given action. Our longer-term goal is to use shame (of one agent in front of the others) as the motor of the dynamics of a multi-agent system (see perspectives in conclusion of this paper).

According to Scherer's multi-componential view [3], emotions are "episodes" having a certain duration (very short but not instantaneous) and a certain dynamics. The following components win almost unanimous support in psychology: *the sentiment* (the feeling of the emotion); *the psychophysiological response* (*e.g.* acceleration of heart rate, body temperature increase); *the motor expression* (*e.g.* face, voice, gestures); *the action tendency* (not to be confused with the action itself); and *the cognitive appraisal*.

In cognitive appraisal theories, this last component causes the other four; it represents the cognitive process of evaluating a given stimulus and triggering a differentiated emotional response (*i.e.* it determines which emotion is triggered). As a result, the cognitive structure of emotions is a mental state, that similarly to belief, desire, intention, *etc.* refers to a state or an object of the world. Therefore emotions are always about something (the object of the appraisal).

In the following, we set up to characterise the cognitive structure of shame. In order to not excessively complicate our study, we do not study the aspects linked to its intensity (on this topic see *e.g.* Lorini [4]). In Section 2 we analyse the emotion of shame; we then present our formal framework in Section 3; we use our formal framework in Section 4 to provide a logical characterisation of shame (mainly following Castelfranchi and Poggi's conceptualisation [5]) and to illustrate different uses of this emotion on some scenarios. Finally we discuss related works in Section 5 before concluding in Section 6.

2 Shame

Shame has been largely studied in psychology [6, 7, 2, 8, 9]. This emotion is perceived as negative, and we are particularly sensitive to it because it makes us focus on our person as a whole, on the damage to our image and to our face (Lewis [10]). Elster [1, pp. 152–153] says that in the case of guilt one sees oneself as having done something bad, while in the case of shame one sees oneself as a bad person. Shame plays a key social role: it has "*the function of cognitive mediators of the individual's social behaviour. (...) Though the unpleasant feelings they inflict they lead one to avoid or remediate possible misfunctioning in one's relationships with other people.*" [5, p. 230]. Lazarus highlights that even if shame can be seen as occurring privately and without any witnesses, it actually always involves other people [8, p. 241].

Shame is mainly linked to the belief of having violated an internalised normative standard[11][3]. Following [9, p. 142–143], this norm is an "important moral value" that one feels committed to respect and whose violation is considered as inexcusable. According to Lazarus [8, p. 240 & 242], shame involves thoughts or actions that violate an "internalised social prescription" and where the blame is

[3] Typically, an agent can be aware of a normative standard (in a general sense, *e.g.* moral value, legal obligations, *etc.*) without internalising it if this agent does not identify with it, *i.e.* it does not consider it important to respect it. This does not mean that an agent necessarily respects all its internalised norms, but it cannot be indifferent to their violation. For example, if one believes that it is forbidden to download music online but still does, it means that they did not internalise that law.

for oneself (see also [9, pp. 136–144]). But finally we agree with Turrini *et al.* when they claim that the norms involved in shame are not necessarily moral values but rather normative standards (*e.g.* being ashamed of one's nose or poverty).

We also agree with Castelfranchi and Poggi [5] on another important aspect of shame: one can feel shame in front of oneself and/or in front of someone else.[4] Elster [1, p. 151] quotes the example of Mathilde de la Mole who is ashamed[5] of being in love with the son of a carpenter (Julien Sorel): ss long as she has not told anyone about her secret, she only feels shame in front of herself; only when she thinks (rightly or wrongly) that other people are aware of her feelings does she feel shame in front of them. A corollary to this is that to feel shame *w.r.t.* others, it is necessary to believe that they are aware of the object of our shame [5]. Of course, as highlighted by these authors, one can project oneself in the future and imagine the shame that one would feel if one's relatives were aware of something. Lazarus [8, p. 241] defends the idea that it is only necessary to imagine how some people would react if they knew what we did or did not do in order to feel shame for it. But in this case, [5] argue that shame in front of one's relatives is not really felt but just imagined, thus contradicting Lazarus. Elster [1, p. 152] imposes a stronger condition by mentioning the "presence of others" but it seems that this condition is not confirmed by experiments in psychology (see [2, p. 14] for example, who showed that a significant number of queried people reported experiences of shame arising when they were alone).

As we can see from this psychological literature review, theories are often vague and/or ambiguous, and do not agree on all details of the definition of shame. We thus had to choose one theory to formalise, and we chose to follow Castelfranchi and Poggi's cognitive analysis [5, p. 233], which seems the most adapted to a BDI logical formalisation. According to this theory, the fact that an agent i feels shame about a fact F in front of an agent j requires four conditions (that we put in parallel with their own example of a doctor ashamed in front of their patient for not knowing a new medicine, making him a bad doctor): (1) agent i believes that j believes that F is true (*e.g.* the doctor believes that his patient believes that he does not know about this new medicine); (2) agent i believes that j believes that if F is true then agent i is negatively appraised *w.r.t.* a certain criterion C (*e.g.* the doctor believes that according to his patient, ignorance of this new medicine makes him a bad doctor); (3) agent i believes that i and j commonly believe that the criterion C is a shared normative standard for them both (*e.g.* the doctor and his patient commonly believe that it is a normative standard to be a good doctor); (4) finally, agent i is not indifferent to j's opinion of him *w.r.t.* C. In other words, i prefers j to have a positive opinion of him with respect to C, *i.e.* to believe that he has this property C (*e.g.* the doctor prefers his patient to think that he is a good doctor).

[4] The expression "in front of" designates in [5] the person (or the group of people), physically present or not, *w.r.t.* whom one feels a given emotion.

[5] Given that she violates a social norm important to her, *i.e.* that a noble woman should not fall in love with someone of an inferior social rank

This last point is in agreement with Lazarus [8, p. 241], according to whom in shame, there is a potentially critical person (regarding the negative state that we are ashamed of) whose approbation is important to us.

It is important to note that when i and j are the same agent, this agent is ashamed in front of itself [5]. Moreover, agent i can be ashamed in front of agent j even if it does not itself share j's beliefs imposed in conditions (1) and (2), as long as it believes that j does have these beliefs (*e.g.* the doctor could believe that ignoring this new medicine does not make him a bad doctor). However, for i to be ashamed (in front of itself or another agent), it is necessary that i shares the normative standard imposed in condition (3), in order to feel concerned by its violation. For example, wiping your nose in public is very impolite in Japan; if one does not know it but realises it while wiping their nose in public, one has no reason to feel ashamed unless one recognises *de facto* this standard as having to be respected. Finally, as explained above, an agent can also be ashamed both in front of itself and in front of someone else, at the same time.

3 Formal framework

3.1 Basic language and mental attitudes

Let AGT be the finite set of agents and $2^{AGT*} = 2^{AGT} \setminus \emptyset$. Let ATM be the set of atomic formulas and $ATM_i \subseteq ATM$ for any $i \in AGT$ the finite set of those representing properties of agent i. The language $\mathcal{L}_{\mathcal{SL}}$ of the logic of shame $\mathcal{SL}$ is defined by the following BNF:

$$\varphi :: p \mid p_i \mid \neg\varphi \mid \varphi \vee \varphi \mid MBel_G\,\varphi \mid Goal_i\,\varphi \mid NStand_i\,\varphi$$

where $p \in ATM$, $p_i \in ATM_i$, $i \in AGT$ and $G \in 2^{AGT*}$. The other classical connectors ($\top$, $\bot$, $\wedge$, $\rightarrow$ and $\leftrightarrow$) are defined in the usual way. p_i reads: p is a property of agent i; $MBel_G\,\varphi$ reads: "the fact that φ is true is a mutual belief of the group of agents G". $Goal_i\,\varphi$ reads: "agent i [has the chosen goal/prefers] that φ". (This is a goal *à la* Cohen&Levesque[6]; see [14].) $NStand_i\,\varphi$ reads: "φ is a normative standard of agent i that is particularly important to i". [7]

We define some abbreviations summarized in Fig. 1:

[6] As in [12], these goals can come from desires (intrinsically endogenous to an individual), from internalised norms, or from exogenous goals imposed on the individual (see [13] for more details). Therefore the satisfaction of a chosen goal is not necessarily positive for the agent, but "less negative" than its non-satisfaction. Moreover, goals are not necessarily realistic: an agent can have a goal without believing that it can be achieved sometime. Finally goals are not necessarily achievement goals: i can have a goal that φ without believing that φ is false, and without wanting to make φ true if it is false. Goals are therefore semantically represented by sets of preferred worlds; we use "(chosen) goal" and "preference" as synonymous.

[7] This means that φ is an internalised standard for i, that is, i commands itself to respect it [15]. In this sense, i is morally responsible for the realisation of φ. The fact that this represents a normative standard particularly important for i is consistent with the type of internalised norms described by [9, p. 142–143] or [5]. The agent is likely to lose face when violating this type of standard.

$$p_G \overset{d\acute{e}f}{=} \bigwedge_{i \in G} p_i \qquad (\mathrm{Def}_{pG})$$

$$p_\emptyset \overset{d\acute{e}f}{=} \bigwedge_{i \in AGT} \neg p_i \qquad (\mathrm{Def}_{p\emptyset})$$

$$Bel_G\, \varphi \overset{d\acute{e}f}{=} \bigwedge_{i \in G} MBel_{\{i\}}\, \varphi \qquad (\mathrm{Def}_{Bel_G})$$

$$Bel_i\, \varphi \overset{d\acute{e}f}{=} MBel_{\{i\}}\, \varphi \overset{d\acute{e}f}{=} Bel_{\{i\}}\, \varphi \qquad (\mathrm{Def}_{Bel_i})$$

$$Goal_G\, \varphi \overset{d\acute{e}f}{=} \bigwedge_{i \in G} Goal_i\, \varphi \qquad (\mathrm{Def}_{Goal_G})$$

$$NStand_G\, \varphi \overset{d\acute{e}f}{=} \bigwedge_{i \in G} NStand_i\, \varphi \qquad (\mathrm{Def}_{NStand_G})$$

Fig. 1. Abbreviations of the langage where $i, j \in AGT$, $G \in 2^{AGT*}$

(Def_{pG}) means that property p is shared by all agents in group G; ($\mathrm{Def}_{p\emptyset}$) means that no agent in AGT has property p; (Def_{Bel_G}) reads: φ is a shared belief of all agents in group G; (Def_{Bel_i}) reads: agent i believes that φ is true; (Def_{Goal_G}) reads φ is a preference shared by all agents in group G; (Def_{NStand_G}) reads φ is a normative standard shared by all agents in group G and particularly important to them.

3.2 Semantics

$\mathcal{SL}$-frames. $\mathcal{SL}$-frames are tuples $F = \langle W, \mathcal{B}, \mathcal{G}, \mathcal{I} \rangle$ where: W is a non-empty set of possible worlds; $\mathcal{B} : AGT \longrightarrow W \times W$ maps each agent i with a transitive euclidean relation $\mathcal{B}_i \subseteq W \times W$ between possible worlds; $\mathcal{G} : AGT \longrightarrow W \times W$ maps each agent i with a serial relation $\mathcal{G}_i \subseteq W \times W$ between possible worlds; $\mathcal{I} : AGT \longrightarrow W \times W$ maps each agent i with a serial relation $\mathcal{I}_i \subseteq W \times W$ between possible worlds. In the following, we note $\mathcal{R}(w) = \{w' \in W : (w, w') \in \mathcal{R}\}$.

$\mathcal{B}_i(w)$ is the belief state of agent i in world w. Each accessibility relation is transitive and euclidean (see the constraints **(SC1)** in Fig. 2).[8] $\mathcal{G}_i(w)$ is the set of preferred worlds of agent i in the world w, and each relation $\mathcal{G}_i$ is serial **(SC2)**. $\mathcal{I}_i(w)$ is the set of ideal worlds of agent i in the world w, and each relation $\mathcal{I}_i$ is serial **(SC3)**. We also impose that each agent is aware of its preferred worlds

[8] Traditionally, this relation is also serial, meaning that $\mathcal{B}_i(w)$ cannot be empty. In other words, if agent i believes φ in w then there necessarily exists a world accessible from w via $\mathcal{B}$ where φ is true. Here, we do not impose this seriality constraint so $\mathcal{B}_i(w)$ can be empty, meaning that an agent can have contradictory beliefs without making the logic contradictory. This technical choice is made necessary by the semantics of public announcements: indeed public announcements can remove accessible worlds, possibly leaving no accessible world at all, which is contradictory with seriality.

(SC4) and of its ideal worlds **(SC5)**: the worlds representing its goals and its standards from w are the same as those accessible from its epistemic worlds.

(SC1). if $w' \in \mathcal{B}_i(w)$ then $\mathcal{B}_i(w) = \mathcal{B}_i(w')$
(SC2). $\mathcal{G}_i(w) \neq \emptyset$
(SC3). $\mathcal{I}_i(w) \neq \emptyset$
(SC4). if $w' \in \mathcal{B}_i(w)$ then $\mathcal{G}_i(w) = \mathcal{G}_i(w')$
(SC5). if $w' \in \mathcal{B}_i(w)$ then $\mathcal{I}_i(w) = \mathcal{I}_i(w')$

Fig. 2. Semantical constraints where $w \in W$, $i \in AGT$

$\mathcal{SL}$-models. $\mathcal{SL}$-models are $M = \langle F, V \rangle$ with F a $\mathcal{SL}$-frame and $V : ATM \longrightarrow 2^W$ a valuation function. For each formula φ, each model M and each world w of this model, $M, w \Vdash \varphi$ reads "φ is true in world w of model M". We denote $M, w \nVdash \varphi$ the fact that $M, w \Vdash \neg\varphi$. Truth conditions are as follows:

- $M, w \Vdash p$ iff $w \in V(p)$;
- $M, w \Vdash \neg\varphi$ iff it is not the case that $M, w \Vdash \varphi$;
- $M, w \Vdash \varphi \wedge \psi$ iff $M, w \Vdash \varphi$ and $M, w \Vdash \psi$;
- $M, w \Vdash MBel_G\, \varphi$ for every $G \in 2^{AGT*}$ iff $M, w' \Vdash \varphi$ for every $w' \in (\bigcup_{i \in G} \mathcal{B}_i)^+(w)$ where $(\bigcup_{i \in G} \mathcal{B}_i)^+$ is the transitive closure of the union of the G's epistemic accessibility relations;
- $M, w \Vdash Goal_i\, \varphi$ iff $M, w' \Vdash \varphi$ for every $w' \in \mathcal{G}_i(w)$;
- $M, w \Vdash NStand_i\, \varphi$ iff $M, w' \Vdash \varphi$ for every $w' \in \mathcal{I}_i(w)$.

A formula φ is true in a $\mathcal{SL}$-model M iff $M, w \Vdash \varphi$ for each world w of M. φ is valid iff φ is true in each $\mathcal{SL}$-model (we then note $\vDash_{\mathcal{SL}} \varphi$). φ is satisfiable iff it is not valid.

3.3 Axiomatics

It follows from our semantics that operators $MBel_G$ (for every $G \in 2^{AGT*}$) are defined in the K4 logic, and operators Bel_i (for $i \in AGT$) in the K45 logic (see [16]). We can prove the validity of the following properties (for each $G \in 2^{AGT*}$ and $i \in G$):

$$MBel_G\, \varphi \rightarrow Bel_G\, \varphi \tag{MBel1}$$

$$MBel_G\, \varphi \rightarrow MBel_{G'}\, \varphi \quad \text{pour tout} \quad G' \in 2^{G*} \tag{MBel2}$$

$$MBel_G\, \varphi \leftrightarrow \bigwedge_{i \in G} Bel_i\, MBel_G\, \varphi \tag{MBel3}$$

$$\neg Bel_i\, \varphi \rightarrow \neg Bel_i\, MBel_G\, \varphi \tag{MBel4}$$

$$\neg Bel_i\, MBel_G\, \varphi \rightarrow \neg MBel_G\, \varphi \tag{MBel5}$$

The operators $Goal_i$ and $NStand_i$ are defined in a normal logic KD and, thanks to the semantic constraints, the logic $\mathcal{SL}$ verifies the following principles:

$$Goal_i\,\varphi \rightarrow Bel_i\; Goal_i\,\varphi \tag{PIgoal}$$
$$\neg Goal_i\,\varphi \rightarrow Bel_i\,\neg Goal_i\,\varphi \tag{NIgoal}$$
$$NStand_i\,\varphi \rightarrow Bel_i\; NStand_i\,\varphi \tag{PInstand}$$
$$\neg NStand_i\,\varphi \rightarrow Bel_i\,\neg NStand_i\,\varphi \tag{NInstand}$$

These properties respectively mean that if an agent has (resp. does not have) a certain preference or a certain internalised normative standard, then it believes that it has it (resp. does not have it).

3.4 Extension to public announcements

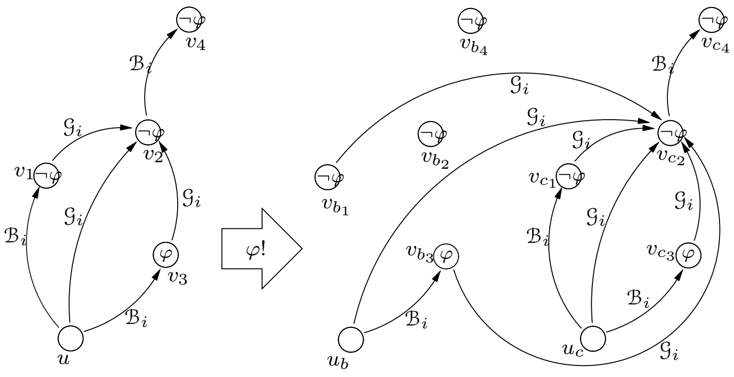
We now want to be able to express properties such as "after the agents have learned some piece of information, agent i will be ashamed", in order to allow reasoning about the dynamics of shame in a MAS. For instance an agent i can want to avoid making some proposition φ true if it believes that when the a group G learns about φ, i will feel ashamed in front of them. Shame-avoidance is thus used as a motivation for (not) acting (see Section 4.1). We therefore extend the language of logic $\mathcal{SL}$ with modal operators of public announcements [17] by adding $[\varphi!]\varphi$ to the BNF defined above. We ground on the framework defined by [18] and extend it with operators of mutual beliefs and internalised normative standards. $[\varphi!]\psi$ reads "ψ is true after the public announcement of φ".

The associated semantics is defined as an update of a $\mathcal{SL}$-model: the update of $M = \langle W, \mathcal{B}, \mathcal{G}, \mathcal{I}, V\rangle$ by $\varphi!$ is the model $M^{\varphi!} = \langle W^{\varphi!}, \mathcal{B}^{\varphi!}, \mathcal{G}^{\varphi!}, \mathcal{I}^{\varphi!}, V^{\varphi!}\rangle$ such that:

$$\begin{aligned}
W^{\varphi!} &= \{u_b : u \in W\} \cup \{u_c : u \in W\}\\
\mathcal{B}^{\varphi!} &= \{(u_b, v_b) : v \in \mathcal{B}(u) \text{ et } M, v \Vdash \varphi\} \cup \{(u_c, v_c) : v \in \mathcal{B}(u)\}\\
\mathcal{G}^{\varphi!} &= \{(u_b, v_c) : v \in \mathcal{G}(u)\} \cup \{(u_c, v_c) : v \in \mathcal{G}(u)\}\\
\mathcal{I}^{\varphi!} &= \ \{(u_b, v_c) : v \in \mathcal{I}(u)\} \cup \{(u_c, v_c) : v \in \mathcal{I}(u)\}\\
V^{\varphi!}(u_b) &= V^{\varphi!}(u_c) = V(u)
\end{aligned}$$

Intuitively, the worlds are duplicated in two groups: one relative to beliefs (u_b) and one relative to preferences and standards (u_c). Regarding accessibility relations, they are integrally reproduced in this latter group while in the former:

- only the elements of the epistemic relation leading to worlds where the announced formula is true are kept;
- the elements of the relations $\mathcal{G}$ and $\mathcal{I}$ are duplicated in such a way that the departure world is a world relative to belief and the arrival world is a preferred world or an ideal world (v_c).



Modèle M Modèle $M^{\varphi!}$

Fig. 3. Example where $M, u \Vdash \neg Bel_i\, \varphi \wedge \neg Bel_i\, \neg\varphi$ while $M^{\varphi!}, u_b \Vdash Bel_i\, \varphi$

Example 1. An example is given in Figure 3. To simplify, the starting model $M = \langle W, \mathcal{B}, \mathcal{G}, \mathcal{I}, V\rangle$ only contains elements of $\mathcal{B}_i$ and of $\mathcal{G}_i$ and we suppose that $W = \{u, v_1, v_2, v_3, v_4\}$. For instance, let φ be a formula that means "it is sunny"; $\neg Bel_i\, \varphi \wedge \neg Bel_i\, \neg\varphi$ then means "agent i does not know if it is sunny or not", and $Bel_i\, \varphi$ means "agent i believes it is sunny". The new model $M^{\varphi!}$ stemming from the announcement of $\varphi!$ has $W^{\varphi!}$ as its set of worlds, where $W^{\varphi!} = \{u_b, v_{b_1}, v_{b_2}, v_{b_3}, v_{b_4}\} \cup \{u_c, v_{c_1}, v_{c_2}, v_{c_3}, v_{c_4}\}$. We can see that all formulas true in M, u are true in $M^{\varphi!}, u_b$ except the fact that: $M, u \Vdash \neg Bel_i\, \varphi \wedge \neg Bel_i\, \neg\varphi$ (while $M, u \nVdash Bel_i\, \varphi$) and $M^{\varphi!}, u_b \Vdash Bel_i\, \varphi$ (while $M^{\varphi!}, u_b \nVdash \neg Bel_i\, \varphi \wedge \neg Bel_i\, \neg\varphi$). In other words, before the update agent i did not know whether φ was true or not (*e.g.* , whether it was sunny or not), and after the update i believes that φ is true (*e.g.* , i believes it is sunny). The announcement has therefore extended the beliefs of agent i.

Proposition 1. *For each formula φ, if M is a $\mathcal{SL}$-model then $M^{\varphi!}$ is also a $\mathcal{SL}$-model.*

and the truth condition associated to public announcements is the following:

- $M, u \Vdash [\varphi!]\psi$ ssi $M^{\varphi!}, u_b \Vdash \psi$

The previous notions of true formula in a $\mathcal{SL}$-model, of validity and satisfiability are extended to take public announcements into account.

We can show that the truth conditions associated to the operators make the following properties valid:

$$\models_{\mathcal{SL}} [\varphi!]p \leftrightarrow p \quad \text{où} \quad p \in ATM \tag{RAp}$$
$$\models_{\mathcal{SL}} [\varphi!]\neg\psi \leftrightarrow \neg[\varphi!]\psi \tag{RAn}$$
$$\models_{\mathcal{SL}} [\varphi!](\psi_1 \wedge \psi_2) \leftrightarrow [\varphi!]\psi_1 \wedge [\varphi!]\psi_2 \tag{RAa}$$
$$\models_{\mathcal{SL}} [\varphi!]Bel_i\,\psi \leftrightarrow Bel_i\,(\varphi \rightarrow [\varphi!]\psi) \tag{RAb}$$
$$\models_{\mathcal{SL}} [\varphi!]Goal_i\,\psi \leftrightarrow Goal_i\,\psi \tag{RAg}$$
$$\models_{\mathcal{SL}} [\varphi!]NStand_i\,\psi \leftrightarrow NStand_i\,\psi \tag{RAv}$$

(RAp) means that a public announcement does not change facts, goals (RAg) or standards (RAv) of an agent. (RAn) means that a formula is false after an announcement iff it is false that this formula is true after the announcement. (RAa) means that two facts are true after an announcement iff each of them is separately true after this announcement. Finally, (RAb) means that a belief of agent i is true after an announcement iff this agent believes that if the content of this announcement is true then after the announcement of this content the objet of this belief will be true.

We can prove from the previous properties that:

$$\models_{\mathcal{SL}} [\varphi!]\top \tag{N$_{[\varphi!]}$}$$

We can also prove that the following rules of equivalence keep their validity:

$$\text{if} \quad \psi \leftrightarrow \psi' \quad \text{then} \quad [\varphi!]\psi \leftrightarrow [\varphi!]\psi' \tag{RE$_{[\varphi!]}$}$$
$$\text{if} \quad \varphi \leftrightarrow \varphi' \quad \text{then} \quad [\varphi!]\psi \leftrightarrow [\varphi'!]\psi \tag{RE$'_{[\varphi!]}$}$$

By definition [16, p. 115] the properties (RAa), (N$_{[\varphi!]}$), (RE$_{[\varphi!]}$) and (RAn) imply that operators $[\varphi!]$ are defined in a KD logic. The equivalences (RAp) to (RAv) and inference rules (RE$_{[\varphi!]}$) and (RE$'_{[\varphi!]}$) above are called “reduction axioms”: they allow to reduce any formula $[\varphi!]\varphi$ to a formula that does not contain any $[\varphi!]$ operator. As shown by [17] there is no reduction axiom for mutual belief and the axiomatics above is this incomplete.

Definition 1 (boolean formula and positive formula). *For each $p \in ATM$, the set of boolean formulas is such that:*

$$P ::= p \mid \neg P \mid P \vee P$$

and the set of positive formulas is such that ($i \in AGT$, $G \in 2^{AGT*}$) :

$$\varphi^+ ::= P \mid \varphi^+ \vee \varphi^+ \mid \varphi^+ \wedge \varphi^+ \mid MBel_G\,\varphi^+ \mid Goal_i\,\varphi^+ \mid NStand_i\,\varphi^+$$

For example, $Bel_i\,\neg p$, $Goal_i\,MBel_G\,(p \vee \neg q)$ and $p \rightarrow Bel_i\,p$ are positive formulas, but not $Bel_i\,Bel_j\,p \rightarrow Bel_i\,p$ (because this formula is equivalent to $\neg Bel_i\,Bel_j\,p \vee Bel_i\,p$ and $\neg Bel_i\,Bel_j\,p$ is not a positive formula).

Finally, we can prove the following properties for each $i \in AGT$, $G \in 2^{AGT*}$:

$$[\varphi!] Bel_i \, \varphi \tag{1}$$
$$[\varphi!] MBel_G \, \varphi \tag{2}$$
$$\varphi^+ \to [\psi!]\varphi^+ \tag{3}$$
$$\neg Bel_i \, \neg P \to [P!] \neg Bel_i \, \neg P \tag{4}$$
$$Bel_i \, [\varphi!]\psi \to [\varphi!] Bel_i \, \psi \tag{5}$$

(1) and (2) respectively mean that after φ has been announced, all agents believe (resp. mutually believe) that φ is true. (3) means that any positive formula stays true after any announcement. See [17] for the proof of these three properties. (5) means that if an agent believes that ψ will be true after the announcement of φ, then after the announcement of φ this agent will believe that ψ is true. (This property can be easily proven from (RAb) and principles of the logic.)

4 Formalisation

Definition 2. *For each agent $i \in AGT$, each group of agents $G \in 2^{AGT*}$ and each formula $p_i \in ATM_i$:*

$$\begin{aligned} Shame_i \, (G, \varphi, p_i) \overset{déf}{=} & Bel_i \, MBel_G \, \varphi \wedge \\ & Bel_i \, MBel_G \, (\varphi \to \neg p_i) \wedge \\ & Bel_i \, MBel_{G \cup \{i\}} \, NStand_{G \cup \{i\}} \, p_i \wedge \\ & Goal_i \, Bel_j \, p_i \end{aligned}$$

$Shame_i \, (G, \varphi, p_i)$ reads: “agent i feels shame in front of group G that φ is true, in relation with property p_i” and the elements of the disjunction correspond to the properties presented in the previous section. (In particular, the last component means that agent i prefers that j believes that i has the property p.) According to these properties, the agent i can feel shame:

- in front of himself only (when G is reduced to $\{i\}$);
- in front of a group G only (and not in front of himself) when it does not belong to it (when $i \notin G$);
- both (when $G = G' \cup \{i\}$ with $G' \neq \emptyset$ and $i \notin G'$).

Definition 3. *For each agent $i \in AGT$, each group of agents $G \in 2^{AGT*}$ and each formula $p_i \in ATM_i$:*

$$Shame_i \, (G, \varphi) \overset{déf}{=} \bigvee_{p_i \in ATM_i} Shame_i \, (G, \varphi, p_i)$$

$Shame_i \, (G, \varphi)$ reads: “agent i feels shame in front of group G that φ is true” and it holds iff this agent feels shame in front of group G that φ is true in relation with at least one property p_i.

Finally, the fact that shame is defined from beliefs of agent i means that this agent can be mistaken and feel shame for something in front of a certain group while this group does not really have the required beliefs or normative standards (agent i has wrong beliefs).

It is easy to prove that:

$$\models_{\mathcal{SL}} \mathit{Shame}_i\,(G,\varphi,p_i) \to \mathit{Shame}_i\,(\{i\},\varphi,p_i) \quad \text{ssi} \quad i \in G \tag{SH1}$$

$$\models_{\mathcal{SL}} \mathit{Shame}_i\,(G,\varphi,p_i) \to \mathit{Bel}_i\,\mathit{MBel}_{G\cup\{i\}}\,\mathit{NStand}_i\,p_i \tag{SH2}$$

$$\models_{\mathcal{SL}} \mathit{Shame}_i\,(G,\varphi,p_i) \to \mathit{Shame}_i\,(G',\varphi,p_i) \quad \text{pour tout} \quad G' \in 2^{G*} \tag{SH3}$$

$$\models_{\mathcal{SL}} \mathit{Shame}_i\,(G,\varphi,p_i) \to \mathit{Bel}_i\,\mathit{Shame}_i\,(G,\varphi,p_i) \tag{SH5}$$

$$\models_{\mathcal{SL}} \neg\mathit{Shame}_i\,(G,\varphi,p_i) \to \mathit{Bel}_i\,\neg\mathit{Shame}_i\,(G,\varphi,p_i) \tag{SH6}$$

(SH1) illustrates the fact that if agent i feels shame in front of a group G that it belongs to, then i feels shame in front of itself. (SH2) does not presuppose that i necessarily belongs to group G and illustrates the fact that even when i feels shame in front of a group G that it does not belong to (i is thus not ashamed in front of himself) there must be a mutual belief between agent i and agents in group G that p_i is a moral value of i. (SH3) represents the fact that if agent i feels shame in front of a group G then i feels shame in front of any non-empty subgroup G' of G. (SH5) and (SH6) illustrate the fact that an agent is aware if what it is, and of what it is not, ashamed of.

Example 2 (Absence of shame). Let's consider $G = \{\mathit{Tom}, \mathit{Maxim}, \mathit{Kenzo}\}$ such that $G \subseteq AGT$, $\mathit{untidy} \in ATM$ meaning that Tom's room is untidy, $\mathit{cool}_{\mathit{Tom}} \in ATM_{\mathit{Tom}}$ meaning that Tom has the property to be cool, and cool_{AGT} meaning that everybody has the property to be cool. Maxim and Kenzo come to Tom's home to play with him, and they see that his room is untidy ($\mathit{MBel}_G\,\mathit{untidy}$). None of them considers that having his room untidy makes Tom "uncool" ($\bigwedge_{i\in G} \neg\mathit{Bel}_i\,(\mathit{untidy} \to \neg\mathit{cool}_{\mathit{Tom}})$) and they believe that it is particularly important to be cool ($\mathit{MBel}_G\,\mathit{NStand}_G\,\mathit{cool}_{AGT}$). Finally, each of them prefers that the others believe he/she is cool ($\bigwedge_{i\in G} \mathit{Goal}_i\,\mathit{Bel}_{G\setminus\{i\}}\,\mathit{cool}_i$).

It is easy to show that Tom is not ashamed in front of his friends that his room is untidy (because he does not believe that this untidines makes him *uncool*). Formally, if we note KB_2 the set of all these facts, this is illustrated by the validity of the following principle:$\models_{\mathcal{SL}} KB_2 \to \neg\mathit{Shame}_{\mathit{Tom}}\,(G, \mathit{untidy}, \mathit{cool}_{\mathit{Tom}})$.

Example 3 (Shame in front of others). In this new example, we replace the first two hypotheses of the previous example with the three following ones (and we keep the other two): everybody has seen Tom's room tidy ($\mathit{MBel}_G\,\mathit{tidy}$) and Tom believes that Maxim and Kenzo mutually believe that this makes him *uncool* ($\mathit{Bel}_{\mathit{Tom}}\,\mathit{MBel}_{\{\mathit{Maxim},\mathit{Kenzo}\}}\,(\mathit{tidy} \to \neg\mathit{cool}_{\mathit{Tom}})$), although he does not share this opinion ($\neg\mathit{Bel}_{\mathit{Tom}}\,(\mathit{tidy} \to \neg\mathit{cool}_{\mathit{Tom}})$). This facts constitute the initial set KB_3. From this example and the principles of our logic, we can show that Tom feels shame in front of Maxim and Kenzo (but not in front of himself) that his room is tidy. So, formally:

1. $\models_{\mathcal{SL}} KB_3 \rightarrow \mathit{Shame}_{\mathit{Tom}}(\{\mathit{Maxim}, \mathit{Kenzo}\}, \mathit{tidy})$.
2. $\models_{\mathcal{SL}} KB_3 \rightarrow \neg \mathit{Shame}_{\mathit{Tom}}(\{\mathit{Tom}\}, \mathit{tidy})$.

Example 4 (Absence of shame due to uninterest in being liked). In this third example, Tom believes that his brother Arthur knows that his room is tidy, and that in Arthur's view this makes Tom *uncool*. It is obvious for Tom and Arthur that ideally one should be cool. But since it is his brother, Tom does not especially prefer at that time that Arthur believes that Tom is cool (that is, $\neg \mathit{Goal}_{\mathit{Tom}}\, \mathit{Bel}_{\mathit{Arthur}}\, \mathit{cool}_{\mathit{Tom}} \wedge \neg \mathit{Goal}_{\mathit{Tom}}\, \neg \mathit{Bel}_{\mathit{Arthur}}\, \mathit{cool}_{\mathit{Tom}}$).

Consequently, Tom does not feel any shame in front of his brother for his room being tidy even if this makes him *uncool*, which is negative *per se* (the formula $\mathit{Shame}_i(\{\mathit{Arthur}\}, \mathit{tidy}, \mathit{cool}_{\mathit{Tom}})$ is false).

4.1 Dynamics of shame

We have said above that shame may be driving some of our behaviours. We thus propose in the following to formally illustrate this aspect. The reactions we adopt when ashamed are context-dependent, so we first set here the frame that will serve as a running example in the sequel of this section.

Example 5 (Shame and belief evolution). Let's again consider Tom and his untidy room (KB5a), of which he is aware (KB5b). Like any teenager, he wants to project a positive self-image to his friends: Maxim and Kenzo of course, but also Lila, his new girlfriend (KB5c). He knows that he and his two mates share the belief that one can have an untidy room and still be *cool* (KB5d). He does not believe either that having an untidy room shows immaturity, but he ignores Lila's opinion on this point (KB5f).

Let $AGT = \{\mathit{Tom}, \mathit{Kenzo}, \mathit{Maxim}, \mathit{Lila}\}$ be the set of all agents and $ATM = \{\mathit{untidy}, \mathit{cool}_{AGT}, \mathit{mature}_{AGT}\}$ the set of all atomic formulas. The initial base of facts KB_5 is as follows:

$$\mathit{untidy} \tag{KB5a}$$

$$\mathit{Bel}_{\mathit{Tom}}\, \mathit{untidy} \tag{KB5b}$$

$$\mathit{Goal}_{\mathit{Tom}}\, \mathit{Bel}_{AGT}\, \mathit{cool}_{\mathit{Tom}} \tag{KB5c}$$

$$\mathit{Bel}_{\mathit{Tom}}\, \mathit{MBel}_{\{\mathit{Tom},\mathit{Maxim},\mathit{Kenzo}\}}\, \neg(\mathit{untidy} \rightarrow \neg \mathit{cool}_{\mathit{Tom}}) \tag{KB5d}$$

$$\mathit{Bel}_{\mathit{Tom}}\, \neg(\mathit{untidy} \rightarrow \neg \mathit{mature}_{\mathit{Tom}}) \tag{KB5e}$$

$$\begin{split}&\neg \mathit{Bel}_{\mathit{Tom}}\, \mathit{Bel}_{\mathit{Lila}}\, (\mathit{untidy} \rightarrow \neg \mathit{mature}_{\mathit{Tom}}) \wedge \\ &\neg \mathit{Bel}_{\mathit{Tom}}\, \neg \mathit{Bel}_{\mathit{Lila}}\, (\mathit{untidy} \rightarrow \neg \mathit{mature}_{\mathit{Tom}})\end{split} \tag{KB5f}$$

$$\mathit{MBel}_{AGT}\, \mathit{NStand}_{AGT}\, (\mathit{cool}_{AGT} \wedge \mathit{mature}_{AGT}) \tag{KB5g}$$

When friends enter Tom's room and find it untidy, this counts for us as a public announcement of $\mathit{untidy}!$. After this announcement, all agents thus mutually believe that Tom's room is untidy and Tom believes that this mutual belief holds ($[\mathit{untidy}!]\mathit{Bel}_{\mathit{Tom}}\, \mathit{MBel}_{AGT}\, \mathit{untidy}$); in particular Tom keeps believing it ($[\mathit{untidy}!]\mathit{Bel}_{\mathit{Tom}}\, \mathit{untidy}$). Of course, the room is still untidy ($[\mathit{untidy}!]\mathit{untidy}$)

and all of Tom's beliefs and preferences are preserved by the announcement (if φ represents one of the facts between (KB5c) and (KB5g) then we have $[untidy!]\varphi$).

By definition of shame, it is immediate to see that Tom does not feel any shame in front of anyone after the announcement *untidy*!. On the contrary, if this first announcement is followed by another announcement: Lila declares that she believes untidiness to be a sign of immaturity (*i.e.* $Bel_{Lila}(untidy \to \neg mature_{Tom})!$ is announced), then Tom will feel shame in front of her regarding this property. It is then quite immediate that:

$$\models_{\mathcal{SL}} KB_5 \to [desordre!][Bel_{Lila}(untidy \to \neg mature_{Tom})!] \\ Shame_{Tom}(\{Lila\}, untidy, mature_{Tom})$$

which means that the initial situation is enough to show that after everybody is informed that Tom's room is untidy, and then that Lila considers that as a lack of maturity, Tom feels shame in front of Lila that his room is not tidy because he believes that this negates his maturity in the eyes of Lila.

From (KB5d) we can also show that at no instant (before the first announcement, between the first and second announcements, and after the second announcement) Tom feels shame in front of himself, Kenzo and/or Maxim about the untidiness of his room (since this untidiness does not make Tom uncool in their eyes).

$$\models_{\mathcal{SL}} KB_5 \to \neg Shame_{Tom}(\{Tom, Kenzo, Maxim\}, untidy, cool_{Tom})$$

$$\models_{\mathcal{SL}} KB_5 \to [desordre!] \\ \neg Shame_{Tom}(\{Tom, Kenzo, Maxim\}, untidy, cool_{Tom})$$

$$\models_{\mathcal{SL}} KB_5 \to [desordre!][Bel_{Lila}(untidy \to \neg mature_{Tom})!] \\ \neg Shame_{Tom}(\{Tom, Kenzo, Maxim\}, untidy, cool_{Tom})$$

5 Related works

Emotions have already been formalised by various authors: for instance [19] provide a first formalisation of emotions; [4] studies the intensity of emotions; Steunebrink and colleagues [20] formalise Ortony, Clore and Collins' theory of emotions. We have also proposed such a formalisation ourselves [21].

In [22], Steunebrink and colleagues formalise shame from Ortony *et al.*'s theory [9]. $Shame_i^T(j{:}\alpha)$ is read "shame about action α of agent j is triggered for agent i". It is logically defined as the fact that: i perceives a performance of an action α by agent j; this action α is blameworthy from i's point of view; and i *identifies* with agent j (Ortony *et al.* talk of "cognitive unit"). They only consider shame about actions of others, and view this emotion as a kind of moral disapprobation about these actions. We believe that this is only a very specific kind of shame. On the contrary our formalisation does not consider any responsibility but only the resulting state (of possible actions of i, j, or any other agent). Moreover, Steunebrink *et al.*'s formalisation does not allow to differentiate between shame in front of oneself and shame in front of others. It seems that in this case, agent i is always ashamed in front of itself (agent j does not play the same role as in our definition, but is just the author of action α).

Shame has been formally investigated by Turrini *et al.* in [23] that is also based on Castelfranchi *et al.*'s work on shame and guilt. They aim to formalise shame and guilt and their associated coping strategies. Our goal is less ambitious than theirs and we have only focused here on appraisal conditions. Following Castelfranchi, in [23] the main appraisal condition of shame is based on the fact that an ashamed agent has "the belief of not having had a capacity to get over a bad state" (see p. 406). We do not agree with this requirement for two reasons.

First, in [23], this *incapacity to get over a bad state* is taken into account in the appraisal conditions of shame by the fact that the agent that is ashamed "believes that there was no good alternative" (p. 413). This fact is formalised as: the agent believes that, after every other action that the agent could have performed instead of the action that he has really performed, the result would have been the same bad state. But it is too strong: experimental results (see [2] for instance) show that *capacity* may also concern willingness: the agent could have prevented this bad effect but he has not had the willingness, the moral strength, for not performing what he performed. Following the example of Clinton and Monica analysed in [23], Clinton may be ashamed about what he did with Monica, but he cannot say that he could not have prevented what happened in the sense that, whatever action he could perform, this action would have led to a state of the world where he had had an inappropriate relationship with Monica. It seems more intuitive to consider that he believes he has not had the moral strength to get over a bad state (whereas, from the point of view of the possible actions, there existed other actions that, if performed, would not have led to the current situation).

Second, from our point of view and related to the first point, the belief that we could not have prevented the current bad state of affairs seems to be more a consequence of shame, a kind of coping strategy, rather than an appraisal condition in itself. Clinton (physically) could have done otherwise, but he did not have the (moral) strength to avoid doing what he should not have done (or to do what he should have done). (One often tries to explain, or rationalise, what one did while one should not have.) Thus, for these two reasons, we do not have this condition in our formalisation. In this sense, we are closer to [5] where this condition is not required either.

6 Conclusion

We showed that shame is a complex emotion but can still be formalised. One can feel shame in front of oneself or other people. Shame involves both strong ideals (generally the necessity not to lose face in front of people whose opinion matters) and a causality between a given situation and its impact on self-image. Contrarily to guilt where one feels responsible for having (or not having) done a certain action, shame does not require this condition (even though it can hold in certain cases). Thus one can feel shame for one's hair colour or for one's origins without feeling responsible for these.

As we said at the start of this paper, shame is a powerful mediator of our social behaviour. Numerous studies, notably in the field of game theory in economy, show that individuals can be more or less sensitive to the feelings of shame and guilt (they talk about "guilt aversion" or "shame aversion"). One of the central aspects to handle this problem is the "action tendency" component. The sequel of this work will consist in introducing physical actions and rules of the type: if agent i has shame-aversion and he believes that after some facts are revealed he will feel shame, then he will adopt the goal to make one of the conditions of this shame false, for example by preventing these facts from being revealed.

We can formally illustrate this point on the example 5 if we suppose that Tom has shame-aversion. Since: (1) on the one hand he believes his friends will come visit his room (and discover it is untidy), and (2) on the other hand he does not know if this untidiness is a criterion of immaturity for *Lila* (by whom he wants to be positively evaluated), we could assign Tom an anticipation behaviour where he would for example clean his room to avoid shame. Another person, having a weaker shame aversion, could be more optimistic and bet on the fact that Lila would not consider untidiness as a sign of immaturity. Finally, someone having very little (or no) shame-aversion could accept to feel ashamed in front of Lila for the untidiness of his room. This aspect will constitute the next step of our research.

References

1. Elster, J.: Alchemies of the Mind: Rationality and the Emotions. Cambridge University Press, Cambridge (1999)
2. Tangney, J.P., Dearin, R.L.: Shame and Guilt. The Guilford Press (2002)
3. Scherer, K.: Emotion as a multicomponent process: a model and some cross-cultural data. Review of personality and social psychology **5** (1984) 37–63
4. Lorini, E.: A Dynamic Logic of Knowledge, Graded Beliefs and Graded Goals and Its Application to Emotion Modelling. In van Ditmarsch, H., Lang, J., Ju, S., eds.: Proceedings of the LORI-III Workshop on Logic, Rationality and Interaction, Guangzhou, P.R.China, 10/10/2011-13/10/2011. Volume 6953 of LNAI., Springer-Verlag (2011) 165–178
5. Castelfranchi, C., Poggi, I.: Blushing as a discourse: Was darwin wrong? In Crozier, W.R., ed.: Shyness and Embarrassment. Cambridge University Press (1990) 230–251
6. Tangney, J.P., Miller, R.S., Flicker, L., Barlow, D.H.: Are shame, guilt, and embarrassment distinct emotions? Journal of Personality and Social Psychology **70**(6) (1996) 1256–1269
7. Tangney, J.P.: The self-conscious emotions: shame, guilt, embarrassment and pride. In Dalgleish, T., Power, M., eds.: Handbook of Cognition and Emotion. John Wiley & Sons (1999)
8. Lazarus, R.S.: Emotion and Adaptation. Oxford University Press (1991)
9. Ortony, A., Clore, G., Collins, A.: The cognitive structure of emotions. Cambridge University Press, Cambridge, MA (1988)

10. Lewis, H.B.: Shame and guilt in neurosis. International Universities Press, New-York (1971)
11. Turrini, P., Meyer, J.J.C., Castelfranchi, C.: Rational agents that blush. In Paiva, A., Prada, R., Picard, R., eds.: ACCI 2007. Volume 4738 of LNCS. Springer (2007) 314–325
12. Conte, R., Castelfranchi, C.: Cognitive and social action. London University College of London Press, London (1995)
13. Searle, J.: Rationality in Action. MIT Press, Cambridge (2001)
14. Cohen, P.R., Levesque, H.J.: Intention is choice with commitment. Artificial Intelligence Journal **42**(2–3) (1990) 213–261
15. Castaneda, H.N.: Thinking and Doing. D. Reidel, Dordrecht (1975)
16. Chellas, B.F.: Modal Logic: an Introduction. Cambridge University Press, Cambridge (1980)
17. van Ditmarsch, H.P., van der Hoek, W., Kooi, B.: Dynamic epistemic logic. Kluwer Academic Publishers (2007)
18. Guiraud, N., Herzig, A., Lorini, E.: Speech acts as announcements (Dagstuhl Seminar on Information processing, rational belief change and social interaction, Dagstuhl, Germany, 23/08/2009-27/08/2009). Science Publications, Dagstuhl Seminar Proceedings 1862-4405, (en ligne). Also presented at LSIR-2 (workshop at IJCAI 2009) (2009)
19. Pörn, I.: Action theory and social science. Synthese Library. Kluwer Academic Publishers, Dordrecht, Holland (1977)
20. Steunebrink, B., Dastani, M., Meyer, J.J.: The OCC model revisited. In Reichardt, D., ed.: Proc. of the 4th Workshop on Emotion and Computing. (2009)
21. Adam, C., Herzig, A., Longin, D.: A logical formalization of the OCC theory of emotions. Synthese **168**(2) (2009) 201–248
22. Steunebrink, B.R., Dastani, M., Meyer, J.J.C.: A formal model of emotion triggers: an approach for bdi agents. Synthese **185** (2012) 83–129
23. Turrini, P., Meyer, J.J.C., Castelfranchi, C.: Coping with shame and sense of guilt: a dynamic logic account. Journal of Autonomous Agents and Multi-Agent Systems **20**(3) (2010) 401–420