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. 314–325, 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]).