Other-Condemning Moral Emotions: Anger, Contempt and Disgust

MEHDI DASTANI and ALEXANDER PANKOV, Utrecht University, The Netherlands

This article studies and analyzes three other-condemning moral emotions: anger, contempt, and disgust. We utilize existing psychological theories—appraisal theories of emotion and the CAD triad hypothesis—and incorporate them into a unified framework. A semiformal specification of the elicitation conditions and prototypical coping strategies for the other-condemning emotions are proposed. The appraisal conditions are specified in terms of cognitive and social concepts such as goals, beliefs, actions, control and accountability, while coping strategies are classified as belief-, goal- and intention-affecting strategies, and specified in terms of action specifications. Our conceptual analysis and semiformal specification of the three other-condemning moral emotions are illustrated by means of an example of trolling in the domain of social media.

CCS Concepts: • **Computing methodologies** → **Intelligent agents**; *Knowledge representation and reasoning;*

Additional Key Words and Phrases: Moral emotions, cognitive models, ontology of emotions

ACM Reference Format:
Mehdi Dastani and Alexander Pankov. 2017. Other-condemning moral emotions: Anger, contempt and disgust. ACM Trans. Internet Technol. 17, 1, Article 4 (January 2017), 24 pages.
DOI: http://dx.doi.org/10.1145/2998570

4

1. INTRODUCTION

There are many reasons emotions in general, and moral emotions in particular, play an important role in rational behavior [Elster 1994; Sloman and Croucher 1981], healthy mental life [Watkins 2008], and maintaining social and moral norms within societies [Gewirth 1981; Prinz 2007; Blackburn 1998]. Emotion is generally thought to be a (cognitive) mechanism that directs one's thought and attention to what is relevant, important, and significant to ensure effective behavior. While emotions in general are concerned with the individual's interest and behavior, moral emotions are concerned with the interest or welfare of other agents or society as a whole [Haidt 2003]. According to Rozin et al. [1999], moral emotions are triggered by the violation of moral norms and motivate morally congruent behavior [Haidt 2003; Vélez García and Ostrosky-Solís 2006]. Moreover, as emphasized by Gewirth [1981], the main characteristic of a moral norm is that it must bear on the interests or welfare of either society as a whole or individuals other than the agent itself. Therefore, moral emotions are viewed as having two prototypical features: disinterested elicitation conditions (the triggering event is related to other agents or the society as a whole) and pro-social action tendencies (benefiting others or the social order). Moral emotions differ in valence (positive vs. negative) and attributed accountability (self vs. other). Based on these differences, one can identify four families of moral emotions: other-condemning (contempt, anger, and

Authors' addresses: M. Dastani and A. Pankov, Department of Information and Computing Sciences, Utrecht University, Princetonplein 5, 3584 CC Utrecht, The Netherlands; emails: M.M.Dastani@uu.nl, A.Pankov@students.uu.nl.

DOI: http://dx.doi.org/10.1145/2998570

disgust), self-conscious (shame, embarrassment, and guilt), other-suffering (compassion), and other-praising (gratitude and elevation) [Haidt 2003]. In our article, we focus only on the other-condemning moral emotions, the type of moral emotions representing negative feelings about the actions or character of others [Haidt 2003].

Although there have been many efforts in the artificial intelligence community to provide a precise specification of emotions [Adam 2007; Adam et al. 2009; Battaglino et al. 2013; Dastani and Meyer 2006; Dastani and Lorini 2012; Gratch and Marsella 2004; Lorini 2011; Lorini and Schwarzentruber 2011; Lorini and Mühlenbernd 2015; Turrini et al. 2010; Steunebrink et al. 2009], there has not been, to our knowledge, a precise specification dedicated to other-condemning moral emotions and their role in dealing with moral transgressions. Other work on moral emotions focuses on other categories than other-condemning moral emotions. For instance, Lorini and Mühlenbernd [2015] study the moral emotion of guilt from the game-theoretic perspective and do not consider the other-condemning moral emotions as we do here. Section 6 provides a more detailed discussion on the similarities and differences between our approach and some existing work on moral emotions.

The aim of our work is to propose a semiformal but precise specification of the appraisal and coping processes involved in the following other-condemning moral emotions: anger, disgust, and contempt. We focus on these three emotions due to their overtly social nature (being concerned with the actions of other agents) and, as a consequence, their potential to influence others' behavior. The specifications can be used to shed light on the rationality and predominance of cooperative, morally congruent behavior. As will be argued, coping with other-condemning moral emotions may affect the adoption of goals that promote responding to the violation of moral norms. We illustrate the specification based on an example from social media. The reason to choose an example from social media is not only the focus of this special issue on the role of emotions in social media but also because we believe that social media is an interesting and unique public virtual environment that provides real data to analyze the role of emotions in human behavior.

The proposed semiformal specification paves the way toward building emotionally aware software agents that operate in multiagent settings. Such software agents have applications ranging from improving education in virtual environments to social media analysis, and from building believable video game characters in both entertaining and serious games to developing multirobot applications. We believe that the interactions among humans, virtual characters, or robotic systems in these applications could be regulated and improved by integrating other-condemning moral emotions in their decision-making modules. Moreover, the semiformal specification allows us to analyze how human subjects may experience emotions and how their mental structures change as a consequence. The semiformal specification enables researchers to disambiguate informal emotion theories and to simulate hypothetical situations (which would otherwise be morally impossible) in order to analyze complex psychological processes such as aggression and depression. Finally, the proposed specification is the first step toward a logical formalization of these emotions.

Our specification of other-condemning moral emotions will be in the spirit of dynamic [Fischer and Ladner 1979] and belief–desire–intention (BDI) [Cohen and Levesque 1990; Rao and Georgeff 1991] models of agency. In particular, it specifies a cognitive model of intelligent agents, capable of experiencing and coping with socially grounded emotions. The main theoretic and empirical support for our specification is from cognitive psychology, more specifically, appraisal and coping theories of emotion [Lazarus and Folkman 1984; Frijda 1986; Ortony et al. 1990; Lazarus 1991; Scherer 2001], as well as the CAD triad hypothesis [Rozin et al. 1999; Haidt 2003]. They have shown support in explaining the relationship between moral norms and emotions [Staller and Petta 2001] and will now be applied to the domain of behavior triggered by moral

emotions. According to these theories, the essential relationship between moral emotions and behavior is in the content of the agent's attitudes underlying the emotion. Different categories of attitudes lead to different emotions and behaviors. This matches perfectly with the BDI paradigm of modeling intelligent agents as entities possessing (uncertain) beliefs about the world and aiming at a desirable state of affairs by means of deliberation and action.

This article builds on and extends our previous work on other-condemning moral emotions [Pankov and Dastani 2015]. In fact, this article extends our previous work by proposing a coping mechanism that specifies how emotions in general and other-condemning moral emotions in particular trigger coping strategies based on action specifications. Furthermore, the number of analyzed coping strategies has been increased to capture some of the different flavors of the other-condemning moral emotions. Finally, the appraisal specification of other-condemning moral emotions has been modified and improved to capture the reviewed literature on emotion.

The structure of this article is as follows. In Section 2, we discuss the overall idea of appraisal and coping in other-condemning moral emotions. Then, in Sections 3, 4, and 5, we provide a detailed description of three moral emotions, that is, anger, disgust, and contempt, together with a semiformal specification of their elicitation and common coping strategies. In Section 6, we provide comparison between our work and some of the existing literature on topics along the same lines. Finally, Section 7 concludes our contribution with some remarks and future directions for further research.

2. APPRAISAL AND COPING IN OTHER-CONDEMNING MORAL EMOTIONS

It is generally believed that the main trigger for the elicitation of an other-condemning moral emotion is a moral transgression [Rozin et al. 1999; Haidt 2003; Vélez García and Ostrosky-Solís 2006]. In this section, we describe a psychological mechanism for emotional response to moral transgressions, in which appraisal and coping play an important role.

The basic premise of appraisal theories is that the agent's evaluation, his or her *appraisal*, of its cognitive condition plays an important role in emotion elicitation and differentiation [Scherer 2001]. Most theorists include goal relevance, agency, novelty, certainty, predictability, and compatibility with social standards to be some of the important appraisals to consider when studying emotion. Appraisal theories then postulate that emotions can be explained by such simpler but still meaningful elements. Once elicited, an emotion attracts the agent's attention and affects his or her behavior and mental attitudes by triggering a *coping* mechanism. Richard Lazarus defines coping as "constantly changing cognitive and behavioral efforts to manage specific external and/or internal demands that are appraised as taxing or exceeding the resources of the person" [Lazarus and Folkman 1984, p. 141]. Note that throughout the article, we prefer using the term *coping strategies* [Lazarus and Folkman 1984] instead of *action tendencies* [Frijda 1986] to denote responses that are promoted by the elicited emotions, although in most of the literature the two have been used interchangeably. The reason for this choice is the deliberative nature of the coping process, which gives it higher potential in modeling different behaviors. Furthermore, emotions in general, and other-condemning moral emotions in particular, motivate behavior in a rational and predictable manner. Coping strategies capture, we think, successfully this quality of emotions and give flexibility in explaining differences between other-condemning moral emotions. Such flexibility comes mainly from the distinction between belief-affecting, goal-affecting, and intention-affecting coping strategies (see Lazarus and Folkman [1984] for the similar, but not crisp, distinction between problem-directed and emotion-directed coping). As the names suggest, *goal-affecting* coping strategies modify the motivations of an agent, *belief-affecting* strategies modify the beliefs of an agent, and *intention-affecting* coping strategies modify the intentions (planned actions)

of an agent. Of course, in many cases a type of coping strategy actually leads to changes in more than one type of attitude. For instance, modifying one's beliefs can lead to also modifying one's goals. However, this does not hold necessarily all the time, and therefore the distinction between the different types of coping is useful in stressing the main tendency behind coping with a specific emotion type.

Many appraisal theorists see social standards as playing a key role in the elicitation of and coping with some emotions, and this observation has been extended to the moral domain. According to the CAD triad hypothesis [Rozin et al. 1999], the three other-condemning moral emotions are typically elicited, across cultures, by violations of three specific categories of moral norms: the *Ethics of Community, Autonomy, and Divinity* [Shweder et al. 1997]. Rozin et al. [1999] provide experimental evidence for this one-to-one correspondence, across cultures, between Shweder's ethics and the three emotions under discussion. Following the literature on moral emotions [Rozin et al. 1999; Haidt 2003; Vélez García and Ostrosky-Solís 2006; Lazarus 1991] and the relation emotions have to norms in human [Elster 1994, 2009; Bicchieri 2006] and artificial societies [Conte and Castelfranchi 1995], we propose a basic mechanism according to which the other-condemning moral emotions get elicited by violations of *internalized* moral norms, which involve inherently the interests of other agents or society as a whole. Moreover, depending on the category of the violated moral norm (e.g., community, autonomy, or divinity norms as distinguished by the CAD triad hypothesis), and thus the specific appraisals involved, different types of moral emotion are elicited (e.g., contempt, anger, or disgust as distinguished by the CAD triad hypothesis). Different coping strategies are then required depending on the specific moral emotion. In some cases the elicitation of moral emotions promotes a sanction-oriented behavior, for it alleviates the negative emotion by dealing directly with its external cause. In other cases, an internal reappraisal of the situation is promoted, for it alleviates the negative emotion by dealing with the agent's appreciation of the situation, modifying his or her beliefs and goals. The choice between these two variants depends on the specificity of the elicited emotions and the availability of resources for coping with it.

Further clarifications are due in order to make the previous picture complete. First, we need to be explicit in defining the conditions under which other-condemning moral emotions occur; that is, the general elicitation conditions of the emotions as well as the psychological appraisals involved in Shweder's ethics need to be specified. According to Shweder's ethics, the other-condemning moral emotions anger, contempt, and disgust are triggered by an agent's beliefs about others who are blameworthy of harm to others, violations of social rules, or contamination. We would like to emphasize that although we adhere here to Shweder's ethics, any distinction based on violations of moral norms will keep the overall emotion mechanism more or less intact. What will change are the types of concerns involved in the elicitation conditions of emotions. Second, we need to describe the coping strategies involved in the other-condemning moral emotions by specifying the coping strategies typical for the emotions under discussion.

It is important to stress here that we stay agnostic about the essence of moral norms or the process of their internalization (we point, however, to Dubreuil and Grégoire [2013] and Andrighetto et al. [2010] for a discussion on these topics). So, whether moral norms are originated from legal or social norms, or if they are formed based on some reasoning or deliberation process, if they are prohibitions or obligations, or how they become a motivational attitude upon which agents act, are not our concern in this article. What is of interest to us is their agreed-upon pro-social nature [Gewirth 1981] and their contents that shape the content of moral emotion and the corresponding coping strategies [Shweder et al. 1997; Rozin et al. 1999]; the rest remains out of scope for this work. We thus assume that, unlike legal or social norms for which an authority or social community is required to issue and monitor norms, moral norms

are essentially internalized by the agents. This view, which is in accordance with the distinction between social and moral norms proposed by Elster [2009] and Bicchieri [2006], suggests that there is no external authority required to issue and maintain moral norms. In some sense, it is rather the agent itself that issues moral norms by internalizing and acting upon them. This view allows us to model an internalized moral norm as a special type of maintenance goal, which motivates agents to act when they believe the norms are (due to be) violated. Thus, we do not assume an authority or component that models moral norms of an agent but consider internalized moral norms as the agent's goals. This may raise the question concerning the interaction between an agent's goals, which represent the agent's desires, and the agent's internalized moral norms, in particular, how possible conflicts between goals and internalized moral norms can be resolved. We believe that such conflicts can be resolved within an agent's deliberation process and by means of the agent's strategy/personality. For example, for moral agents, the activation of internalized moral norms overrules the activation of an agent's personal goal, while for an egoistic agent the order of overruling is reversed [Broersen et al. 2002]. Modeling an internalized moral norm as an agent's goal ensures that the violation of such a norm triggers other-condemning moral emotions in the agent, which in turn may trigger the agent to cope with the emotion. Note that this is in accordance with the Elster's view [Dubreuil and Grégoire 2013] to see emotions as a mechanism underlying norm compliance.

Let us first illustrate the three other-condemning moral emotions and their related coping strategies by means of a popular example from the domain of social media: *trolling*. Trolling, which is quite similar to flaming and cyberbullying, is often defined as a provocative behavior of posting inflammatory, offensive, or off-topic messages in social media. A troll is then the agent that performs such behavior. There are several recent studies from the psychological literature that provide insight on the cognition of a troll. First, a positive correlation between trolling behavior and personality traits such as sadism (strongest), psychopathy, and Machiavellianism have been shown [Buckels et al. 2014]. Some of these traits have been associated with inability or unwillingness to follow social norms [Cleckley 1964; Hare and Hart 1993]. Second, Johnson et al. [2009] have shown that there is a strong correlation between the inflammatory (flaming) nature of trolling and unfairness, harm, and anger. Finally, in popular culture, trolling is said to "promote antipathetic emotions of disgust and outrage" [Redmond 2014, p. 105]. From all these, we believe that trolling can serve as an interesting testbed for our study on moral emotions.

Example 1 (*Trolling in Social Media*). Imagine a participant in a social media discussion forum posting a comment on a given topic and receiving a trolling reply. In case the provocative trolling comment is an offense aimed at the person who posted the original comment, then one would not be surprised if some of the observing participants react with *anger*, verbally attacking the troll or reporting him or her to the site administrators to be banned. Similarly, if the trolling comment simply uses foul language without attacking someone in particular, one would expect the response of reporting or banning the *disgusting* offender, not trying to argue with him or her, as any such attempt might lead to more foulness. Finally, a trolling comment may not be offending but simply off-topic. In such case, banning seems quite harsh and a more *contemptuous* reaction of ignoring the comment can be expected. In these cases, trolling elicits in the participants who witness the trolling behavior an emotion condemning the behavior and leads to behavior that promotes the agreed-upon norms. In the rest of this article, we use the following names, propositions, and actions to formally describe this example scenario.

In addition to these elements and in order to analyze this example scenario, we will use some emotion-related propositions throughout the article. For now it suffices to

Agent names	*poster*	the agent who posts the comment
	troll	the troll agent
	obs	the observing participant
Proposition	*noOffLang*	no offensive language is used
	discussNoOff	the discussion proceeds without offenses
	discussNoFoul	the discussion proceeds without foulness
	discussOnTopic	the discussion proceeds according to its topic
Actions	*postComment*	post a comment
	offensiveComment	post an offensive comment
	foulComment	post a foul comment
	offTopicComment	post an off-topic comment

note that the previously described scenario starts with the *poster* agent who posts a comment (performs *postComment*), followed by the *troll* agent who posts a reply (either *offensiveComment*, *foulComment*, or *offTopicComment*).

In the next three sections, for each emotion in the other-condemning family, we first review the psychological literature on its elicitation conditions and typical coping strategies. We then analyze its moral flavor by identifying the content of the moral norm category being violated. Finally, we provide detailed definitions of the three other-condemning emotions and provide a semiformal specification of their elicitation conditions and coping strategies.

3. ANGER

The first to provide systematic treatment of anger, with surprisingly strong cognitive flavor, was Aristotle. In his *Rhetoric, Bk 2, Ch.2,* he writes: "Anger may be defined as a belief that we [...] have been unfairly slighted, which causes in us both painful feelings and a desire or impulse for revenge." His definition points out some key features: the negative nature of anger, its provocation by slight, and its motivational power for aggression.

3.1. Elicitation

In recent literature on emotion, anger has been viewed as the main motivator of aggressive behavior, and as triggered by the frustration or thwarting of a goal commitment (see Lazarus [1991]). In our trolling example, this amounts to saying that the original poster's wish to present and discuss his or her opinion without being offended has been thwarted by an offensive comment. This broad view has been refined by appraisal theories according to which *any* negative emotion can arise from goal incongruence; therefore, it is important to specify what makes the provocation of anger different from other negatively valanced emotional states, such as sadness, guilt, and remorse. To address this question, most appraisal theorists incorporate the agent's attribution of *blame* to another person [Lazarus 1991; Frijda 1986]. As a result, blame toward someone else becomes necessary for anger, for without the attribution of blame we can expect emotion such as sadness instead of anger; and with attribution of blame, but toward oneself, we can expect, for instance, guilt or remorse.

What does it mean, however, to blame someone for his or her deeds? According to Lazarus [1991], blame is an appraisal based on *accountability* and imputed *control*. To attribute accountability is to know who caused the relevant goal-frustrating event, and to attribute control is to believe that the accountable agent could have acted differently without, therefore, causing the goal incongruence. Therefore, to blame, instead of simply hold someone responsible, is to think that the blameworthy agent could have acted

otherwise. The difference is apparent in the case of trolling, where the person posting the offensive comment is to blame because he or she could, obviously, have refrained from commenting.

Attribution of blame is crucial to the elicitation of anger, but is that all there is to it? Lazarus argues that secondary appraisal plays a role as well: it "maximize[s] the possibilities of success" [Lazarus 1991, p. 227] when coping with the threatening situation and therefore influences which emotion gets elicited. According to him, (1) if *coping potential* (evaluation of the possibility to actualize personal commitments) favors attack as viable, then anger is facilitated; and (2) if future expectancy is positive about the environmental response to attack, then anger is facilitated. Similarly, Scherer [2001] writes about the coping ability of the agent in terms of an appraisal of power (availability of resources to act and anticipated effort) and adjustment ability (possibility/cost of changing/dropping goals). Both theorists seem to refer to the same mechanisms, which we will group under the title of coping potential, a type of secondary appraisal, to use Lazarus's term.

3.2. Coping

Most psychologists agree that the innate coping strategy in anger is *aggression* toward the blameworthy agent [Averill 1982, 1983]. Frijda seems to agree and calls the action tendency (coping strategy in his terms) underlying aggressive behavior "agonistic" [Frijda 1986, p. 88]. Supposedly, such behavior includes *attack* and *threat* as actions, with the goal being the removal of the obstruction that caused anger. However, secondary appraisal influences the selection of strategies of attack, and they can differ greatly in content [Lazarus 1991]. Furthermore, when planning an attack, the agent chooses between types of attack (e.g., verbal vs. physical, or punishment vs. warning) based on coping potential. For instance, in our trolling example, the participant's decision to attack the offender and report the post to an administrator could be based on his or her expectation of his or her inability to argue with the offender: an estimate of coping potential.

We can conclude that in most cases of anger, the applied coping strategy aims at attacking the cause of goal incongruence (intention-affecting coping) instead of reappraisal (belief-affecting or goal-affecting coping). The main reason for this seems to be the nature of anger: it gets promoted in cases when attack is viable and aggression is needed [Lazarus 1991].

3.3. Moral Anger

In many cases, anger is triggered by moral concerns and becomes an instance of a moral emotion. A main distinguishing factor between anger and moral anger is that moral anger typically involves a third party (e.g., an agent or the society as a whole) whose goal is threatened. For instance, consider a modified social media scenario where an agent posts a comment, which is believed by a second observing agent to be offensive to a third agent. In this case, the third agent can rightfully be angry at the first agent because of the appraised offense, without any of his or her moral views being offended. However, the second observing agent may get angry at the first agent because he or she believes some of his or her moral principles have been violated; that is, the second agent may believe that some of his or her internalized norms (e.g., no offensive language is used) are violated by the first agent. Of course, it may be possible that the second and third agents are one and the same, in which case the second/third agent is both angry and morally angry at the first agent because not only is he or she offended by the first agent but also one of his or her moral norms is violated.

Moral anger is a type of anger that arises when *harm* has been done to someone else whose rights have been violated [Prinz 2007]. The relationship between this definition and Shweder's ethics of autonomy has been demonstrated in Rozin et al. [1999] (as part

of the CAD triad hypothesis). As already mentioned in our discussion on the psychological mechanisms behind the moral emotions, Shweder's autonomy norms are best seen as norms pertaining to harm against persons. Shweder et al. [1997, p. 98] write: "The ethics of autonomy aims to [. . .] promote the exercise of individual will in the pursuit of personal preferences." Combining this aspect of moral anger with the elicitation conditions of core anger allows us to define moral anger in psychological terms:

Elicitation (moral anger). Displeasure from thwarting an internalized moral norm aimed at preserving the autonomy of agents, combined with attribution of blame for the goal-thwarting state of affairs to another agent and an estimate of one's own coping potential as favoring punishment of the blameworthy agent.
Coping (moral anger). Intention-affecting strategies aimed at sanctioning the blameworthy agent by means of attack or threat.

Due to the complex nature of moral anger, we first provide a semiformal specification of the appraisal of the core notion of anger in the next subsection. We then use this specification to present the semiformal specification of the appraisal of moral anger in the subsequent subsection.

3.4. Anger: Appraisal Specification

Assuming ϕ as a proposition that denotes a state of affairs, we use the special proposition $Control_i(\phi)$, which should be read as "agent i has control over the state of affairs ϕ." We define $Control_i(\phi)$ as there exists an action such that ϕ will be false after agent i executes the action, regardless the actions of other agents. In other words, agent i can make ϕ false on its own. An instance of the $Control_i(\phi)$ proposition is $Control_{troll}(discussNoOff)$, where *troll* denotes the agent from our trolling example and *discussNoOff* denotes the state of affairs where discussion has taken place without offenses. The proposition $Control_{troll}(discussNoOff)$ can thus be read as "*troll* agent has control over the discussion to proceed without offenses." Of course, we assume that the *troll* agent is the only one that can act as a troll. When other potential trolling agents exist, we should introduce various propositions (e.g., $discussNoOffTroll_1$, $discussNoOffTroll_2$, . . .) to denote different states of affairs ($discussNoOffTroll_i$ should be read as "discussion has taken place without offenses from $troll_i$ agent").

Moreover, we use special proposition $Account_i(a, \phi)$, which should be read as "agent i is accountable for the state of affairs ϕ by doing action a." So, we assume that the state of affairs ϕ is caused by action a and that other agents and the environment did not contribute in making ϕ true. In our running example, $Account_{troll}$ ($offensiveComment$, $\neg discussNoOff$) is an instance of this special proposition and should be read as "*troll* is accountable for not having a discussion without offenses by posting an offensive comment." Control and accountability, as defined here, are not viewed as epistemological but as ontological concepts representing causal relationships between events. It is their appreciation by an agent that provides the necessary inside on the agent's epistemic state, including his or her attribution of blame. Although similar concepts have been previously analyzed from a logical perspective [Lorini et al. 2013], here we only focus on their role in moral anger, disgust, and contempt.

We can now define blameworthiness, which will be represented by a special proposition $Blame_{i,j}(a, \phi)$. This proposition should be read as "agent i blames agent j for doing a and causing the state of affairs ϕ." The blameworthiness is defined as agent i believes that the state of affairs ϕ is true now, agent j is accountable for the state of affairs ϕ by doing action a, and before performing action a agent j had control over the state of

affairs $\neg\phi$. We specify blame in a semiformal manner as follows:

$$Blame_{i,j}(a,\phi) \stackrel{def}{=} Belief_i(\ \phi \wedge Account_j(a,\phi) \wedge [-a]_j Control_j(\neg\phi)\ ). \quad (1)$$

In this specification, $[-a]_j$ indicates "before performing action a by agent j."

As explained before, another appraisal condition for anger, also called secondary appraisal, is the possibility of coping potential; that is, the agent gets in moral anger if he or she has the practical possibility to remove the obstruction that caused the anger. This requires a way of specifying the practical possibility of an agent to realize a state of affairs. In our example, this can be understood as an observing participant being able to restore the no-offense nature of the discussion by, say, reporting the offender and leading to the removal of the offensive comment. For this purpose, we use a special proposition $Pos_i(\phi)$, which should be read as "there is a practical possibility for agent i to make the state of affairs ϕ true," and define it as there exists an action such that when it is performed by agent i, the state of affairs ϕ will be true. In our running example, proposition $Pos_{poster}(discussNoOff)$ indicates that the *poster* agent has the practical ability to realize a discussion without offenses.

We are now ready to specify the anger emotion, which will be represented by a special proposition $Anger_{i,j}(a,\phi)$ and read as "agent i is angry at agent j for doing a and preventing i from maintaining the state of affairs ϕ." In other words, agent i is angry at agent j because agent j has performed action a and thereby has prevented agent i from maintaining its desirable state of affair ϕ. We therefore define $Anger_{i,j}(a,\phi)$ as agent i has the maintenance goal ϕ (desirable state of affairs), blames agent j for performing action a and thereby preventing agent i from maintaining the desirable state of affairs ϕ, and believes there is still a practical possibility of getting back to the desirable state of affairs ϕ:

$$Anger_{i,j}(a,\phi) \stackrel{def}{=} Goal_i(\phi) \wedge Blame_{i,j}(a,\neg\phi) \wedge Belief_i(Pos_i(\phi)). \quad (2)$$

In this specification, we use proposition $Goal_i(\phi)$ to indicate that agent i desires to maintain the state of affairs ϕ and $Belief_i(Pos_i(\phi))$ to indicate that agent i believes he or she has the practical possibility of getting back to the state of affairs ϕ. According to this specification, thwarting goal ϕ, as expected for a negatively valanced emotion, is represented by the belief of agent i that ϕ does not hold, although agent i believes this was the case before action a was performed by agent j (see specification of blame). The belief of agent i about the practical possibility for realizing ϕ, which may not have been considered before, highlights the positive evaluation by agent i of his or her coping potential—the type of secondary appraisal claimed to be an indispensable part of anger.

Example 2 (*Anger*). Getting back to our running example, we assume that the *poster* agent performs *postComment* and that the *troll* agent replies with *offensiveComment*. We further assume the following facts to hold:

(1) $Goal_{poster}(discussNoOff)$: *poster* agent wants to have a discussion without offense.
(2) $Belief_{poster}(Pos_{poster}(discussNoOff))$: *poster* has practical possibility to have a discussion without offense.
(3) $Belief_{poster}(\neg discussNoOff \wedge Account_{troll}(offensiveComment,\neg discussNoOff) \wedge [-offensiveComment]_{troll} Control_{troll}(discussNoOff))$: *poster* agent believes that *troll* agent is accountable for the discussion with offenses (i.e., $\neg discussNoOff$) by posting *offensiveComment* and that before posting it he or she had control over the discussion to proceed without offenses.

Following the definition of blame from Equation (1), the third items can be reformulated as $Blame_{poster,troll}(offensiveComment,\neg discussNoOff)$, which means that

poster agent blames *troll* agent to prevent him or her from having a discussion without offenses by posting *offensiveComment*. Finally, following the definition of anger from Equation (2) and the aforementioned items, one can conclude $Anger_{poster,troll}(offensiveComment, discussNoOff)$, which means that *poster* agent is angry at *troll* agent for posting *offensiveComment* and preventing *poster* from having a discussion without offenses.

3.5. Moral Anger: Appraisal Specification

Proceeding to moral anger, we reassert that it is a flavor of anger with its content related to other agents and their autonomy. Autonomy was reduced to the exercise of individual choice in the pursuit of personal preferences. We surmise that the concept of *harm* captures this meaning: violating the autonomy of an agent means harming the agent. Although there are different types of harm distinguished in the literature [Ohbuchi et al. 1989; Helwig et al. 2001], what they all have in common is the violation of personal preferences by others. In case of physical harm, we can say the personal preference is for protecting one's own body, while in the case of psychological harm, the personal preference can be viewed as protecting one's beliefs. We use a special proposition $Harm_{i,j}(a, \phi)$, which should be read as "agent i harms agent j by doing action a and preventing i from maintaining the state of affairs ϕ," and define it as agent j having the goal ϕ and agent i being accountable for eliminating the practical possibility of agent j to maintain the state of affairs ϕ by doing action a:

$$Harm_{i,j}(a, \phi) \stackrel{def}{=} Goal_j(\phi) \wedge Account_i(a, \neg Pos_j(\phi)). \tag{3}$$

It should be noted that the accountability is defined with respect to the maintenance of a state of affairs. In the case of harm, this state of affairs is represented by $\neg Pos_j(\phi)$, denoting a state in which the possibility to maintain the state of affairs ϕ is eliminated. In our running example, the *troll* agent harms the *poster* agent by the action *offensiveComment* (i.e., action of posting an offensive comment) and preventing *poster* from having a discussion without being offended. This is represented by $Harm_{troll,poster}(offensiveComment, discussNoOff)$.

We now specify moral anger $MAnger_{i,j,k}(a, \phi, \psi)$, which should be read as "agent i is morally angry at agent j for harming k by doing action a and preventing agent k from maintaining his or her desirable state of affairs ψ, and thereby preventing agent i from maintaining its desirable state ϕ." In this reading, the desirable state of affairs ϕ for agent i is assumed to be related to the autonomy of other agents and created through a process of internalization of the moral norm concerning autonomy of agents. In the case of the trolling example, the internalization of moral norm concerning autonomy of agents by *obs* agent is assumed to create the goal of having no offensive language use and to relate this goal to any attempts for preventing *poster* agent from maintaining its goal to have a discussion without offenses. The moral anger is thus defined in semiformal representation as (1) $Anger_{i,j}(a, \phi)$—that is, agent i is angry at agent j for doing action a and thereby preventing him or her from maintaining his or her goal ϕ (which is created by the internalization of a moral norm), and 2) agent i believes $Harm_{j,k}(a, \psi)$ and $Harm_{j,k}(a, \psi) \rightarrow \neg\phi$. The second proposition specifies that agent i believes that agent j has harmed agent k by restricting his or her autonomy with respect to his or her goal ψ and, more importantly, agent i believes that this violation of k's autonomy is disadvantageous for him- or herself as this implies that i's goal (an internalized moral norm) ϕ is not maintained:

$$MAnger_{i,j,k}(a, \phi, \psi) \stackrel{def}{=} Anger_{i,j}(a, \phi) \wedge Belief_i(Harm_{j,k}(a, \psi) \wedge (Harm_{j,k}(a, \psi) \rightarrow \neg\phi)). \tag{4}$$

In this definition, the violation of autonomy of agent k by agent j is represented by $Harm_{j,k}(a, \psi)$, which implies the violation of the moral norm that is internalized by agent i as the goal ϕ. It is the violation of this internalized moral norm that makes agent i morally angry. The norm internalization here ensures that any violation of k's autonomy is experienced as a negative event for the morally angry i. Note that respecting agents' autonomy is one of the moral categories according to Shweder. It is also important to observe that what matters for the elicitation of moral anger is ϕ's relation to the autonomy of agents. It is this relation with the autonomy of agents that gives a moral accent to ϕ.

We would like to emphasize that in our semiformal presentation, we do not explicitly distinguish between goals that are created by the internalization of moral norms and goals that are originated from an agent's desires. Note, for example, that for the notion of core anger, the prevention of goals that are originated from desire may also be the cause of anger. However, for the notion of moral anger, we explicitly assume that the goal of the morally angry agent (i.e., goal ϕ of agent i) is created by the internalization of a moral norm. We did not introduce an explicit distinction between various types of goals in our semiformal presentation in order to keep the presentation simple.

We can see how the previous definition captures our analysis of the concept of moral anger, namely, as a type of anger with content related to harm done to someone else. Here the proposition $Harm_{j,k}(a, \psi)$ represents the autonomy aspect of moral anger, whereas $Harm_{j,k}(a, \psi) \rightarrow \neg\phi$ captures its negative consequences for the morally angry agent i. Two important points are due here. First, in our specification of anger and moral anger, as well as the other two emotions from the other-condemning family, we include the beliefs of agent i (e.g., the belief that harm has been done: $Belief_i(Harm_{j,k}(a, \psi))$). Of course, these beliefs do not have to be true in order for an emotion to be elicited. It is quite possible and natural, and in accord with appraisal theories of emotion, for an agent to be morally angry even though, and contrary to the agent's beliefs, there has not been any harm done. Second, it should be noted that our specification of moral anger concerns a basic notion, which can be further refined with, for example, the intentionality of the violator's action. Thus, if agent i believes that agent j has intentionally or knowingly has performed action a to harm agent k, then agent i can be said to be morally furious (or more intensely morally angry) at agent j. For the purpose of this article, we do not focus on such refinements of other-condemning moral emotions.

Example 3 (*Moral Anger*). As in Example 2, we assume *poster* agent performs *postComment*, *troll* agent replies with *offensiveComment*, and *obs* agent observes these two actions. We further assume the following facts to hold:

—$Goal_{obs}(noOffLang)$: *obs* agent wants to have no offensive language use.
—$Belief_{obs}(Pos_{obs}(noOffLang))$: *obs* agent believes it can maintain its goal.
—$Belief_{obs}(\neg noOffLang \wedge Account_{troll}(offensiveComment, \neg noOffLang) \wedge [-offensiveComment]_{troll} Control_{troll}(noOffLang))$: *obs* agent believes that *troll* agent is accountable for violating the maintenance of *noOffLang* by posting *offensiveComment* and that before posting it he or she had control over norm compliance (goal state).
—$Belief_{obs}(Goal_{poster}(discussNoOff))$: *obs* agent believes that *poster* agent wants to have a discussion without offenses.
—$Belief_{obs}(Account_{troll}(offensiveComment, \neg Pos_{poster}(discussNoOff)))$: *obs* agent believes that *troll* agent, by posting *offensiveComment*, is accountable for preventing *poster* agent of not having a discussion without offenses.

—$Belief_{obs}(Harm_{troll,poster}(offensiveComment, discussNoOff) \rightarrow \neg noOffLang)$: *obs* agent believes that restricting the autonomy of *poster* agent by *troll* agent has a negative implication for the maintenance of his or her goal *noOffLang*.

Like Example 2, from the first three items we conclude $Anger_{obs,troll}(offensiveComment, noOffLang)$, which means that *obs* agent is angry at *troll* agent for posting *offensiveComment* and violating its internalized norm *noOffLang*. Following the definition of harm from Equation (3), we conclude from the fourth and fifth facts $Belief_{obs}(Harm_{troll,poster}(offensiveComment, discussNoOff))$. From this and the sixth item, we can conclude $Belief_{obs}(Harm_{troll,poster}(offensiveComment, discussNoOff) \wedge (Harm_{troll,poster}(offensiveComment, discussNoOff) \rightarrow \neg noOffLang))$. Finally, following the definition of moral anger from Equation (4), we conclude $MAnger_{obs,troll,poster}(offensiveComment, noOffLang, discussNoOff)$, which means that *obs* agent is morally angry at *troll* agent for his or her post *offensiveComment* that prevented *poster* agent from having a discussion without offenses and thereby violating *obs*'s internalized norm *noOffLang*.

3.6. Moral Anger: Coping Specification

The elicitation of emotions in general, and moral emotions in particular, directs an agent's thought and attention to what is relevant and important to the agent. We interpret the focus of an agent's thought and attention as focusing on a subset of action repertoire that constitutes the agent's decision choices. An emotion can thus trigger a set of actions and thereby direct an agent's thought and attention. The coping strategy is then conceived as the mechanism that determines the subset of action repertoire for further deliberation and decision choices. It should be emphasized that different emotions can be elicited at the same time in an agent depending on the cognitive state of the agent. The subset of actions selected by the coping strategy should therefore be seen as possible strategies for all elicited emotions. We assume the dominant emotion that determines the further deliberation of an agent can be further specified by means of emotion intensity. In this article, we ignore this important aspect of emotion theory in order to simplify the presentation of our emotion specifications.

In order to specify the coping mechanism, we propose to specify the effect of an agent's action on its cognitive state and its environment. An agent's action can cause its beliefs, goals, intentions, or environment to change. In general, given ϕ as denoting the agent's mental state or the state of its environment, a set of actions Act, we write $cause(\alpha, \phi)$ to indicate that action $\alpha \in Act$ causes the realization of state ϕ. For example, $cause(remove_{obs}(troll), \neg Belief_{obs}(Harm_{troll,poster}(offensiveComment, discussNoOff)))$ indicates that removing *troll* agent by *obs* agent from a social media discussion forum ensures that *troll* agent cannot harm *poster* agent by offensive comments.

A coping strategy can then be specified by a set of rules that determine the triggering conditions for actions for some given emotions. The general form of such a rule is "An emotion triggers an action when the action has certain properties that influence the cause of the emotion." For example, the following rule specifies a coping strategy for agent i, which indicates that moral anger triggers an action α_i when α_i causes agent i not to believe that harm is done:

$$triggers(MAnger_{i,j,k}(a, \phi, \psi)\ ,\ \alpha_i) \text{ when } cause(\alpha_i, \neg Belief_i(Harm_{j,k}(a, \psi))). \quad (5)$$

Following this approach, the elicitation of moral anger—and anger in general—commonly leads to behavior targeted at resolving the psychological tension that triggered it. In our model, this amounts to an intention-affecting coping strategy aimed at removing the cause of moral anger. Therefore, for an agent i, we specify that coping with moral anger involves adopting the intention of performing an action α_i for which it

is known to make agent i not to believe $Harm_{j,k}(a, \psi)$. This way, successfully triggering the defined coping strategy removes the presence of moral anger—a property necessary for successful coping [Lazarus and Folkman 1984; Watkins 2008].

Example 4 (*Coping with Moral Anger*). In our running example, one should expect that morally angry *obs* agent initiates an attack behavior by removing offensive comments from the social media, banning the offender, arguing with the offender, or a combination of these actions toward *troll* agent. This way, the problem of harming *poster* agent will be mitigated by allowing the discussion to continue or defending *poster* agent. We assume the following facts hold:

(1) $MAnger_{obs,troll,poster}(offensiveComment, noOffLang, discussNoOff)$.
(2) $cause(remove_{obs}(troll), \neg Belief_{obs}(Harm_{troll,poster}(offensiveComment, discussNoOff)))$.

Following the triggering rule of moral anger from Equation (5), these two facts cause action $remove_{obs}(troll)$ to be triggered, the performance of which will ban *troll* agent from the social media. When there are more alternative facts of the form $cause(\alpha_i, \neg Belief_i(Harm_{j,k}(a, \psi)))$, we have more possible coping strategies. We assume that the order of these facts represents some kind of plausibility order such that action suggested by the first fact will be used as the actual coping response of *obs* agent.

4. DISGUST

Research on disgust has gained popularity since the 1990s with some of the main contributors being Paul Rozin and his colleagues [Rozin and Fallon 1987; Haidt et al. 1997; Rozin et al. 1999, 2008]. According to their theory, disgust has its evolutionary origins in helping people decide what to eat and is usually viewed as based on a *distaste* response found also in other animals [Rozin et al. 1999]. This evolutionary-old response has then been shaped by natural selection to become a more generalized "guardian of the temple of the body" [Rozin et al. 2008, p. 764]. In that context, distaste refers to the sensory-motor functions of smelling and tasting. Similar to anger, disgust has simpler (core disgust) and more complex (moral disgust) forms [Rozin et al. 2008].

4.1. Elicitation

At its core, disgust provides a mechanism for protecting against dangerous types of objects and behaviors: food, body products, animals, sex, death, body envelope violations, and bad hygiene [Haidt et al. 1997]. It is argued that what makes all these a concern for the agent are their dangerous physical products and their contradiction with the agent's motivation to protect oneself from *contamination*. Not surprisingly, Lazarus defines disgust as a negative emotion triggered by a "risk of being contaminated by a 'poisonous idea'" [Lazarus 1991, p. 260]. The logic of contamination is then expressed as the statement that an agent gets contaminated by coming into contact with another contaminated object or agent. Other authors describe similar elicitation conditions. For instance, in the OCC model, disgust is elicited by disliking an unfamiliar aspect (of an object) [Ortony et al. 1990], whereas Oatley and Johnson-Laird say that disgust gets triggered by a gustatory goal being violated [Oatley and Johnson-Laird 1996].

Furthermore, it is commonly accepted that during biological and cultural evolution, disgust expanded beyond its role as guardian of the body from contamination and became a suitable reaction both to physical objects and to social violations [Rozin et al. 2008; Ortony et al. 1990; Haidt 2003; Lazarus 1991]. Lazarus unites the physical and social aspects of disgust by referring to it as "taking in or being too close to an indigestible object or idea (metaphorically speaking)" [Lazarus 1991, p. 260]. This and other definitions [Ortony et al. 1990; Rozin and Fallon 1987] focus on the mouth and

dislike toward physical objects. They suggest that some class of nonphysical objects can also cause a similar feeling. Paul Rozin and his collaborators argue that disgust grew out of a distaste response to become coupled to a motivation to protect oneself from any sort of contamination, including contamination of ideas. Disgust plays a role also in sexuality analogous to its role in food selection by guiding people to the narrow class of culturally acceptable sexual partners and sexual acts [Haidt 2006]. As a consequence of its sensitivity to social violations, disgust is often recruited to support many of the norms, rituals, and beliefs that cultures use to define themselves [Haidt et al. 1997].

4.2. Coping

All forms of disgust include a motivation to avoid, expel, or otherwise break off contact with the offending entity, often coupled with a motivation to purify or otherwise remove residues of any physical contact that was made with the entity [Rozin et al. 2008]. This motivation is clearly adaptive when dealing with potentially lethal contamination of food, but it appears to have made the transition into our moral and symbolic life as well [Rozin et al. 2008]. Similarly, according to Frijda [1986], the typical action tendency of disgust is rejecting. One can conclude that coping with disgust usually requires intention-affecting strategies to realize the required result, which is purity. It involves behaviors such as expulsion, separation, and cleansing.

Although in most cases actions are required to deal with the feeling of disgust, it has been argued that disgust has also the ability to "extinguish desire" [Haidt 2006, p. 186]. For example, think of the effect a disgusting situation has on desires such as hunger or sexual drive. Generally speaking, this amounts to saying that goal-affecting strategies of reducing the strength of the violated goal are suitable for coping with disgust. Of course, in complex real-world scenarios, several types of coping strategies will typically function simultaneously.

4.3. Moral Disgust

The variation of disgust, called moral disgust, is triggered by people who violate local social rules for how to use their bodies, particularly in the domains of sex, drugs, and body modification [Haidt 2003]. Rozin and his colleagues have demonstrated that moral disgust derives from core disgust by showing that it has the same bodily basis and the same logic of contamination: we do not like to have contact with objects that have touched a person we deem morally disgusting [Rozin et al. 2008]. For example, we would not like to live in the former home of a condemned pedophile, or, following our running example, we would not like to argue with a person posting only comments containing foul language.

Furthermore, according to the CAD triad hypothesis [Rozin et al. 1999], we can make a link between disgust and Shweder's ethics of divinity: moral norms concerning the natural order. What follows is that disgust gets triggered by violations of such norms. In explaining the ethics of divinity, Shweder et al. [1997, p. 99] write: "The ethics of divinity protect the soul, the spirit, the spiritual aspects of the human agent and nature from degradation." Interestingly, none of the moral transgressions under the "divinity" label used in forming the CAD triad hypothesis [Rozin et al. 1999] have to do with religious violations. Thus, we conclude that the name of this category should not be taken literally; instead, it should be understood as referring to purity and the natural order of things—with the divine being an instance of the natural order. Our methodology, then, requires us to combine this result with the standard appraisal theory account of the elicitation and coping with disgust, resulting in the following definition:

Elicitation (moral disgust). Displeasure from the thwarting of an internalized moral norm aimed at protecting against contamination, including contamination of ideas.

Coping (moral disgust). Intention-affecting strategies aimed at avoiding, expelling, or otherwise breaking off contact with the offending entity; goal-affecting strategies aimed at extinguishing desire.

Due to the relatively simple nature of moral disgust, we present the semiformal specification of the appraisal for both the core notion of disgust and moral disgust in the following subsection.

4.4. (Moral) Disgust: Appraisal Specification

In this section, we specify the appraisal process for moral disgust. Similar to our discussion on anger, we use primitive concepts such as agent goals, beliefs, and actions, together with the concept of accountability. The difference, compared to anger, will be the introduction of the special atoms C_i, which should be read as "agent i is contaminated." We assume that this special atom holds whenever some of i's moral norms concerning the natural order is violated.

We first specify the core notion of disgust, which will be represented by a special proposition $Disgust_i(\phi)$ and read as "agent i is disgusted by ϕ." We define $Disgust_i(\phi)$ as agent i has the goal $\neg\phi$ (ϕ representing an undesirable state of affairs), believes ϕ holds, and believes C_i holds if ϕ holds:

$$Disgust_i(\phi) \stackrel{def}{=} Goal_i(\neg\phi) \wedge Belief_i(\phi \wedge (\phi \rightarrow C_i)). \tag{6}$$

In this specification, goal $\neg\phi$ is considered to be a desirable state of affairs for agent i. Thwarting this goal, as expected for a negatively valanced emotion, is represented as the belief of agent i in ϕ. Finally, his or her belief in $\phi \rightarrow C_i$ captures the property of disgust of being about a state of the world that the agent believes to cause contamination.

As was the case with anger and its moral flavor, we specify moral disgust as a type of disgust, triggered by the actions of others, and represent it with the special proposition $MDisgust_{i,j}(a, \phi)$. The formula $MDisgust_{i,j}(a, \phi)$ should be read as "agent i is morally disgusted by agent j doing a and causing ϕ." In accordance with this reading, we define $MDisgust_{i,j}(a, \phi)$ as agent i is disgusted by ϕ and believes agent j is accountable for ϕ by performing action a:

$$MDisgust_{i,j}(a, \phi) \stackrel{def}{=} Disgust_i(\phi) \wedge Belief_i(Account_j(a, \phi))). \tag{7}$$

Here, due to the generality of the definition, there is no need for specifying a third agent, as we did with moral anger: ϕ from the definition can describe a contaminating contact with an agent as well as a physical state. The moral flavor of the emotion remains, for it is concerned with the behavior of agent j. We would like to emphasize that the goal ϕ in this specification is assumed to be created by the internalization of a moral norm. As was the case with moral anger, we will not explicitly distinguish goals that are created by the internalization of norms and goals that are originated from an agent's desires.

Example 5 (*Moral Disgust*). Applying the previous definition to our running example should clarify the idea behind moral disgust. If the trolling comment from the example contained language utterances considered foul (dirty) by some participant, he or she is expected to be disgusted by it. Here again we assume that *poster* performs *postComment*, *troll* agent posts a nasty reply *foulComment*, and some agent, for

example, *poster* agent (or otherwise an arbitrary *obs* agent), reads the nasty reply. We also assume the following facts hold:

(1) $Goal_{poster}(discussNoFoul)$: *poster* agent wants to have a discussion without foul language use.
(2) $Belief_{poster}(\neg discussNoFoul)$: *poster* agent believes the discussion includes foul language use.
(3) $Belief_{poster}(\neg discussNoFoul \rightarrow C_i)$: *poster* agent believes a discussion with foul language use contaminates him or her.
(4) $Belief_{poster}(Account_{troll}(foulComment, \neg discussNoFoul))$: *poster* agent believes *troll* agent is accountable because the discussion includes foul language use by his or her *foulComment* action.

Following the definition of moral disgust from Equation (6) and based on the first three facts, we conclude $Disgust_{poster}(\neg discussNoFoul)$. This together with the fourth fact allows us to conclude $MDisgust_{poster,troll}(foulComment, \neg discussNoFoul)$, which means that *poster* agent is morally disgusted by *troll* agent due to his or her *foulComment*. In this case, the contamination is purely contamination of ideas, but this, as we stated in our informal discussion, is to be expected for the moral flavor of disgust.

4.5. Moral Disgust: Coping Specification

We said that the prototypical coping strategy when dealing with disgust is an intention-affecting strategy to try and expel the source of contamination. An agent i feeling disgust will try performing an action (e.g., expelling the source of contamination) if he or she thinks it will remove the cause of contamination. The following rule specifies such a coping strategy:

$$triggers(MDisgust_{i,j}(a, \phi),\ \alpha_i) \text{ when } cause(\alpha_i, \neg Belief_i(\phi)), \tag{8}$$

where the performance of α_i is assumed to make i believe that ϕ is not the case anymore.

Example 6 (*Coping with Moral Disgust First Alternative*). In our trolling example, one should expect behavior that restores the nonfoul nature of the social media discussion, by either removing the trolling comment or preventing further contamination by banning/reporting the offending agent, but not, for instance, arguing with him or her, for this will only cause further contamination. We assume the following facts hold:

(1) $MDisgust_{obs,troll}(foulComment, \neg discussNoFoul)$.
(2) $cause(removeComment_{obs}(troll), \neg Belief_{obs}(\neg discussNoFoul))$.

Following the triggering rule in Equation (8), these two facts lead to the triggering of the action $removeComment_{obs}(troll)$. The performance of this action will remove the foul reply of *troll* agent and therefore make *obs* agent not believe that the discussion includes foul comments.

A second possibility we identified for coping with disgust was a goal-affecting strategy aimed at reducing the strength of the thwarted goal. In such case, an agent i will update his or her goal set by reconsidering goal ϕ:

$$triggers(MDisgust_{i,j}(a, \phi)\ ,\ \alpha_i) \text{ when } cause(\alpha_i, \neg Goal_i(\neg\phi)), \tag{9}$$

where $cause(\alpha_i, \neg Goal_i(\neg\phi))$ indicates that performing epistemic action α_i reconsiders the agent's goal $\neg\phi$. An example of α_i with the property to reconsider goals is the action of dropping a goal. On an account where goals have desirability levels, to reconsider goal ϕ can also refer to an action that reduces the desirability level of ϕ, without

completely removing it from the agent's goal set. It should be noted that in this case the action α_i is categorically different to the expel action previously. Here the action is purely epistemic and aimed at affecting the mental state of the agent by modifying the agent's goal base.

Example 7 (*Coping with Moral Disgust Second Alternative*). The goal-affecting strategy specified here can be understood in terms of our running example. In case of a trolling comment using foul language utterances, one could expect that the original poster re-evaluate his or her goal to have a discussion without foul comments as a less important goal in order to disengage from the discussion and, therefore, protect him- or herself from future exposure to foulness. We assume the following facts hold:

(1) $MDisgust_{poster,troll}(foulComment, \neg discussNoFoul)$.
(2) $cause(deleteG_{poster}(discussNoFoul), \neg Goal_{poster}(discussNoFoul))$.

Following the triggering rule in Equation (9), these two facts lead to the triggering of the action $deleteG_{poster}(discussNoFoul)$. The performance of this action will drop the *poster*'s goal of having a discussion without foul comments.

Moreover, it should be clear that the two strategies for coping with moral disgust specified earlier are also applicable to disgust emotion as well. The reason, of course, is that moral disgust is a type of disgust and thus triggers the strategies for coping with the core disgust. As a consequence, coping with the core disgust behind moral disgust alleviates the feeling of moral disgust as well.

Additionally, moral disgust allows for a third type of coping strategy that is worth exploring briefly, namely, one that affects the belief behind moral disgust and revises its beliefs accordingly:

$$triggers(MDisgust_{i,j}(a, \phi),\ \alpha_i) \text{ when } cause(\alpha_i, \neg Belief_i(Account_j(a, \phi))), \qquad (10)$$

where the performance of α_i causes agent i not to believe, or otherwise reduce the belief, that agent j is accountable for the disgusting state of affairs ϕ. Here again α_i is an epistemic action that affects the beliefs of agent i by making him or her reconsider his or her belief regarding who is accountable for realizing state ϕ. Such a strategy should be viewed as a type of reconsideration/reappraisal of the situation on the side of the agent. On an account with graded beliefs, the property of reconsidering beliefs can also be defined as reducing the level of the agent's belief of the cause for ϕ.

Example 8 (*Coping with Moral Disgust Third Alternative*). The belief-affecting strategy specified here can be understood in terms of our running example. In case of a trolling comment using foul language utterances, one could expect that the original poster revise his or her beliefs regarding the accountability of *troll* agent. We assume the following facts hold:

(1) $MDisgust_{poster,troll}(foulComment, \neg discussNoFoul)$.
(2) $cause(deleteB_{poster}(Account_{troll}(foulComment, \neg discussNoFoul)),$
$\neg Belief_{poster}(Account_{troll}(foulComment, \neg discussNoFoul)))$.

Following the triggering rule in Equation (10), these two facts cause action $deleteB_{poster}(Account_{troll}(foulComment, \neg discussNoFoul))$ to be triggered. The performance of this action will delete the belief of *poster* agent regarding the accountability of *troll* agent.

5. CONTEMPT

Contempt is one of the least discussed emotions in the psychological literature [Haidt 2003, Table 1]. If research on the facial expression of contempt is excluded, there is almost no other empirical research on contempt. In most discussions, it falls in between

anger and disgust, and is sometimes said to be a blend of the two [Plutchik 1980], folded into the anger family [Lazarus 1991], or else said to be part of anger [Ortony et al. 1990]. Here, however, it is discussed separately because of its important role as the only moral emotion from the other-condemning family not having a core/immoral variant: all instances of contempt are triggered by violations of moral norms related to conforming to social hierarchies.

5.1. Elicitation

For our discussion, we adopt the view that contempt is part of the reproach emotions family and is elicited by disapproving of someone else's *blameworthy action* [Ortony et al. 1990, p. 145]. This is quite similar to what we said about the triggering conditions of anger. This is also the reason anger's elicitation condition is seen as a blend between those of a reproach emotion (such as contempt) and a negative event-based emotion (such as distress) [Ortony et al. 1990]. In this work, it is emphasized that anger is not a compound emotion; instead, its elicitation conditions have an overlap with those of distress and contempt.

As stated in the introduction, there is evidence [Rozin et al. 1999] for the relation between contempt and violations of Shweder's ethics of community [Shweder et al. 1997]. Shweder writes [Shweder et al. 1997, p. 98]:

> The ethics of community [...] aims to protect the moral integrity of the various stations or roles that constitute a society or community.

The main concepts discussed in Shweder et al. [1997] regarding the ethics of community are those of *hierarchy* and *duty*. We consider hierarchy and duty to be specified by a set of social roles. Violations of hierarchy and duties are then viewed as violations of the required, by these social roles, behavior. Such an abstraction, we think, covers the basic idea behind hierarchy and duty, and can be used to specify contempt. For example, when participating in social media discussions, one can distinguish two roles: the poster of the original comment and the participants. Their relationship (in terms of hierarchy and duties) can then be captured by a mechanism that indicates whether some behavior (e.g., posting off-topic comments or replying in a different language by the participants) is a violation of the required behavior. Furthermore, we introduce the concept of *significant others* [Higgins 1987] and take it to define a kind psychological attachment between agents. We use the concept of significant others as a constraint on the scope of contempt: an agent is contemptuous only toward agents that are "significant others" to him or her.

5.2. Coping

Contempt motivates neither attack nor withdrawal; rather, it seems to cause social-cognitive changes such that the object of contempt will be treated with less *warmth*, *respect*, and *consideration* in future interactions [Oatley and Johnson-Laird 1996]. There is a lot one can say about these concepts, but we only stipulate that warmth, respect, and consideration all supervene on the perceived significance of the other agent. Thus, less (more) perceived significance means less (more) warmth, respect, and consideration in future interactions. As a result, all belief changes for coping with contempt become bound to reduction of the level of belief in the "social significance" of the other agent. In our running example, this would amount to saying that in response to an off-topic comment by a participant agent (in this case *troll* agent), other participants (e.g., *poster* agents) will change their appreciations of the importance that the participant (i.e., *troll* agent) has to them.

As with disgust, contempt suggests a belief-affecting coping strategy. This makes contempt significantly different than moral anger and, to some extent, moral disgust, both of which had intention-affecting strategies for coping as well. This results in the following description of the emotion contempt:

Elicitation (contempt). Displeasure from the thwarting of an internalized moral norm aimed at preserving the social hierarchy and duty, combined with the attribution of blame to a significant other.
Coping (contempt). Belief-affecting strategies for changing the personal significance of the blameworthy agent.

5.3. Contempt: Appraisal Specification

As usual, we will specify contempt using the primitive concepts of beliefs, goals, and actions. Additionally, we will use the special atoms V_i to denote that agent i violates the behavior that is required by hierarchy and duties, and $Sig_{i,j}$ to denote that agent j is significant to agent i. As stated earlier, contempt is a negative emotion triggered by violation of a goal (internalized norm) concerned with preserving hierarchy and duty, together with the attribution of blame for the goal-thwarting state of affairs to a significant other. The appraisal of blame has already been defined in Section 3.4 and can be used directly. Preserving hierarchy and duty will be modeled as a maintenance goal whose violation leads to neglecting hierarchy and duty.

We specify the contempt emotion, which will be represented by a special proposition $Contempt_{i,j}(a, \phi)$ and read as "agent i is contemptuous toward agent j for doing a and making ϕ false." We define $Contempt_{i,j}(a, \phi)$ as agent i has the maintenance goal ϕ, blames agent j for performing the action a and making ϕ false, and believes j to be a significant other ($Sig_{i,j}$) and that $\neg\phi$ violates hierarchy and duty required from agent j:

$$Contempt_{i,j}(a, \phi) \stackrel{def}{=} Goal_i(\phi) \wedge Blame_{i,j}(a, \neg\phi) \wedge Belief_i(Sig_{i,j} \wedge (\neg\phi \rightarrow V_j)). \quad (11)$$

In this specification, ϕ denotes the desirable state of affairs that agent i wants to maintain. As with moral anger and disgust, we assume ϕ to be a goal created by the internalization of a moral norm. Thwarting this goal, as expected for a negatively valanced emotion, is represented as the belief of agent i in $\neg\phi$. Finally, his or her beliefs in $Sig_{i,j}$ and $\neg\phi \rightarrow V_j$ capture the property of contempt that agent j is significant to agent i, and that j has violated behavior required by hierarchy and duty, respectively.

Here we should note the similarity of the previous definition to that of moral anger. Both include an appraisal of blame as well as a belief in a violation of a moral rule. In the case of anger, the violation of the autonomy of other agents has been modeled simply as causing harm, whereas in the case of contempt, the violation is represented by an externally defined criterion: a violation of the social hierarchy by a significant other.

Example 9 (*Contempt*). Similar to the previous two emotions, we assume *poster* agent performs *postComment* and that *troll* agent replies with *offTopicComment*. We further assume that the following facts hold:

(1) $Goal_{poster}(discussOnTopic)$: *poster* agent wants to have a focused discussion.
(2) $Blame_{poster,troll}(offTopicComment, \neg discussOnTopic)$: *poster* agent blames *troll* agent of thwarting his or her goal of have an on-topic discussion by posting an off-topic comment.
(3) $Belief_{poster}(Sig_{poster,troll})$: *poster* agent believes *troll* agent is significant to him or her.

(4) $Belief_{poster}(\neg discussOnTopic \rightarrow V_{troll})$: *poster* agent believes *troll* agent has violated the behavior required from participants by damaging the focus of the discussion.

Following the definition of contempt from Equation (11) and based on the previous four facts, we conclude $Contempt_{poster,troll}(offTopicComment, discussOnTopic)$, which means that *poster* agent is contemptuous toward *troll* agent for posting an off-topic comment and thwarting his or her maintenance goal to have a focused discussion.

5.4. Contempt: Coping Specification

Contempt has the interesting characteristic of affecting one's appreciation of the other agent's significance, without having direct influence on one's behavior [Oatley and Johnson-Laird 1996]. We can specify such a coping strategy as the contemptuous agent i reconsidering his or her belief by deleting (or reducing) the significance of j (i.e., his or her belief in the formula $Sig_{i,j}$):

$$triggers(Contempt_{i,j}(a, \phi),\ \alpha_i) \text{ when } cause(\alpha_i, \neg Belief_i(Sig_{i,j})), \tag{12}$$

where the performance α_i causes agent i not to believe, or otherwise reduce the belief, that agent j is significant. Note that, although removal of (or reduction in) the significance of the offending agent might also be possible when coping with anger and disgust, we think such strategy is essential for coping with contempt.

Once more, we have a coping strategy aimed at resolving the cognitive tension by promoting a special type of action α_i, which removes one of the emotion elicitors from the agent's beliefs. Trying out this definition in our example, we see its immediate logic: dealing with off-topic comments (the trigger of contempt) involves ignoring, instead of fighting, them by reducing the importance of the people making those comments.

Similarly to the other two emotions under discussion, it is possible to extend the coping specification of contempt by letting the coping strategy deal with different parts of the emotion elicitors. For instance, to address the blameworthiness of the agent:

$$triggers(Contempt_{i,j}(a, \phi)\ ,\ \alpha_i) \text{ when } cause(\alpha_i, \neg Belief_i(Account_j(a, \neg\phi))). \tag{13}$$

One can immediately see that by deleting the belief in $Account_j(a, \neg\phi)$, one of the conjuncts from the definition of contempt, namely, $Blame_{i,j}(a, \neg\phi)$, becomes false and, therefore, copes successfully with the feeling of contempt. Yet, another option would be to use a goal-affecting strategy for dropping the goal ϕ. Similarly to the case of disgust, this strategy will cope with the feeling of contempt by forcing the agent to reconsider the desirability of the state of affairs ϕ:

$$triggers(Contempt_{i,j}(a, \phi)\ ,\ \alpha_i) \text{ when } cause(\alpha_i, \neg Goal(\phi)). \tag{14}$$

The working of these coping strategies can be illustrated in our running example.

Example 10 (*Coping with Contempt*). The belief-affecting coping strategies for contempt can be to delete the significance of *troll* agent or his or her accountability. In particular, agent *poster* can delete his or her beliefs concerning the significance or accountability of *troll* agent. Moreover, the goal-affecting coping strategy can be to delete the goal to maintain having a focused discussion. We assume the following facts:

(1) $Contempt_{poster,troll}(offTopicComment, discussOnTopic)$
(2) $cause(deleteB_{poster}(Sig_{poster,troll}), \neg Belief_{poster}(Sig_{poster,troll}))$.
(3) $cause(deleteB_{poster}(Account_{troll}(offTopicComment, \neg discussOnTopic)),$
$\neg Belief_{poster}(Account_{troll}(offTopicComment, \neg discussOnTopic)))$.
(4) $cause(deleteG_{poster}(discussOnTopic), \neg Goal_{poster}(discussOnTopic))$.

Following the triggering rule in Equations (12), (13), and (14), these facts trigger the selection of actions $deleteB_{poster}(Sig_{poster,troll})$, $deleteB_{poster}(Account_{troll}(offTopicComment, \neg discussOnTopic))$, or $deleteG_{poster}(discussOnTopic)$, after which *poster* agent respectively does not believe *troll* agent is significant, does not believe *troll* agent is accountable for the unfocused discussion, or does not desire to maintain having a focused discussion.

6. RELATED WORK

The most important feature of the present analysis of other-condemning emotions is its multiagent flavor and its inclusion of coping strategies in analyzing moral emotions. Although the importance of coping in emotion has been stressed by appraisal theorists, most of the formal models in the literature have ignored it [Adam 2007; Adam et al. 2009; Dastani and Meyer 2006; Lorini 2011; Lorini and Schwarzentruber 2011; Steunebrink et al. 2009; Battaglino et al. 2013].

An inspiration for our current work has been Dastani and Lorini [2012], who propose a formal system in which emotion intensities based on belief and goal strengths can be modeled. Support of emotion intensities is an obvious advantage over our current work, but we have reasons to believe that extending it with means of talking about belief and goal strengths will straightforwardly allow for modeling emotion intensities much in the spirit of Dastani and Lorini [2012]. We have avoided doing this here, for it would have increased significantly the complexity and length of the article. Furthermore, the model in Dastani and Lorini [2012] is based on a single agent and does not offer means of talking about past events and their effects; therefore, it does not directly allow for modeling the other-condemning moral emotions as specified here.

Another influencing work on the topic has been Steunebrink et al. [2009]. Inspired by Frijda [1986], they provide a formal model of emotions extended with intensities and action tendencies. Steunebrink et al. [2009] take emotion intensity as primitive, without explaining how it depends on belief and goal strengths. Furthermore, Steunebrink et al. [2009] do not discuss moral emotions specifically and do not offer suggestions on how their "moral" flavor can be modeled in the offered framework.

We should also mention some of the work on modeling shame and guilt. We stress again that these two emotions are not part of the other-condemning family of moral emotions analyzed here. However, they share some characteristics specific to moral emotions in general, and therefore we discuss them here. The formal system proposed in Turrini et al. [2010] is based on modal logic and, similarly to our work, analyzes emotions from a multiagent perspective with concern for coping strategies and the attribution of social significance between agents. However, their model of coping strategies seems to be tailored to the two emotions of shame and guilt, which makes it difficult to extend to other moral emotions. We believe our analysis of coping emotions is quite generic and offers the possibility of capturing other emotion types.

Lorini and Mühlenbernd [2015] provide a game-theoretic analysis of guilt and, similarly to our work, argue for its relation with internalized moral norms. Although their setup is quite different than ours, which makes comparing the two difficult, we believe there are some important conceptual differences. For instance, Lorini and Mühlenbernd [2015] focus on a specific kind of utilitarian fairness norm, which promotes behavior beneficial to the less advantaged agents. Based on it, they are able to derive a measure of responsibility as the deviation between the ideality of the current state of affairs—defined by means of an external measure of ideality—and that of all other states in which the agent acted differently. This reminds us of our definitions of responsibility and blame used in specifying moral anger, but also differs from them, for we do not base our analysis on an external ideality function when considering the effects of actions. Instead, we base our analysis on Schweder's ethics, where the goodness of

the situation is determined based on more primitive concepts such as harm. In a way, this allows us to determine the material conditions behind the internalized moral rule [Haidt 2003]. Furthermore, in our analysis, moral anger includes the possibility for the agent to change the goal-thwarting state of affairs, whereas in the game-theoretic model of guilt, this aspect is not a consideration at all.

Finally, Gratch and Marsella [2004] propose a computational model of emotions that incorporates coping in a multiagent setting. However, the authors do not provide any details on the underlying logic, which makes comparing the two approaches difficult.

7. CONCLUSION

In this work, we provide a semiformal specification of the elicitation conditions and coping strategies of a set of socially grounded emotions, dubbed moral. The specification is based on appraisal theories of emotion and the CAD triad hypothesis, and is grounded in a multiagent BDI framework. In this system, emotions are defined based on agents' actions and cognitive attitudes (including beliefs, goals, and intentions). The moral aspect of the modeled emotions is based on Shweder's ethics and is represented using concepts grounded in the agents' beliefs and goals. Coping strategies are represented as belonging to several categories depending on their effects on the cognitive attitudes of agents, and are applied using a triggering mechanism based on the elicitation conditions of the emotion, plus an estimate of their potential for alleviating the emotion that triggered them.

The result should be viewed as twofold. First, the current conceptualization contributes to building a precise ontology of emotions, by incorporating cognitive theories into existing intelligent agent models. Second, it paves the way toward building and analyzing emotionally and morally aware agents capable of coexisting in a dynamic multiagent environment. Our analysis specifies when other-condemning emotions arise and consequently which behaviors are selected. These specifications can be used to design and develop emotion-driven software agents.

We consider this work as only the first step toward a complete formal specification and operationalization of the attitudes behind moral emotions. The current specification is based on concepts such as control, accountability, contamination, violation, and social significance. These concepts are used without further specification of their internal logical structures. We left a detailed analysis and further specification of these concepts for future research. We also intend to extend the set of emotions as well as the variety of coping strategies in future work. Furthermore, we ignored some aspects of the coping process that may be important in implementing real-world scenarios. These include the concepts of coping power (availability of resources) and adjustment ability (possibility and cost of changing/dropping goals) found in the literature. An important point to be addressed in the future is a mechanism for triggering coping strategies using thresholds on the emotion intensity. A possible extension to the base formalism is the introduction of complex actions. In the present work, moral norms have been modeled in a simplistic manner without representing their logical structure. Future work will address this by extending the base language with means of talking about norms.

REFERENCES

C. Adam. 2007. *The Emotions: From Psychological Theories to Logical Formalization and Implementation in a BDI Agent*. PhD Thesis, IRIT, Toulouse.

C. Adam, A. Herzig, and D. Longin. 2009. A logical formalization of the OCC theory of emotions. *Synthese* 168, 2 (2009), 201–248.

G. Andrighetto, D. Villatoro, and R. Conte. 2010. Norm internalization in artificial societies. *Ai Communications* 23, 4 (2010), 325–339.

J. R Averill. 1982. *Anger and Aggression: An Essay on Emotion*. Springer Verlag Gmbh, New York.

J. R. Averill. 1983. Studies on anger and aggression: Implications for theories of emotion. *American Psychologist,* 38, 1 (1983), 1145–1160.

C. Battaglino, R. Damiano, and L. Lesmo. 2013. Emotional range in value-sensitive deliberation. In *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-agent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 769–776.

C. Bicchieri. 2006. *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press, New York.

S. Blackburn. 1998. *Ruling Passions*. Clarendon Press, Oxford.

J. Broersen, M. Dastani, J. Hulstijn, and L. van der Torre. 2002. Goal generation in the BOID architecture. *Cognitive Science Quarterly* 2, 3–4 (2002), 428–447.

E. E. Buckels, P. D. Trapnell, and D. L. Paulhus. 2014. Trolls just want to have fun. *Personality and Individual Differences* 67 (2014), 97–102.

H. M. Cleckley. 1964. *The Mask of Sanity: An Attempt to Clarify Some Issues About the So Called Psychopathic Personality*. Aware Journalism.

P. R. Cohen and H. J. Levesque. 1990. Intention is choice with commitment. *Artificial Intelligence* 42, 2–3 (March 1990), 213–261.

R. Conte and C. Castelfranchi. 1995. Understanding the functions of norms in social groups through simulation. *Artificial Societies: The Computer Simulation of Social Life,* N. Gilbert (Ed.). UCL Press, London.

M. Dastani and E. Lorini. 2012. A logic of emotions: From appraisal to coping. *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*. 1133–1140.

M. Dastani and J.-J. Ch Meyer. 2006. Programming agents with emotions. In *ECAI*. 215–219.

B. Dubreuil and J.-F. Grégoire. 2013. Are moral norms distinct from social norms?: A critical assessment of Jon Elster and Cristina Bicchieri. *Theory and Decision* 75, 1 (2013), 137–152. DOI:http://dx.doi.org/10.1007/s11238-012-9342-3

J. Elster. 1994. Rationality, emotions, and social norms. *Synthese* 98, 1 (1994), 21–49.

J. Elster. 2009. Social norms and the explanation of behavior. In *The Oxford Handbook of Analytical Sociology.* OUP Oxford, Oxford, 195–217.

M. J. Fischer and R. E. Ladner. 1979. Propositional dynamic logic of regular programs. *Journal of Computer and System Sciences* 18, 2 (1979), 194–211.

N. H. Frijda. 1986. *The Emotions*. Cambridge University Press, Cambridge.

A. Gewirth. 1981. *Reason and Morality*. University of Chicago Press, Chicago 60637.

J. Gratch and S. Marsella. 2004. A domain-independent framework for modeling emotion. *Cognitive Systems Research* 5, 4 (2004), 269–306.

J. Haidt. 2003. The moral emotions. *Handbook of Affective Sciences* 11 (2003), 852–870.

J. Haidt. 2006. *The Happiness Hypothesis: Finding Modern Truth in Ancient Wisdom*. Basic Books, New York.

J. Haidt, P. Rozin, C. McCauley, and S. Imada. 1997. Body, psyche, and culture: The relationship between disgust and morality. *Psychology & Developing Societies* 9, 1 (1997), 107–131.

R. D. Hare and S. D. Hart. 1993. Psychopathy, mental disorder, and crime. In *Psychopathy, Mental Disorder, and Crime.* Sage Publications.

C. C. Helwig, P. D. Zelazo, and M. Wilson. 2001. Children's judgments of psychological harm in normal and noncanonical situations. *Child Development* 72, 1 (2001), 66–81.

E. T. Higgins. 1987. Self-discrepancy: A theory relating self and affect. *Psychological Review* 94, 3 (1987), 319.

N. A. Johnson, R. B. Cooper, and W. W. Chin. 2009. Anger and flaming in computer-mediated negotiation among strangers. *Decision Support Systems* 46, 3 (2009), 660–672.

R. S. Lazarus. 1991. *Emotion and Adaptation*. Oxford University Press, Oxford.

R. S. Lazarus and S. Folkman. 1984. *Stress, Appraisal, and Coping*. Springer Publishing Company, New York.

E. Lorini. 2011. A dynamic logic of knowledge, graded beliefs and graded goals and its application to emotion modelling. *Logic, Rationality, and Interaction* 6953 (2011), 165–178.

E. Lorini, D. Longin, and E. Mayor. 2013. A logical analysis of responsibility attribution: Emotions, individuals and collectives. *Journal of Logic and Computation* 9387 (2013), ext072.

E. Lorini and R. Mühlenbernd. 2015. The long-term benefits of following fairness norms: A game-theoretic analysis. In *Principles and Practice of Multi-Agent Systems (PRIMA'15)*. Springer, 301–318.

E. Lorini and F. Schwarzentruber. 2011. A logic for reasoning about counterfactual emotions. *Artificial Intelligence* 175, 3–4 (March 2011), 814–847.

K. Oatley and P. N. Johnson-Laird. 1996. The communicative theory of emotions: Empirical tests, mental models, and implications for social interaction. In *Striving and Feeling: Interactions Among Goals, Affect, and Self-Regulation,* L. L. Martin and A. Tesser (Eds.). Psychology Press, New York, 363–393.

K.-I. Ohbuchi, M. Kameda, and N. Agarie. 1989. Apology as aggression control: Its role in mediating appraisal of and response to harm. *Journal of Personality and Social Psychology* 56, 2 (1989), 219.

A. Ortony, G. L. Clore, and A. Collins. 1990. *The Cognitive Structure of Emotions*. Cambridge University Press, Cambridge.

A. Pankov and M. Dastani. 2015. Towards a formal specification of moral emotions. In *2nd International Workshop on Emotion and Sentiment in Social and Expressive Media (ESSEM at AAMAS'15)*. 3–18.

R. Plutchik. 1980. *Emotion: A Psychoevolutionary Synthesis*. Harper & Row, New York.

J. Prinz. 2007. *The Emotional Construction of Morals*. Oxford University Press, Oxford.

A. S. Rao and M. P. Georgeff. 1991. Modeling rational agents within a BDI-architecture. In *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning.* 91 (1991), 473–484.

S. Redmond. 2014. *Celebrity and the Media*. Palgrave Macmillan, New York.

P. Rozin and A. E. Fallon. 1987. A perspective on disgust. *Psychological Review,* 94, 1 (Jan. 1987), 23–41.

P. Rozin, J. Haidt, and C. R. McCauley. 1999. Disgust: The body and soul emotion In *Handbook of Cognition and Emotion* T. Dalgleish and J. Mick (Eds.). John Wiley & Sons, New York. 429–445.

P. Rozin, J. Haidt, and C. R. McCauley. 2008. Disgust. In *Handbook of emotions* (3rd ed.), M. Lewis, J. M. Haviland-Jones, and L. F. Barrett (Eds.). Guilford Press, New York. 757–776.

P. Rozin, L. Lowery, S. Imada, and J. Haidt. 1999. The CAD triad hypothesis: A mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). *Journal of Personality and Social Psychology* 76 (1999), 574–586.

K. R. Scherer. 2001. Appraisal considered as a process of multilevel sequential checking: A component process approach. *Appraisal Processes in Emotion: Theory, Methods, Research* 92 (2001), 120.

R. A. Shweder, N. C. Much, M. Mahapatra, and L. Park. 1997. The "big three" of morality (autonomy, community, divinity) and the "big three" explanations of suffering. *Morality and Health* 119 (1997), 119–169.

A. Sloman and M. Croucher. 1981. Why robots will have emotions. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence*.

A. Staller and P. Petta. 2001. Introducing emotions into the computational study of social norms: A first evaluation. *Journal of Artificial Societies and Social Simulation* 4, 1 (2001), U27–U60.

B. Steunebrink, M. Dastani, and J. J. Meyer. 2009. A formal model of emotion-based action tendency for intelligent agents. *Progress in Artificial Intelligence* 5816 (2009), 174–186.

P. Turrini, J. J. C. Meyer, and C. Castelfranchi. 2010. Coping with shame and sense of guilt: A dynamic logic account. *Autonomous Agents and Multi-Agent Systems* 20, 3 (2010), 401–420.

A. E. Vélez García and F. Ostrosky-Solís. 2006. From morality to moral emotions. *International Journal of Psychology* 41, 5 (2006), 348–354.

E. R. Watkins. 2008. Constructive and unconstructive repetitive thought. *Psychological Bulletin* 134, 2 (2008), 163.

Received December 2015; revised August 2016; accepted September 2016