Towards a Formal Specification of Moral Emotions

Alexander Pankov and Mehdi Dastani

Department of Information and Computing Sciences
Utrecht University, The Netherlands
a.pankov@students.uu.nl , m.m.dastani@uu.nl

Abstract. We propose a semi-formal specification of the elicitation conditions and prototypical coping strategies for three of the 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. Key features of the approach, such as its dynamic and epistemic natures, allow for modeling qualitative, quantitative and dynamic aspects of the moral emotions. We show that successful conceptualization is not only possible, but can shed light on the rationality behind moral emotions, as well as their importance to building socially aware agents.

Keywords: moral emotions, appraisal theories, coping strategies, formal specification

1 Introduction

Moral emotions respond to violations of internalized moral rules, and motivate morally congruent behavior [17,43]. According to Gewirth, the main characteristic of a moral rule is that it must bear on the interests or welfare either of society as a whole or of individuals other than the judge or agent [16]. Therefore, moral emotions are viewed as having two prototypical features: disinterested elicitation conditions (self having no direct stake in the triggering even) and pro-social action tendencies (benefiting others or the social order) [17]. According to the CAD triad hypothesis, and supported by experimental evidence [36], three moral emotions – *contempt*, *anger* and *disgust* – are typically elicited, across cultures, by violations of three specific categories of moral rules advocated by Richard Shweder: ethics of *community*, *autonomy* and *divinity* [38]. Furthermore, there are reasons to think that emotions in general, and moral emotions in particular, play important role in rational behavior [39], healthy mental life [44], and in maintaining social and moral norms [12,16,30,4] within societies.

Although there have been many efforts in the Artificial Intelligence (AI) community to provide a precise specification of emotions [10,9,23,24,42,41], there have not been, to our knowledge, a precise specification dedicated to these three moral emotions and their role in dealing with moral transgressions. The aim of our work is to propose a semi-formal specification of the appraisal and coping process involved in the *other-condemning* – about actions or character of others – moral emotions: anger, contempt and disgust. We focus on these three emotions due to their overtly social nature (being concerned with other agents), and, as a consequence, their potential to influence others' behavior. The choice of Shweder's ethics as the underlying moral theory is warranted by the convincing experimental evidence showing a one-to-one correspondence, across different cultures, between Shweder's ethics and the three emotions under discussion [36]. The proposed specification will, first, allow to operationalize and build emotionally aware software agents with applications ranging from improving education in virtual environments to social media analysis, and building believable video game characters. We note that social media as a unique public, and at the same time, virtual, environment, can be especially useful in analyzing the role emotions play in real human behavior. Second, the specification allows us to analyze how humans and other animate subjects may experience emotions, and how their mental structures change as a consequence. This second aspect enables researchers to disambiguate informal emotion theories and simulate hypothetical situations (morally impossible otherwise) and analyze complex psychological processes, such as aggression, depression and others that have been related to specifics in the appraisal and coping processes. Moreover, it is interesting to see if such formal model can shed light on the rationality and predominance of cooperative, morally congruent, behavior: it will be suggested that coping with moral emotions affects the adoption of goals promoting sanctioning of moral violations - a mechanism for maintaining and reinforcing the social status of moral rules. Last, but not least, the proposed specification is the first step towards a logical formalization of these emotions and can fuel future work by providing a framework in which other emotions can be analyzed.

The approach will be in the spirit of dynamic [14] and belief–desire–intention (BDI) [7,31] models, and, as a result, will provide a cognitive model of intelligent agents capable of experiencing and coping with socially-grounded emotions. The main theoretic and empirical support from cognitive psychology will be the appraisal and coping theories of emotion [22,15,28,21,37], as well as the CAD triad hypothesis [36,17]. Such a support cast – especially appraisal theories – have shown promise in explaining the relationship between social norms and emotions [40], 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 behind the emotion. Different categories of attitudes (such as those concerned with Shweder's ethics) 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 desirable state of affairs by means of deliberation and action.

In what follows, we first present in Sect. 2 an overall mechanism for coping with the other-condemning moral emotions (i.e., anger, disgust, contempt). Then, in Sect. 3, 4 and 5 we provide a detailed description of each of the three emotions, together with a semi-formal specification of their elicitation and common coping strategies. Finally, Sect. 6 delivers concluding remarks on the results of this endeavor.

2 Mechanism for Coping with Moral Transgressions

At the outset, we asserted that the main trigger for an other-condemning moral emotion is a moral transgression. We now ask what is the psychological mechanism accounting for the individual's appraisal and behavior when dealing with moral transgressions. We believe that an answer to this question, and a general account of the similarities and differences between the moral emotions, can be given based on a theory of emotion elicitation and coping. Following the literature on moral emotions [36,17,21] and the relation emotions have to norms in human [12] and artificial [8] societies, we propose the following basic mechanism.

The other-condemning moral emotions get elicited by violations of internalized moral norms. Depending on the category (e.g., community, autonomy, divinity) of the violated moral norm, and thus the specific appraisals involved, different type of moral emotion, requiring different coping strategies, occurs. In most cases a sanction-oriented behavior is promoted, for it alleviates the negative emotion by dealing with concerns that triggered it. As a consequence of this behavior, the status of the violated norm may be reinforced.

Further clarifications are due in order to make the above picture complete. We need to, first, be more explicit in defining the conditions under which moral emotions occur: their general elicitation conditions, and the psychological appraisals behind Shweder's ethics [1]. Second, we need to describe the coping dynamics involved in the moral emotions in such a way that they actually make sense in light of the sanctioning behavior alluded to.

Let us, first, illustrate the proposed mechanism by means of a popular example from the domain of social media: *trolling*. Trolling is usually defined as a provocative behavior of posting inflammatory, offensive, or off-topic messages, and as quite similar to the concepts of flaming and cyberbullying. A troll, in that context, is the agent performing such behavior. There are several recent studies from the psychological literature that provide inside on the cognitive content behind trolling. First, a positive correlation between trolling behavior and personality traits such as sadism (strongest), psychopathy, and Machiavellianism have been shown [5]. Some of these traits have been associated with inability or unwillingness to follow social norms [6,18]. Second, a study have shown a strong correlation between the inflammatory (flaming) nature of trolling and unfairness, harm, and anger [20]. Finally, in popular culture trolling is said to "promote antipathetic emotions of disgust and outrage" [32]. From all this we conclude that trolling can serve as an interesting testbed for our model of the

[1] Here we adhere to Shweder's ethics; however, it should be clear that any such distinction based on the norm content will keep the overall coping mechanism more-or-less intact. What will change are the types of concerns (virtues) involved in the elicitation conditions

moral emotions. For example, imagine a participant in social media discussion posting a comment on a given topic, and receiving a trolling reply. In case the provocative comment is an offense aimed at the person who posted the original comment, then one would not be surprised if some of the participants react with *anger*, verbally attacking the offender or reporting him to the site administrators to be banned. Similarly, if the trolling comment simply uses foul language without attacking someone in particular, one would expect response of reporting or banning the *disgusting* offender; not trying to argue with him, as any such attempt might lead to more foulness. Finally, one can imagine a trolling comment that is not 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 all cases, in accord with the proposed basic mechanism and the cited literature, trolling elicits in the participants an emotion condemning the behavior, and leads to behavior that promotes the agreed upon norm.

Couple of remarks are required before we proceed. Note that throughout we prefer using the term "coping strategies" [22] instead of "action tendencies" [15], 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. What is more important to our discussion, is that emotions in general, and 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 moral emotions. Such flexibility comes mainly from the distinction between *belief-affecting*, *goal-affecting* and *intention-affecting* coping strategies (see [22] for the similar, but not crisp, distinction between problem-directed and emotion-directed coping). As the names suggest goal-affecting coping strategies modify directly the desires of the agent, whereas belief-affecting strategies work on the level of beliefs (still being able to subsequently change the goals and behavior of the agent). Intention-affecting coping function by modifying directly the intentions (planned actions) of the agent.

It is also important to stress here, that we stay agnostic about the essence of moral rules or the process of their internalization (we point, however, to [11] and [1] for a discussion on these topics). What is of interest to us, is their agreed upon pro-social nature [16] and categorization based on content [38,36], the rest remains out of scope for this work.

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, then we 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 semi-formal 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,* 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.

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 (for an overview see [21, pp. 218]). In our trolling example, this amounts to saying that the original poster’s wish to present and discuss his 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, remorse. To address this question, most appraisal theorists incorporate the agent’s attribution of *blame* to another person [21,15]. As a result, blame towards 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 towards oneself, we can expect, for instance, guilt or remorse.

What does it mean, however, to blame someone for his deeds? According to [21], 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 belief 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 could, obviously, have refrained from commenting.

Obviously, attribution of blame is crucial to the elicitation of anger, but is it all there is to it? Lazarus argues that secondary appraisal processes can favor “maximizing the possibilities of success” in coping with the threatening situation, and therefore, influence 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, [37] 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’ term.

Coping Most psychologists agree that the innate coping strategy in anger is *aggression* towards the blameworthy agent [2,3]. Frijda calls the action tendency (in his terms) underlying aggressive behavior "agonistic" [15, pp. 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 [21, pp. 227]. Furthermore, when planning an attack the agent chooses between types of attack (e.g., verbal versus physical, or punishment versus warning) based on coping potential. For instance, in our trolling example, the participant's decision to report the post to an administrator is based on the evaluation of his inability to argue with the offender: an estimate of his coping potential

From this 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 re-appraisal (belief-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 needed [21, pp. 226, Table 6.1].

Moral anger Anger is usually viewed as an immoral emotion, but in many instances it is actually triggered by moral concerns. Of course, it does not mean that anger is always a moral emotion. For instance, consider a modified social media scenario where someone creates a post considered offensive by someone else. In this case, that someone else, can rightfully be angry because of the appraised offense, without any of his moral views being offended.

Moral anger, on the other hand, is a type of anger that arises when *harm* has been done to someone else and his rights have been violated [30, pp. 70]. The relationship between this definition and Shweder's ethics of autonomy has been demonstrated in [36] (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. [38, pp. 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 of a personal goal 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.*

3.1 Anger: Appraisal Specification

Assuming φ as denoting a state of affairs, we use $Control_i(\varphi)$, which should be read as "agent i has control over φ", and define it as there exists an action such that φ will be false after agent i executes the action. In other words, "agent i

can prevent φ from being true". An instance of the $Control_i(\varphi)$ formula can be $Control_{troll}(discussNoOff)$, where $troll$ denotes the agent from our trolling example, and $discussNoOff$ denotes the state of affairs where discussion proceeds with no offenses.

Moreover, we use $Account_i(a,\varphi)$, which should be read as "agent i is accountable for (caused) φ by doing a", and define it as φ is true now and was not true before i performed a.[2] Again, $Account_{troll}(offComment, \neg discussNoOff)$ can be an instance of this formula. 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 attribution of blame. Although similar concepts have been previously analyzed from a logical perspective [25], here we only focus on their role in anger and contempt.

Therefore, we can now define $Blame_{i,j}(a,\varphi)$, which should be read as "agent i blames agent j for doing a and causing φ", as agent i believes that agent j is accountable for φ by doing a, and that before doing a, j had control over φ. Finally, before defining anger, we need a way of talking about the practical possibility of an agent to realize a state of affairs. In our example, this can be understood as a 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 we use $Pos_i(\varphi)$, which should be read as "there is a practical possibility of agent i to make φ true", and define it as there exists an action a, such that if performed by i, φ will be true (e.g., $Pos_{obs}(discussNoOff)$).

We now introduce $Anger_{i,j}(a,\varphi)$, which should be read as "agent i is angry at agent j for doing a and preventing i from achieving φ as planned", and define it as agent i has an achievement goal φ, blames agent j for performing action a, thereby preventing i's plan from achieving φ, and believes there is still a practical possibility of achieving φ. For example, a participant in a social media discussion can be angry at the troll for posting an offensive comment and preventing the discussion (i.e., $Anger_{obs,troll}(offComment, discussNoOff)$).

In this specification, the achievement of goal φ captures the prototypical feature of all emotions, i.e., to be about a desired goal state. Thwarting this goal, as expected for a negatively-valanced emotions, is represented as the agent's belief not to be able to achieve his goal as planned, although agent i believes this was possible before action a was performed by agent j. The belief of i about the practical possibility for achieving φ by some other, not considered before, means highlights the positive evaluation by the agent of his coping potential–the type of secondary appraisal claimed to be an indispensable part of anger.

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 exercise of individual choice in the pursuit of personal preferences. We surmise that the concept of *harm* captures this meaning: preserving one's autonomy means not harming him. Although there are different types of harm distinguished in the literature [27,19], what they all have in common is the violation of per-

[2] We assume that only one agent acts at each moment in time.

sonal preferences by others. In case of physical harm, we can say the personal preference is for protecting one's own body. In case of psychological harm, the personal preference can be viewed as about (not) having certain types of beliefs.

We use $Harm_{i,j}(a, \varphi)$, which should be read as "agent i harmed agent j by doing a and preventing him from achieving φ", and define it as i is accountable by doing action a for j not being able to achieve his goal φ. For example, the troll harmed the original poster by posting an offensive comment and preventing him from discussing the topic without being offended (e.g., $Harm_{troll,poster}(offComment, discussNoOff)$. This definition is quite similar to the one for anger, for we can view anger as triggered by harm to oneself.

We now specify moral anger $MAnger_{i,j,k}(a, \varphi, \psi)$, which should be read as "agent i is morally angry at j for harming k by doing a, preventing k from achieving ψ and preventing i from following his moral norm φ", and define it as 1) $Anger_{i,j}(a, \varphi)$ (i.e., agent i is angry at agent j for doing a and thereby preventing him from achieving the moral norm φ), and 2) agent i believes $Harm_{j,k}(a, \psi)$ with $\varphi \to \psi$ (i.e., φ being the case requires ψ to be the case as well). Note that we refer to i's goal φ as a moral norm, for it implies no harm to k, therefore preserving k's autonomy, one of the moral categories according to Shweder. However, 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, i.e., the preservation of agents' autonomy is considered as a moral rule.

We can see how the above 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 formula $Harm_{j,k}(a, \psi)$ represents the harm aspect of moral anger, whereas $\varphi \to \psi$ captures the logical relationship between the internalized moral rule φ and the violated personal preference ψ.

To illustrate, let us again take our social media example. In its first case, that of directly offending a participant of an online discussion, k from our definition could become the agent posting the original comment, j could be the troll and i can be the observing participant (experiencing the moral anger). Furthermore, for this scenario, ψ could be the original poster's wish to present and discuss his opinion without being offended, φ could represent the "no-offensive language" rule of conduct when posting comments, and the action a would be the actual act of posting an offensive comment. All to the effect of the following moral anger being elicited: $MAnger_{obs,troll,poster}(offComment, noOffLang, discussNoOff)$.

3.2 Anger: Coping Specification

The elicitation of anger – including moral anger – 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 anger preconditions. The prototypical action is attack towards the blameworthy agent.

Furthermore, moral anger is elicited by violation of the autonomy of other agents. We reduced the concept of autonomy to that of harm. Therefore, we specify that coping with moral anger involves adopting the intention of performing an action a for which it is known to lead to $Harm_{j,k}(a, \psi)$ not being true.

This way successfully triggering the thus defined coping strategy removes the presence of moral anger – a property necessary for successful coping [22,44].

In our running example, this amounts to saying that in case of moral anger one should expected attacking behavior (banning, arguing) towards the trolling agent. This way the problem of harming the original poster will be mitigated by allowing the discussion to continue or defending the character of the poster.

4 Disgust

Disgust is an emotion that, from an evolutionary perspective, can be viewed as based on *distaste* - a term referring to the sensory-motor functions of smelling and tasting. Similar to anger, it has simpler (core disgust) and more complex (moral disgust) forms [35]. Research on disgust has gained popularity in the 1990s with some of the main contributors being Paul Rozin and his colleagues [33,34,35].

Elicitation Disgust is considered a response both to physical objects and to social violations [35,28,17]. Lazarus unites the physical and social aspects of disgust by defining it as "taking in or being too close to an indigestible object or idea (metaphorically speaking)"[21, pp. 260]. This and other definitions [28,33] focus on the mouth and dislike towards physical objects, and then suggest that some class of non-physical objects can cause a similar feeling. Furthermore, [35] argue that disgust grew out of a distaste response found also in other animals, which was then shaped by evolution to become a more generalized "guardian of the temple of the body". Thus, getting coupled to, and triggered by, motivation to protect oneself from any sort of *contamination*, including of ideas. Contamination, in this discussion, will have one important property: an agent gets contaminated by coming into contact with another contaminated agent.

Coping All forms of disgust include a motivation to avoid, expel, or otherwise break off contact with the offending entity, often coupled to a motivation to purify, or otherwise remove residues of any physical contact that was made with the entity [35]. 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 [35]. Thus making moral disgust (see below) a powerful drive for action when dealing with norm violations.

As with anger, coping with disgust usually requires intention-affecting (action-directed) strategies to achieve the required result, purity. This does not mean that belief-affecting strategies are not possible, but that in most cases actions are required to deal with the feeling of disgust.

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 domains of sex, drugs, and body modification [17]. Rozin and his colleagues have demonstrated that moral disgust derives from physical 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 [35]. 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 [36], we can make a link between disgust and Shweder's ethics of divinity: social norms concerning the natural order. What follows is that disgust gets triggered by violations of such norms. In explaining the ethics of divinity, [38] write: "[T]he 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 [36], 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 a personal goal aimed at protecting the perceived natural order among agents, including protecting against contamination.*

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

4.1 Disgust: Appraisal Specification

Here we apply more-or-less the same strategy as with anger: use primitive concepts such as goals, beliefs and actions together with the more complex one, the appraisal of accountability. The difference will be in introducing the special atoms C_i, which should be read as "agent i is contaminated".

We use $Disgust_i(a, \varphi)$, which should be read as "agent i is disgusted from experiencing a which caused φ", and define it as agent i has an avoidance goal φ, believes a to have caused φ, and believes that φ leads to the contamination state C_i. Again, the avoidance goal φ captures the prototypical feature of any emotion: to be about a(n) (un)desired state φ, whereas a and C_i capture the property of disgust of being about a kind of contamination of the agent.

As was the case with anger and its moral flavor, moral disgust is actually a type of disgust, with the moral aspect coming from concerns about the actions of others. Therefore we use $MDisgust_{i,j}(a, \varphi)$, which should be read as "agent i is disgusted from agent j doing a which caused φ", and define it as agent i has an avoidance goal φ, believes j to have caused φ by doing a, and believes that φ leads to the contamination state C_i. Here, due to the generality of the definition, there is no need of specifying a third agent, as we did with moral anger, for the appraised contamination triggering disgust can be on any object, not necessarily an agent.

Applying the above definition to our running example should clarify. If the trolling comment from the example contained language considered foul (dirty) by some participant, he is expected to be disgusted by it. In our definition this amounts to saying that j is the troll, i is the participant reading the nasty comment, a is the action of posting a comment containing fault language, and φ expresses i's exposure to dirty language. Then, from assuming that i does not want to be exposed to dirty language, it directly follows that i would experience disgust towards the troll and his comment, which is expressed by the fact $MDisgust_{obs,troll}(foulComment, foulLang)$. In this case the contamination we talk about is purely one of contamination of ideas, but this, as we stated before, is to be expected for the moral flavor of disgust.

4.2 Disgust: Coping Specification

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 from doing a will try performing an action (e.g., expelling the source of contamination) if he thinks it will remove the contamination itself (i.e., C_i). As defined, this coping strategy trigger applies to core disgust. However, having in mind that moral disgust is a type of disgust after all, we see that such a coping strategy would work for the moral variant as well.

Finally, in our trolling example with foul language and elicited disgust, one should expect actions that somehow prevent further contamination. This includes reporting/banning the offending agent, but not arguing with him, for this will only cause further contamination.

5 Contempt

Contempt is one of the least discussed emotions in the psychological literature [17, 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 [29], folded into the anger family [21], or else said to be part of anger [28]. 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 social - in most cases, moral - norms related to obeying social hierarchies.

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* [28, pp. 145]. This is quite similar to what we said about the triggering conditions of anger. This is also the reason why [28] see anger's elicitation conditions as a blend between those of a reproach emotion (such as contempt) and a negative event-based emotion (such as distress). [28] emphasize, however, that anger is not a compound emotions, instead its elicitation conditions have an overlap with those of distress and contempt.

As stated in the introduction, there is evidence [36] for the relation between contempt and violations of Shweder's ethics of community [38]. Shweder writes [38, pp. 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 by [38] regarding the ethics of community are those of *hierarchy* and *duty*. Detailed account of hierarchy and duty in societies is not the aim of this work, however, we suggest these two concepts can be abstracted away in a meaningful way. Hierarchy we consider to be a set of roles, which define a special kind of relation between agents. We call it a *social significance* relation, and should be seen as a relation capturing the potential effects of one's actions on the wellbeing of others', or society as a whole. Violations of one's duties are then indicated by this relation for each possible situation. Such an abstraction, we think, covers the basic cognitive content behind roles and duties, and can serve us in conceptualizing contempt. For example, when participating in social media discussions, one can distinguish two roles: the poster of the original comment and the participant. Their relationship (it terms of social hierarchy and duties) can then be captured by a mechanism to indicate if each action performed is a violation of the duties (e.g., following the topic, writing in the same language) derived from the these two roles.

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 [26]. We are sure there is a lot one can say about these concepts, but we simplify the matter by stipulating that warmth, respect and consideration all supervene on the perceived social significance of the other agent. Thus, less (more) perceived social 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 off-topic comment by an agent, participants will change their appreciation of the importance that participant has to the discussion. His role, including his and others' duties, during the discussion will change.

Note that contempt offers the first example of a belief-affecting coping strategy among moral emotions. This makes contempt significantly different than moral anger and moral disgust. However, we argue that despite its "passive" nature, contempt is still capable of reinforcing the social status of moral norms by indirectly sanctioning moral violators. The corresponding mechanism goes much in the spirit of [13]: becoming aware of others' disapproval, can cause negative emotion (shame) in the subject. Therefore, coping with contempt can lead to epistemic changes that can stimulate the expression of disapproval, which can trigger negative feelings (e.g., shame) in the moral violator, which, on its own, can serve as a sanction for his behavior. Nevertheless, this shaming function, although important as a mechanism for reinforcing the status of social norms, will remain out of scope for our proposed framework. In what follows we will assume the following about contempt.

Elicitation (contempt): *Displeasure from the thwarting of a personal goal concerned with preserving the social hierarchy, combined with the attribution of blame for the goal-thwarting state of affairs.*

Coping (contempt): *Belief-affecting strategies for changing the level of the personal social significance of the blameworthy agent.*

5.1 Contempt: Appraisal Specification

Here we use the special atoms V_i and $Sig_{i,j}$ for talking about violations of duties by agent i, and social hierarchies (in this case agent j is significant to agent i), respectively. As stated above contempt is a negative emotion triggered by violation of a goal concerned with preserving the social hierarchy, together with the attribution of blame for the goal-thwarting state of affairs to someone else. The appraisal of blame has already been defined in previous sections and can be used directly. Preserving the social hierarchy will be modeled as an avoidance goal whose violation leads to breaking the social hierarchy by a significant other.

We now specify contempt $Contempt_{i,j}(a, \varphi)$, which should be read as "agent i is contemptuous towards agent j for doing a and making φ true", and define it as agent i has an avoidance goal φ, blames agent j for performing the physical action a, thus making φ true (i.e., $Blame_{i,j}(a, \varphi)$), believes j to be a significant other ($Sig_{i,j}$) and that φ violates a duty derived from the social hierarchy (i.e., $\varphi \to V_i$). The above definition captures several key components of contempt: goal-incongruence, violation of a norm concerned with preserving the social hierarchy and the attribution of blame. This attribution of blame is what contempt shares with anger and is why [28] have considered them similar.

As with the previous two emotions, let us see how this definition fairs with our running example. In terms of roles, it suffices to say again that there are two roles involved: poster and participant. Poster's duty is to start a topic by clearly stating a proposition, whereas the participant's duty is to contribute to that topic with his opinion or new information, but not to change it. Assuming this simplistic social structure, it becomes obvious how posting an off-topic (trolling) comment can trigger contempt: φ from the above definition becomes the norm of participants not changing the original topic and a the action of actually posting a comment that does: $Contempt_{obs,troll}(offComment, onTopic)$.

5.2 Contempt: Coping Specification

Contempt has the interesting characteristic of affecting one's appreciation of the other agent's social significance, without having direct influence on one's behavior [26]. We specify this prototypical coping strategy as agent i feeling contempt towards agent j will reduce his belief in the social significance of j (i.e., his belief in the formula $Sig_{i,j}$). Note that, although reduction in the social significance of the offending agent might also be possible when coping with anger or disgust, in our work we treat only prototypical coping mechanisms. Such reduction in the social significance is essential to coping with contempt, whereas it is not in the case of anger or disgust.

Again, by trying out this definition in our example, we see its immediate logic: dealing with off-topic comments (the trigger of contempt) involves ignoring them, instead of fighting them, which will only trigger some aggression and further pollute the discussion underway.

6 Conclusion

In this work we provide a semi-formal 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 dynamic, multi-agent BDI framework. In this system, emotions are defined based on agents' actions and attitudes (including graded 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 attitudes of agents, and are applied using a triggering mechanism based on the elicitation conditions of the emotion, plus an estimates 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 towards building and analyzing emotionally and morally aware agents capable of coexisting in a dynamic multi-agent environment.

We consider this work as only the first step towards a complete formal specification and operationalization of the attitudes behind moral emotions. We 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 rules 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 and obligations.

References

1. Andrighetto, G., Villatoro, D., Conte, R.: Norm internalization in artificial societies. Ai Communications 23(4), 325–339 (2010)
2. Averill, J.R.: Anger and aggression: An essay on emotion (1982)
3. Averill, J.R.: Studies on anger and aggression: Implications for theories of emotion. American Psychologist, 38(1), 1145–1160 (1983)
4. Blackburn, S.: Ruling passions. Clarendon Press Oxford (1998)
5. Buckels, E.E., Trapnell, P.D., Paulhus, D.L.: Trolls just want to have fun. Personality and individual Differences 67, 97–102 (2014)
6. Cleckley, H.M.: The mask of sanity: An attempt to clarify some issues about the so called psychopathic personality. Aware Journalism (1964)
7. Cohen, P.R., Levesque, H.J.: Intention is choice with commitment. Artificial Intelligence 42(2–3), 213–261 (Mar 1990)
8. Conte, R., Castelfranchi, C.: Understanding the functions of norms in social groups through simulation. Artificial societies: The computer simulation of social life (1995)
9. Dastani, M., Lorini, E.: A logic of emotions: from appraisal to coping. Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 2 pp. 1133–1140 (2012)
10. Dastani, M., Meyer, J.J.C.: Programming agents with emotions pp. 215–219 (2006)
11. Dubreuil, B., Grégoire, J.F.: Are moral norms distinct from social norms?: A critical assessment of jon elster and cristina bicchieri. Theory and Decision 75(1), 137–152 (2013)
12. Elster, J.: Rationality, emotions, and social norms. Synthese 98(1), 21–49 (1994)
13. Elster, J.: Alchemies of the Mind. Cambridge Univ Press (1999)
14. Fischer, M.J., Ladner, R.E.: Propositional dynamic logic of regular programs. Journal of computer and system sciences 18(2), 194–211 (1979)
15. Frijda, N.H.: The emotions. Cambridge Univ Pr (1986)
16. Gewirth, A.: Reason and morality. University of Chicago Press (1981)
17. Haidt, J.: The moral emotions. Handbook of affective sciences pp. 852–870 (2003)
18. Hare, R.D., Hart, S.D.: chap. Psychopathy, mental disorder, and crime. Sage Publications, Inc (1993)
19. Helwig, C.C., Zelazo, P.D., Wilson, M.: Children's judgments of psychological harm in normal and noncanonical situations. Child Development 72(1), 66–81 (2001)
20. Johnson, N.A., Cooper, R.B., Chin, W.W.: Anger and flaming in computer-mediated negotiation among strangers. Decision Support Systems 46(3), 660–672 (2009)
21. Lazarus, R.S.: Emotion and adaptation. Oxford University Press, USA (1991)
22. Lazarus, R.S., Folkman, S.: Stress, appraisal, and coping. Springer Publishing Company (1984)
23. Lorini, E.: A dynamic logic of knowledge, graded beliefs and graded goals and its application to emotion modelling. Logic, Rationality, and Interaction pp. 165–178 (2011)
24. Lorini, E., Schwarzentruber, F.: A logic for reasoning about counterfactual emotions. Artificial Intelligence (2010)
25. Lorini, E., Longin, D., Mayor, E.: A logical analysis of responsibility attribution: emotions, individuals and collectives. Journal of Logic and Computation p. ext072 (2013)

26. Oatley, K., Johnson-Laird, P.N.: The communicative theory of emotions: Empirical tests, mental models, and implications for social interaction. Martin, Leonard L. (Ed); Tesser, Abraham (Ed), (1996). Striving and feeling: Interactions among goals, affect, and self-regulation. , (pp. 363-393). Hillsdale, NJ, England rtin, Leonard L(6), 363–393 (1996)
27. Ohbuchi, K.i., Kameda, M., Agarie, N.: Apology as aggression control: its role in mediating appraisal of and response to harm. Journal of personality and social psychology 56(2), 219 (1989)
28. Ortony, A., Clore, G.L., Collins, A.: The cognitive structure of emotions. Cambridge Univ Pr (1990)
29. Plutchik, R.: Emotion: A psychoevolutionary synthesis. Harper & Row New York (1980)
30. Prinz, J.: The emotional construction of morals. Oxford University Press (2007)
31. Rao, A.S., Georgeff, M.P.: Modeling rational agents within a bdi-architecture. KR 91, 473–484 (1991)
32. Redmond, S.: Celebrity and the media. Palgrave Macmillan (2014)
33. Rozin, P., Fallon, A.E.: A perspective on disgust. Psychological Review, 94(1), 23–41 (Jan 1987)
34. Rozin, P., Haidt, J., McCauley, C.R.: Disgust: The body and soul emotion. Dalgleish, Tim (Ed); Power, Mick J. (Ed), (1999). Handbook of cognition and emotion. , (pp. 429-445). New York, NY, US: John Wiley & Sons Ltd, xxi, 843 pp. doi lgleish, Tim(9), 429–445 (1999)
35. Rozin, P., Haidt, J., McCauley, C.R.: Disgust. Lewis, Michael (Ed); Haviland-Jones, Jeannette M. (Ed); Barrett, Lisa Feldman (Ed), (2008). Handbook of emotions (3rd ed.). , (pp. 757-776). New York, NY, US: Guilford Press, xvi, 848 p wis, Michael(8), 757–776 (2008)
36. Rozin, P., Lowery, L., Imada, S., Haidt, J., et al.: 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, 574–586 (1999)
37. Scherer, K.R.: Appraisal considered as a process of multilevel sequential checking: A component process approach. Appraisal processes in emotion: Theory, methods, research 92, 120 (2001)
38. Shweder, R.A., Much, N.C., Mahapatra, M., Park, L.: The “big three” of morality (autonomy, community, divinity) and the “big three” explanations of suffering. Morality and health pp. 119–169 (1997)
39. Sloman, A., Croucher, M.: Why robots will have emotions. Proc 7th Int. Joint Conf. on AI (1981)
40. Staller, A., Petta, P.: Introducing emotions into the computational study of social norms: A first evaluation. Journal of artificial societies and social simulation 4(1), U27–U60 (2001)
41. Steunebrink, B., Dastani, M., Meyer, J.J.: A formal model of emotion-based action tendency for intelligent agents. Progress in Artificial Intelligence pp. 174–186 (2009)
42. Turrini, P., Meyer, J.J.C., Castelfranchi, C.: Coping with shame and sense of guilt: a dynamic logic account. Autonomous Agents and Multi-Agent Systems 20(3), 401–420 (2010)
43. Vélez García, A.E., Ostrosky-Solís, F.: From morality to moral emotions. International Journal of Psychology 41(5), 348–354 (2006)
44. Watkins, E.R.: Constructive and unconstructive repetitive thought. Psychological Bulletin 134(2), 163 (2008)