The Other-Condemning Moral Emotions: A Modal Logic Approach

Alexander Pankov

Contents

Abstract

We propose a 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 modal logical 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 maintaining social norms and building socially aware agents.

1 Introduction

Moral emotions are emotions that respond to violations of internalized moral rules, and motivate morally congruent behavior (Haidt, 2003; Vélez García and Ostrosky Solís, 2006). 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 (Gewirth, 1981). 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) (Haidt, 2003). According to the CAD triad hypothesis, and supported by experimental evidence (Rozin et al., 1999b), 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* (Shweder et al., 1997). Furthermore, there are reasons to think that emotions in general, and moral emotions in particular, play important role in rational behavior (Sloman and Croucher, 1981), healthy mental life (Watkins, 2008), and in maintaining social and moral norms (Elster, 1994; Gewirth, 1981; Prinz, 2007; Blackburn, 1998) within societies.

Although there have been many efforts in the Artificial Intelligence (AI) community to provide a precise specification of emotions (Dastani and Meyer, 2006; Dastani and Lorini, 2012; Lorini, 2011; Lorini and Schwarzentruber, 2010; Turrini et al., 2010; Steunebrink et al., 2009), 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 formalization of the appraisal and coping process involved in the *other-condemning* – about actions or character of others – moral emotions: anger, contempt and disgust. Their overtly social nature (being concerned with other agents) and their potential to influence others' behavior, make them interesting subjects to investigate. 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 (Rozin et al., 1999b). The proposed formalization will, first, allow to operationalize and eventually build emotionally aware software agents. Applications range from improving education in virtual environments to social media analysis, and building believable video game characters. Second, the formalization allows us to analyze how humans and other animate subjects experience emotions, and how their mental structures change as a consequence. This second aspect enables researchers to disambiguate informal emotion theories, simulate hypothetical situations (morally impossible otherwise) and analyze complex psychological processes, such as aggression, depression and psychopathy that have been related to specifics in the appraisal and coping processes (Watkins, 2008; Damasio, 2005). 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 logical formalization of these emotions can fuel future work by providing a framework in which other emotions can be analyzed.

The approach 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, 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 (Lazarus and Folkman, 1984; Frijda, 1986; Ortony et al., 1990; Lazarus, 1991; Scherer, 2001), as well as the CAD triad hypothesis (Rozin et al., 1999b; Haidt, 2003). Such a support cast – especially appraisal theories – have shown promise in explaining the relationship between social 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 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 we provide a detailed description of each of the three emotions, together with a specification of their elicitation conditions and common coping strategies in the form of informal definitions. In Sect. 4 we pave the way towards a formal specification of the moral emotions in question, by presenting the syntax and semantics of a *dynamic multi-agent logic of graded attitudes* (DMAL-GA). Then, in Sect. 5 we tackle the main goal of our work: grounding the informal definition from Sect. 3 into the formal system of DMAL-GA. Finally, Sect. 7 and Sect. 6 deliver concluding remarks on the results and unique features of this endeavor.

2 A 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 (Haidt, 2003; Vélez García and Ostrosky Solís, 2006; Rozin et al., 1999b; Lazarus, 1991) and the relation emotions have to norms in human (Elster, 1994) and artificial (Conte and Castelfranchi, 1995) 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 (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, a study have shown a strong correlation between the inflammatory (flaming) nature of trolling and unfairness, harm, and anger (Johnson et al., 2009). Finally, in popular culture trolling is said to "promote antipathetic emotions of disgust and outrage" (Redmond, 2014). From all this we conclude that trolling can serve as an interesting test-bed for our model of the moral emotions. For example, imagine a participant in social media discussion

[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

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.

There are, of course, other domains from human life that can serve as an example of the proposed mechanism. These include the fairly accepted, at least among human societies[2], moral transgression - *theft.* Imagine an observer witnessing an act of stealing from a physically challenged, say blind, person. The question we would like to ask, then, is: What emotion is most likely to be felt by the observing person, and what, if any, behavior this emotion will promote? First, assuming that the witnessing person is not a thief himself, a psychopath, or in some way socially uninvolved, we can expect him to get *angry* at the transgressor, for he is to *blame* for the *harm* done to the blind man. Second, if the person is one who believes in human rights adherence, he might consider the situation as one in which *the autonomy* of the blind person has been violated. Finally, we might expect that anger in the observer will lead to action: Running after the culprit to stop him, or calling the police, are both possibilities for dealing with the negative emotion and mitigating the harm done. Other domains where one can expect to see the causes for and effects of moral emotions are cases of *consensual incest* or *cursing*. In these cases the elicitation of *disgust* or *contempt*, correspondingly, instead of anger, are to be expected in an observer (for more details see Prinz (2007, pp. 121) and some of the scenarios used in Rozin et al. (1999b)).

Couple of remarks are required before we proceed. Note that throughout we prefer using the term "coping strategies" (Lazarus and Folkman, 1984) instead of "action tendencies" (Frijda, 1986), 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 Lazarus

[2] Here we are suggesting that morality is present in non-human societies as well. This is a well-accepted claim in the fields of evolutionary biology and primatology (Tomasello and Vaish, 2013), but is often treated with suspicion elsewhere (e.g. philosophy, cognitive science).

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 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 Dubreuil and Grégoire (2013) and Andrighetto et al. (2010) for a discussion on these topics). What is of interest to us, is their agreed upon pro-social nature (Gewirth, 1981) and categorization based on content (Shweder et al., 1997; Rozin et al., 1999b), the rest remains out of scope for this work.

In the next section, 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. We then provide detailed definitions of the three other-condemning emotions, and a semi-formal specification of their elicitation conditions and coping strategies.

3 The Other-Condemning Moral Emotions

Following our discussion of the intuitions behind the other-condemning moral emotions, in this section we focus on each one of them separately. Table 1 is a concise summary of the concepts to be discussed.

	Moral anger	**Moral disgust**	**Contempt**
Appraisal	harm, blame	distaste	blame
Coping	attack	withdrawal	reduced social significance
Shweder's ethic	autonomy	divinity	community
Principle virtues	rights, justice	purity, natural order	social hierarchy

Table 1: The other-condemning moral emotions

Outlined there is our approach to conceptualizing the moral emotions, which should provide context and reference to the reader when going through the next three subsections. Each of the other-condemning moral emotions will be analyzed, first, in terms of the essential cognitive appraisals involved. Here we follow well-established theorists from psychology in providing coherent and, as much as possible, minimal definitions, aimed at being implementable in a BDI framework. Similarly, the prototypical coping strategies are conceptualized using the same appraisal theories, together with the important distinction between belief- and goal-affecting coping. Then, we discuss the moral flavor of the emotion by pointing out the relevant details (according to Schweder's ethics) of the content (virtues) behind the norms associated with it. Finally, we conclude with a definition of the moral emotion which will then serve as basis for formalization.

3.1 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 Lazarus (1991, 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 (Lazarus, 1991; Frijda, 1986). 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 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 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, 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' term.

Coping

Most psychologists agree that the innate coping strategy in anger is *aggression* towards the blameworthy agent (Averill, 1982, 1983). Frijda calls the action tendency (in his terms) underlying aggressive behavior "agonistic" (Frijda, 1986, 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 (Lazarus, 1991, 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 (Lazarus, 1991, 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 (Prinz, 2007, pp. 70). The relationship between this definition and Shweder's ethics of autonomy has been demonstrated in Rozin et al. (1999b) (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, 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.*

3.2 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 (Rozin et al., 2008). Research on disgust has gained popularity in the 1990s with some of the main contributors being Paul Rozin and his colleagues (Rozin and Fallon, 1987; Rozin et al., 1999a, 2008).

Elicitation

Disgust is considered a response both to physical objects and to social violations Rozin et al. (2008); Ortony et al. (1990); Haidt (2003). 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)"(Lazarus, 1991, pp. 260). This and other definitions (Ortony et al., 1990; Rozin and Fallon, 1987) 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, Rozin et al. (2008) 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 (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). 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 (Haidt, 2003). 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 (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., 1999b), 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, Shweder et al. (1997) 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, 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.*

3.3 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 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* (Ortony et al., 1990, pp. 145). This is quite similar to what we said about the triggering conditions of anger. This is also the reason why Ortony et al. (1990) 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). Ortony et al. (1990) 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 (Rozin et al., 1999b) for the relation between contempt and violations of Shweder's ethics of community (Shweder et al., 1997). Shweder writes (Shweder et al., 1997, 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 Shweder et al. (1997) 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 (Oatley and Johnson-Laird, 1996). 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 Elster (1999): 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.*

4 Towards a Logic of the Moral Emotions

After spending significant amount of time discussing mainly topics from the domain of cognitive psychology (appraisal theories of emotion, the CAD triad hypothesis), we can now proceed to the actual aim of this work: a logical framework for modeling the elicitation and coping process involved in the other-condemning moral emotions.

We can view the resulting model abstractly in input-output terms. The input being a description of the external world composed of set of agents, atomic propositions and a specification of the agents' attitudes - beliefs, goals, intentions - in respect to this world (including other agents). Whereas the output is a description of the emotional state of the agents triggered by all ensuing events, together with a description of the resulting behavior caused by the dynamics of the system.

In what follows we define a multi-agent dynamic logic with special operators for graded beliefs, goals and intentions, and means for defining different types of coping strategies. Which then, in Sect. 5, serves as the basis for the formalization of the other-condemning moral emotions - anger, contempt and disgust.

4.1 DMAL-GA: dynamic multi-agent logic of graded attitudes

In this section we present the syntax and semantics of the logic DMAL-GA (*Dynamic Multi-Agent Logic of Graded Beliefs*). The inspiration comes from work done by Dastani and Lorini (2012) on their DL-GA (*Dynamic Logic of Graded Beliefs*). As the name clearly suggests, the currently proposed formal system can be viewed as a multi-agent version of DL-GA. However, there are significant changes to the semantics of the system. Here the semantical approach to modeling the effects of actions is in the spirit of classical PDL (Propositional Dynamic Logic) (see Blackburn et al., 2002), whereas Dastani and Lorini (2012) take a syntactic approach similar to that of Situation Calculus (Reiter, 2001). As a result, DMAL-GA becomes an extension of the multi-agent logic of dynamic knowledge (BDL) introduced in Schmidt and Tishkovsky (2002). The relationship with BDL is of crucial importance as it can be used for obtaining decidability results. [3]

Syntax

Following Dastani and Lorini (2012) we assume a finite set of agent variables Agt with $|Agt| > 1$, a finite set of physical action variables $PAct = \{a, b, \ldots\}$, a finite set of propositional variables $Atm = \{p, q, \ldots\}$, and a finite set of natural numbers $Num = \{x \in \mathbb{N} \ : \ 0 \leq x \leq max\}$ with $max \in \mathbb{N} \backslash \{0\}$. Let us also

[3] For more detailed discussion on the similarities and differences of the current approach to previous work on these topics, see the section on related work.

have $Num^- = \{-x \,:\, x \in Num \backslash \{0\}\}$. Finally, define the set of propositional formulae $Prop$ to be the set of all Boolean combinations of atomic propositions.

Then, the language $\mathcal{L}$ of DMAL-GA is the smallest set containing all formulae conforming to the following grammar in Backus-Naur Form (BNF):

$$\begin{array}{lcll}
Act & : & \alpha & ::= a \,|\, {-a} \\
Lit & : & l & ::= p \,|\, \neg p \\
CStr & : & \beta & ::= \varphi \uparrow^B \;|\; \varphi \downarrow^B \;|\; l \uparrow^D \;|\; l \downarrow^D \;|\; a^+ \,|\, a^- \\
Fml & : & \varphi & ::= p \,|\, V_i \,|\, C_i \,|\, exc_i^h \,|\, Sig_{i,j} \,|\, Des_i^k l \,|\, Int_i a \,|\, \neg\varphi \,|\, \varphi \wedge \varphi \,| \\
 & & & \qquad\qquad [K_i]\varphi \,|\, [\alpha]_i\varphi \,|\, [\beta]_i\varphi
\end{array}$$

where a ranges over $PAct$, i, j range over Agt, p ranges over Atm, h ranges over Num and k ranges over $Num \cup Num^-$.

The other Boolean constructions on formulae ($\vee$, $\rightarrow$, $\leftrightarrow$, $\top$ and $\bot$) are defined in the standard way using $\neg$ and $\wedge$. The dual modal operators are defined in the usual manner as well: $\langle K_i \rangle \varphi \overset{def}{=} \neg[K_i]\neg\varphi$ and $\langle \alpha \rangle_i \varphi \overset{def}{=} \neg[\alpha]_i\neg\varphi$, where $\varphi \in Fml$.

Some clarification is needed on the intended reading of the different components of the language. The set of actions Act includes physical actions and converse physical actions in the form $-a$, introduced in Parikh (1978). The converse construct allows to express facts about states before the current: states before a physical action was executed. This way giving the possibility of defining axioms such as $p \rightarrow [-a]_i \langle a \rangle_i p$, which should be read as "if p is true in the current state, then in the state before executing the action a, it was true that after performing a p might be true". Note, also, that there is no possibility of combining primitive actions to form composite ones. Although this can be considered as a limitation to the expressiveness of the current approach, it does not prevent us from successfully modeling the concepts involved in the moral emotions. It simply calls for a slightly different appreciation of the meaning behind primitive actions: namely, as representing predefined plans instead of simple, indivisible into smaller parts, actions. Furthermore, this simplification is mainly for ease of exposition, for preliminary investigation shows that complex actions can be introduced to the current system without decidability suffering (see section on future work). Finally, in should be clear that actions are always executable and do not have explicit propositional preconditions. The intuitive idea of an action requiring a specific state of the world in order to be executable is encoded in the effects of the action: if the state of the world does not allow for executing a specific action, then executing the action simply has no effect.

The set of literals Lit contains all atomic propositions together with their negations, and is the domain of the Des_i^k operator, which means they are the target of agents' desires.

The set of coping strategies $CStr$ includes three different types of coping strategies: *belief-affecting* ($\varphi \uparrow^B$ and $\varphi \downarrow^B$), *goal-affecting* ($l \uparrow^D$ and $l \downarrow^D$) and *intention-affecting* (a^+ and a^-). The belief-affecting strategies increase ($\varphi \uparrow^B$) or decrease ($\varphi \downarrow^B$) the belief grade of the executing agent that φ is true.

Similarly, the goal-affecting strategies increase ($l \uparrow^D$) or decrease ($l \downarrow^D$) the desirability of the state of affairs l. Finally, as might be expected, the intention-affecting strategies generate (a^+) or remove (a^-) the intention to perform an action a. Sometimes for $\beta \in CStr$ and $n \in \mathbb{N}$ we will write $[\beta]_i^n$ as a shorthand for nesting $[\beta]_i$ n times: $[\beta]_i^2$ is the same as $[\beta]_i[\beta]_i$.

The set of formulae Fml contains special constructions exc_i^h, V_i, C_i, $Sig_{i,j}$, $Des_i^k l$ and $Int_i a$ which are used to represent agents' mental states. Formulae exc_i^h are used to identify the degree of plausibility of a given world for a given agent i. Following Spohn (1988), the worlds that are assigned the smallest numbers are the most plausible. Therefore, formula exc_i^h can be read as "the current world has a degree of exceptionality h for agent i" or "the current world has a degree of plausibility $max - h$ for agent i". In the spirit of Turrini et al. (2010), the formulae V_i, C_i and $Sig_{i,j}$ are special atoms that will receive special semantics. Such atoms are used to describe concerns assumed primitive in the current formalization, but needed for the definition of some emotional states. $Sig_{i,j}$ should be read as "i is significant other for j", V_i should be read as "the current world is considered bad by i due violation of social hierarchy by someone else", whereas C_i should be read as "the current world is considered bad by i due to contamination of his body or mental state" or simply "agent i is contaminated". The formula $Des_i^k l$ represents the desires, or preferences, of agent i and has to be read as "the state of affairs l has a degree of desirability k for agent i". For notational convenience, in what follows, the following abbreviations are used: $AchG_i^k l \overset{def}{=} Des_i^k l$ for $k > 0$ and $AvdG_i^k l \overset{def}{=} Des_i^{-k} l$ for $k > 0$, where $AchG$ and $AvdG$ respectively stand for *achievement goal* and *avoidance goal*. Formula $Int_i a$ represent the agents' intentions or commitments and are assumed to apply only to physical actions. In that respect, $Int_i a$ should be read as "agent i intends to perform the physical action a".

Furthermore, the logic has an epistemic operator K_i for each agent. The formula $[K_i]\varphi$ should be read as "agent i knows that φ is true". This concept of knowledge is the standard S5-notion and represents an absolutely unrevisable belief that is stable under belief revision with any new evidence. For a more relaxed version of belief, namely graded belief based on the plausibility relation represented by the exc_i^h formulae, see below.

The formula $[\alpha]_i\varphi$ covers the dynamic nature of the formalism by referring to the state of the world after the execution of actions. It should be read as "after agent i performs action α, φ will be true". As can be seen by the reference to an agent i, the formula associates performed actions with the agent performing it. That is the agent causing the physical world to change.

Finally, the formula $[\beta]_i\varphi$ uses a different type of dynamic operator ($[\beta]_i$) for referring to the state of the world after the execution of coping strategy β by agent i. More precisely, $[\beta]_i\varphi$ should be read as "after agent i executes β, φ will be true".

Graded belief

Another crucial aspect of the language is the possibility of defining *graded beliefs* using the formulae exc_i^h and the epistemic operators K_i. First, we introduce the following abbreviation: $exc_i^{\leq k} \overset{def}{=} \bigvee_{0 \leq l \leq k} exc_i^l$ for all $i \in Agt$ and $k \in Num$. Now, following Spohn (1988) and Laverny and Lang (2005), we define the following doxastic concepts:

$$[B_i]\varphi \overset{def}{=} [K_i](exc_i^0 \rightarrow \varphi)$$

The formula $[B_i]\varphi$ says that an agent believes a formula φ *if and only if* φ is true in all worlds that are maximally plausible (or minimally exceptional) for the agent.

$$[B_i^{\geq h}]\varphi \overset{def}{=} [K_i](exc_i^{\leq h-1} \rightarrow \varphi)$$

The formula $[B_i^{\geq h}]\varphi$ says that an agent believes a formula φ with strength$\geq h$ *if and only if* φ is true in all worlds with exceptionality degree for the agent of $< h$.

$$[SB_i]\varphi \overset{def}{=} [K_i](exc_i^{\leq max-1} \rightarrow \varphi)$$

The formula $[SB_i]\varphi$ says that an agent strongly believes a formula φ *if and only if* φ is true in all worlds with exceptionality degree for the agent of $< max$. The value max is excluded here to avoid collapsing the concept of *strong belief* to that of *knowledge*: if we assume, as expected, $\bigvee_{h \in Num} exc_i^h$, then $[K_i](exc_i^{\leq max} \rightarrow \varphi) \leftrightarrow [K_i]\varphi$.

$$[B_i^h]\varphi \overset{def}{=} \begin{cases} [B_i^{\geq h}]\varphi \wedge \neg[B_i^{\geq h+1}]\varphi & 1 \leq h < max \\ [B_i^{\geq max}]\varphi & h = max \end{cases}$$

The formula $[B_i^h]\varphi$ says that an agent believes that φ exactly with strength h *if and only if* the agent believes that φ with strength at least h and it is not the case that the agent believes that φ with strength at least $h+1$.

Structure

The language $\mathcal{L}$ has a possible world semantics with special functions for exceptionality, desirability and intentions. It is interpreted on structures of the following type:

Definition 1 (Model)**.** *DMAL-GA model* is a structure

$$\mathfrak{M} = \Big\langle Agt,\, W,\, \{\sim_i |i \in Agt\},\, \{\mathcal{K}_i|i \in Agt\},\, \{\mathcal{D}_i|i \in Agt\},\, \{\mathcal{I}_i|i \in Agt\},\, \{\mathcal{R}_i^a|i \in Agt,\, a \in PAct\},\, Aut,\, Cont,\, \mathcal{V}\Big\rangle$$

where:

- Agt is a nonempty set of agents;
- W is a nonempty set of worlds or states;
- $\sim_i \subset W \times W$ is an equivalence (reflexive, transitive and symmetric) relation between worlds in W for every $i \in Agt$;
- $\mathcal{K}_i : W \longrightarrow Num$ is a total function from the set of possible worlds to the set of natural numbers Num for every $i \in Agt$;
- $\mathcal{D}_i : W \times Lit \longrightarrow Num \cup Num^-$ is a total function from the set of possible worlds and the set of literals to the set of integers $Num \cup Num^-$ for every $i \in Agt$;
- $\mathcal{I}_i : W \longrightarrow 2^{PAct}$ is a total function called commitment function, mapping worlds to sets of physical actions for every $i \in Agt$;
- $\mathcal{R}_i^a \subset W \times W$ is a relation between worlds in W for every $i \in Agt$ and $a \in PAct$;
- $Aut : Agt \times Agt \times W \longrightarrow 2^W$ is a function that associates a set of worlds to a pair of agents and a world;
- $Cont : Agt \times W \longrightarrow 2^W$ is a function that associates a set of worlds to an agent and a world;
- $\mathcal{V} : W \longrightarrow 2^{Atm}$ is a valuation function, i.e. maps worlds to a set of atomic propositions. $\dashv$

Agt is a countable set representing agents. It is used to interpret agent specific formulae, as well as to index all agent specific structures in the model.

$\sim_i$ is an equivalence relation used to interpret the epistemic operator K_i. The set $\sim_i (w) = \{v \in W \mid w \sim_i v\}$ is the agent's information state at world w: the set of worlds the agent considers possible at world w. As $\sim_i$ is an equivalence relation, if $w \sim_i v$, then $\sim_i (w) = \sim_i (v)$: being at w or v is indistinguishable for agent i.

The function $\mathcal{K}_i$ is the plausibility grading of the possible worlds for agent i, and is used to interpret the atomic formulae exc_i^h. $\mathcal{K}_i(w) = h$ means that, according to agent i, the world w has a degree of exceptionality h, or alternatively, degree of plausibility $max - h$. The function $\mathcal{K}_i$, together with the epistemic equivalence relation, allow to model the notion of graded belief: among the worlds agent i can not distinguish from, there are worlds the agent considers more plausible.

The function $\mathcal{D}_i$ is the plausibility grading of literals for agent i, and is used to interpret the atomic formulae $Des_i^k l$. $\mathcal{D}_i(w, l) = k$, means that, at world w, for agent i, l has a degree of desirability k. Positive values of k denote positive desirability, whereas negative values of k denote negative desirability (undesirability). A value of 0 means that agent i is indifferent about l at world w.

The function I_i represents the intended actions of agent i at every possible world, and is used to interpret the atomic formulae $Int_i a$.

The relation $\mathcal{R}_i^a$ connects the current world to the one resulting in agent i executing the physical action a. It is used for interpreting the formulae $[a]_i\varphi$ and $[-a]_i\varphi$.

The function Aut is used for interpreting the special formulas V_i and $Sig_{i,j}$.

The function $Cont$ is used for interpreting the special formula C_i.

Constraints on models

DMAL-GA models satisfy the following *normality* condition in respect to the $\mathcal{K}_i$ functions: for every $i \in Agt$ and $w \in W$, there is $v \in W$ such that $w \sim_i v$ and $\mathcal{K}_i(v) = 0$.

In modeling intelligent agents a key concern is the connection between action and knowledge. The most well-known and natural connection found in the literature are the properties of *no learning* (NL) and *perfect recall* (PR)(Hoek, 2001). These two properties can be formulated using the axiom schemas $[a]_i[K_i]p \rightarrow [K_i][a]_ip$ and $[K_i][a]_ip \rightarrow [a]_i[K_i]p$, correspondingly. (NL) says that agent i knows the result of his action in advance – no learning happens during the execution of action a. (PR), on the other hand, expresses the persistence of the agent's knowledge after the execution of an action. Incorporating these two properties in DMAL-GA, requires the following condition on models: $\sim_i \circ \mathcal{R}_i^a = \mathcal{R}_i^a \circ \sim_i$ for every $i \in Agt$ and $a \in PAct$. Here $\circ$ denotes relational composition: $\sim_i \circ \mathcal{R}_i^a = \{(x, z) \in W \times W \mid \exists y \in W \; s.t. \; (x, y) \in \mathcal{R}_i^a \; and \; (y, z) \in \sim_i\}$.

Semantics

Having specified the structures used for interpreting the DMAL-GA language, we now define what does it mean for a formula to be satisfied (or true).

Definition 2 (Truth conditions)**.** Given a DMAL-GA model $\mathfrak{M}$, a world w and a formula $\varphi \in \mathcal{L}$, $\mathfrak{M}, w \vDash \varphi$ will mean that φ is *satisfied* (or *true*) in $\mathfrak{M}$ at state w. The rules defining inductively the *truth conditions* of formulae are as follows:

- $\mathfrak{M}, w \models p \; iff \; w \in \mathcal{V}(p)$;
- $\mathfrak{M}, w \models Des_i^h l \; iff \; h = \mathcal{D}_i(w, l)$;
- $\mathfrak{M}, w \models exc_i^h \; iff \; h = \mathcal{K}_i(w)$;
- $\mathfrak{M}, w \models V_i \; iff \; w \in \bigcup_{j \in Agt} Aut(i, j, w)$;
- $\mathfrak{M}, w \models C_i \; iff \; w \in Cont(i, w)$;
- $\mathfrak{M}, w \models Sig_{i,j} \; iff \; j \in \{k \mid \exists v \text{ s.t. } v \in Aut(i, k, \{w\})\}$;
- $\mathfrak{M}, w \models Int_i a \; iff \; h \in \mathcal{I}_i(w)$;

- $\mathfrak{M}, w \models \neg\varphi$ *iff* not $\mathfrak{M}, w \models \varphi$;
- $\mathfrak{M}, w \models \varphi \wedge \psi$ *iff* $\mathfrak{M}, w \models \varphi$ and $\mathfrak{M}, w \models \psi$;
- $\mathfrak{M}, w \models [K_i]\,\varphi$ *iff* $\mathfrak{M}, v \models \varphi$ for all v s.t. $v \sim_i w$;
- $\mathfrak{M}, w \models [a]_i\,\varphi$ *iff* $\mathfrak{M}, v \models \varphi$ for all v s.t. $w\mathcal{R}_i^a v$;
- $\mathfrak{M}, w \models [-a]_i\,\varphi$ *iff* $\mathfrak{M}, v \models \varphi$ for all v s.t. $v\mathcal{R}_i^a w$;
- $\mathfrak{M}, w \models [\beta]_i\,\varphi$ *iff* $\mathfrak{M}_i^\beta, w \models \varphi$.

where $\mathfrak{M}_i^\beta$ is defined according to Definitions 3, 4 and 5. $\dashv$

Writing $\models \varphi$ will mean that φ is satisfied (or true) in any DMAL-GA model, at any state.

Definition 3 (Update via coping strategy on beliefs)**.** Given a DMAL-GA model

$$\mathfrak{M} = \langle Agt,\, W,\, \{\sim_i\}_{i\in Agt},\, \{\mathcal{K}_i\}_{i\in Agt},\, \{\mathcal{D}_i\}_{i\in Agt},\, \{\mathcal{I}_i\}_{i\in Agt},\, \{\mathcal{R}_i^a\}_{i\in Agt, a\in PAct},\, \mathcal{V}\rangle$$

and $\beta \in \{\varphi \uparrow^B, \varphi \downarrow^B\}$, the update of $\mathfrak{M}$ by agent j's coping strategy β is defined as:

$$\mathfrak{M}_j^\beta = \Big\langle Agt,\, W,\, \{\sim_i\}_{i\in Agt},\, \{\mathcal{K}_i\}_{i\in Agt;\, i\neq j} \cup \mathcal{K}_j^\beta,\, \{\mathcal{D}_i\}_{i\in Agt},\, \{\mathcal{I}_i\}_{i\in Agt},\, \{\mathcal{R}_i^a\}_{i\in Agt, a\in PAct},\, \mathcal{V}_j\Big\rangle$$

where for all $w \in W$:

$$\mathcal{K}_j^\beta = \begin{cases} \mathcal{K}_j(w) & \textit{if } \mathfrak{M}, w \models \varphi \\ Cut_B(\mathcal{K}_j(w) + \delta) & \textit{if } \mathfrak{M}, w \models \neg\varphi \wedge [B_j]\,\varphi \textit{ and } \beta = \varphi\uparrow^B \\ \mathcal{K}_j(w) & \textit{if } \mathfrak{M}, w \models \neg\varphi \wedge \neg\,[B_j]\,\varphi \textit{ and } \beta = \varphi\uparrow^B \\ Cut_B(\mathcal{K}_j(w) - \delta) & \textit{if } \mathfrak{M}, w \models \neg\varphi \wedge [B_j]\,\varphi \textit{ and } \beta = \varphi\downarrow^B \\ \mathcal{K}_j(w) & \textit{if } \mathfrak{M}, w \models \neg\varphi \wedge \neg\,[B_j]\,\varphi \textit{ and } \beta = \varphi\downarrow^B \end{cases}$$

where $\delta \in Num\backslash\{0\}$ and

$$Cut_B(x) = \begin{cases} x & \textit{if } 0 \leq x \leq max \\ max & \textit{if } x > max \\ 0 & \textit{if } x < 0 \end{cases}$$

$\dashv$

Definition 4 (Update via coping strategy on goals)**.** Given a DMAL-GA model $\mathfrak{M} = \langle Agt,\, W,\, \{\sim_i\}_{i \in Agt},\, \{\mathcal{K}_i\}_{i \in Agt},\, \{\mathcal{D}_i\}_{i \in Agt},\, \{\mathcal{I}_i\}_{i \in Agt},\, \{\mathcal{R}_i^a\}_{i \in Agt, a \in PAct},\, \mathcal{V}\rangle$ and $\beta \in \{l \uparrow^B,\, l \downarrow^B\}$, the update of $\mathfrak{M}$ by agent j's coping strategy β is defined as:

$$\mathfrak{M}_j^\beta = \Big\langle Agt,\, W,\, \{\sim_i\}_{i \in Agt},\, \{\mathcal{K}_i\}_{i \in Agt},\, \{\mathcal{D}_i\}_{i \in Agt;\, i \neq j} \cup \mathcal{D}_j^\beta,\, \{\mathcal{I}_i\}_{i \in Agt},$$
$$\{\mathcal{R}_i^a\}_{i \in Agt, a \in PAct},\, \mathcal{V}_j\Big\rangle$$

where for all $w \in W$:

$$\mathcal{D}_j^\beta(w, l^{'}) \quad = \quad \begin{cases} Cut_D(\mathcal{D}_j(w, l^{'}) + \delta) & if\ \beta = l \uparrow^D \ \ and\ l^{'} = l \\ Cut_D(D_j(w, l^{'}) - \delta) & if\ \beta = l \downarrow^D \ \ and\ l^{'} = l \\ \mathcal{D}_j(w, l^{'}) & if\ l^{'} \neq l \end{cases}$$

where $\delta \in Num \backslash \{0\}$ and

$$Cut_D(x) \quad = \quad \begin{cases} x & if\ -max \leq x \leq max \\ max & if\ x > max \\ -max & if\ x < -max \end{cases}$$

⊣

Definition 5 (Update via coping strategy on intentions)**.** Given a DMAL-GA model $\mathfrak{M} = \langle Agt,\, W,\, \{\sim_i\}_{i \in Agt},\, \{\mathcal{K}_i\}_{i \in Agt},\, \{\mathcal{D}_i\}_{i \in Agt},\, \{\mathcal{I}_i\}_{i \in Agt},\, \{\mathcal{R}_i^a\}_{i \in Agt, a \in PAct},\, \mathcal{V}\rangle$ and $\beta \in \{a^+,\, a^-\}$, the update of $\mathfrak{M}$ by agent j's coping strategy β is defined as:

$$\mathfrak{M}_j^\beta = \Big\langle Agt,\, W,\, \{\sim_i\}_{i \in Agt},\, \{\mathcal{K}_i\}_{i \in Agt},\, \{\mathcal{D}_i\}_{i \in Agt},\, \{\mathcal{I}_i\}_{i \in Agt;\, i \neq j} \cup \mathcal{I}_j^\beta,$$
$$\{\mathcal{R}_i^a\}_{i \in Agt, a \in PAct},\, \mathcal{V}_j\Big\rangle$$

where for all $w \in W$:

$$\mathcal{I}_j^\beta(w) \quad = \quad \begin{cases} \mathcal{I}_j(w) \backslash \{a\} & if\ \beta = a^- \\ \mathcal{I}_j(w) \cup \{a\} & if\ \beta = a^+ \end{cases}$$

⊣

Some DMAL-GA validities

Converse actions For every $i \in Agt$, $\varphi \in Fml$ and every $a \in PAct$, we have:

$$\models \varphi \rightarrow [a]_i \langle -a \rangle_i \varphi \tag{1}$$

$$\models \varphi \rightarrow [-a]_i \langle a \rangle_i \varphi \tag{2}$$

Validities (1) and (2) capture the dependence between physical actions and their converse counterparts.

Knowledge and action For every $i \in Agt$, $\varphi \in Fml$ and every $a \in PAct$, we have:

$$\models [a]_i[K_i]\varphi \leftrightarrow [K_i][a]_i\varphi \tag{3}$$

Validity (3) represents the relationship between knowledge and the effects of physical actions: the *no learning* and *perfect recall* properties from above are valid in any DMAL-GA model, at any state.

Beliefs As in Dastani and Lorini (2012) we have the following set of validities related to the belief modality:

For every $i \in Agt$, $p \in Prop$, $\varphi, \psi \in Fml$ and every $h, k \in Num$ such that $h \geq 1$ and $k \geq 1$, we have:

$$\models [K_i]\varphi \rightarrow [B_i^{\geq h}]\varphi \tag{4}$$

$$\models [B_i]\varphi \leftrightarrow [B_i^{\geq 1}]\varphi \tag{5}$$

$$\models [SB_i]\varphi \leftrightarrow [B_i^{\geq max}]\varphi \tag{6}$$

$$\models \neg([B_i]\varphi \wedge [B_i]\neg\varphi) \tag{7}$$

$$\models ([B_i^{\geq h}]\varphi \wedge [B_i^{\geq k}]\psi) \rightarrow [B_i^{\geq min[h,k]}](\varphi \wedge \psi) \tag{8}$$

$$\models ([B_i^{\geq h}]\varphi \wedge [B_i^{\geq k}]\psi) \rightarrow [B_i^{\geq max[h,k]}](\varphi \vee \psi) \tag{9}$$

Validities (4) to (9) highlight some interesting properties of beliefs.

Coping strategies We inherit from Dastani and Lorini (2012) also the following set of validities related to the coping strategy affecting beliefs and desires:

For every $i \in Agt$, $p \in Prop$, $\varphi, \psi \in Fml$ and every $h, k \in Num$ such that $h \geq 1$ and $k \geq 1$, we have:

$$\models [B_i^{\geq h}]\varphi \rightarrow [\varphi \uparrow^B]_i[B_i^{\geq Cut_B(h+\delta)}]\varphi \tag{10}$$

$$\models [B_i^{\geq h}]\varphi \rightarrow [\varphi \downarrow^B]_i [B_i^{\geq Cut_B(h-\delta)}]\varphi \text{ if } Cut_B(h-\delta) > 0 \tag{11}$$

$$\models [B_i^{\geq h}]\varphi \rightarrow [\varphi \downarrow^B]_i \neg [B_i]\varphi \text{ if } Cut_B(h-\delta) = 0 \tag{12}$$

$$\models Des_i^h l \rightarrow [l \uparrow^D] Des_i^{Cut_D(h+\delta)} l \tag{13}$$

$$\models Des_i^h l \rightarrow [l \downarrow^D] Des_i^{Cut_D(h-\delta)} l \tag{14}$$

Validities (10) to (14) highlight some interesting properties of the attitudes update via coping strategies.

4.2 Axiomatization

A Hilbert-style axiomatization of DLMA-GA is given by the following three definitions.

Definition 6 (Rules of proof)**.** Let $\varphi, \psi \in Fml$, $a \in PAct$ and $i \in Agt$, then The *rules of proof* of DMAL-GA are:

$$\varphi, \varphi \rightarrow \psi \vdash \psi \qquad (\textit{Modus ponens})$$

$$\varphi \vdash [\alpha]_i \varphi \qquad (\textit{Generalization for actions})$$

$$\varphi \vdash [K_i]\varphi \qquad (\textit{Generalization for } K)$$

$$\psi_1 \leftrightarrow \psi_2 \vdash \varphi \leftrightarrow \varphi[\psi_1/\psi_2] \qquad (\textit{Substitution})$$

$\dashv$

Definition 7 (Axioms)**.** Let $\varphi, \psi \in Fml$, $a \in PAct$, $\beta \in CStr$ and $i, j \in Agt$, then the set of axioms of DMAL-GA are *all instances of propositional tautologies* plus:

PDL-like axiom schemas for physical actions and their converse.

$$[\alpha]_i(p \rightarrow q) \rightarrow [\alpha]_i p \rightarrow [a]_i q \tag{A1}$$

$$\langle \alpha \rangle_i \, p \rightarrow \neg [\alpha]_i \neg p \tag{A2}$$

$$p \rightarrow [a]_i \, \langle -a \rangle_i \, p \tag{A3}$$

$$p \rightarrow [-a]_i \, \langle a \rangle_i \, p \tag{A4}$$

S5 axiom schemas for the knowledge operator.

$$[K_i](p \rightarrow q) \rightarrow ([K_i]p \rightarrow [K_i]q) \tag{A5}$$

$$\langle K_i \rangle \, p \rightarrow \neg [K_i] \neg p \tag{A6}$$

$$[K_i]p \rightarrow p \tag{A7}$$

$$\langle K_i \rangle \langle K_i \rangle p \rightarrow \langle K_i \rangle p \tag{A8}$$

$$p \rightarrow [K_i] \langle K_i \rangle p \tag{A9}$$

Theory of the agents' mental states.

$$\bigvee_{h \in Num} exc_i^h \tag{A10}$$

$$\bigvee_{k \in Num \cup Num^-} Des_i^k l \tag{A11}$$

$$exc_i^h \rightarrow \neg exc_i^l \text{ if } h \neq l \tag{A12}$$

$$Des_i^k l \rightarrow \neg Des_i^m l \text{ if } k \neq m \tag{A13}$$

$$\langle K_i \rangle exc_i^0 \tag{A14}$$

Reduction axiom schemas for the operators $[\beta]_i$.

$$[\beta]_i p \leftrightarrow p \tag{A15}$$

$$[\beta]_i Int_j a \leftrightarrow \begin{cases} \top & \text{if } \beta = a^+ \text{ and } i = j \\ \bot & \text{if } \beta = a^- \text{ and } i = j \\ Int_j a & \text{if otherwise} \end{cases} \tag{A16}$$

$$[\beta]_i exc_j^h \leftrightarrow \begin{cases} ... & \text{if } \beta = \varphi \uparrow^B \text{ and } i = j \\ ... & \text{if } \beta = \varphi \uparrow^B \text{ and } i = j \\ exc_j^h & \text{if otherwise} \end{cases} \tag{A17}$$

$$[\beta]_i Des_i^k l \leftrightarrow \begin{cases} \bigvee_{k = Cut_D(l+\omega)} Des_j^m l & \text{if } \beta = \varphi \uparrow^D \text{ and } i = j \\ \bigvee_{k = Cut_D(l-\omega)} Des_j^m l & \text{if } \beta = \varphi \downarrow^D \text{ and } i = j \\ Des_i^k l & \text{if otherwise} \end{cases} \tag{A18}$$

$$[\beta]_i \neg \psi \leftrightarrow \neg [\beta]_i \psi \tag{A19}$$

$$[\beta]_i (\psi_1 \wedge \psi_2) \leftrightarrow ([\beta]_i \psi_1 \wedge [\beta]_i \psi_2) \tag{A20}$$

$$[\beta]_i [K_j] \psi \leftrightarrow [K_j][\beta]_i \psi \tag{A21}$$

⊣

Definition 8 (DMAL-GA-proof)**.** A *DMAL-GA-proof* is a finite sequence of formulas, each of which is an axiom, or follows from one or more earlier items in the sequence by applying a rule of proof. ⊣

As a consequence, DMAL-GA can be seen as axiomatized as an extension of a simpler version (with no complex actions) of the logic BDL introduced in Schmidt and Tishkovsky (2002). The extension being the theory of the agents' mental states, together with the reduction axioms for the $[\beta]_i$ operators (see Definition 7).

5 Formalizing the Other-Condemning Moral Emotions

After having defined the basic language of DMAL-GA, we are now well-equipped to move on to the task of formalizing the other-condemning moral emotions: anger, disgust and contempt. This task will require translating the definitions at the end of section 3.1 into the formal language from the previous section. We will introduce several new abbreviations and formal concepts describing the cognitive content of those emotions on top of the basic concepts of actions, goals and beliefs.

5.1 Elicitation

Anger

As we saw in the discussion on anger in Sect. 3.1, the crucial appraisal behind the elicitation of anger is *blame*. Following appraisal theories of emotion, we concluded that there are two more basic concepts behind blame: *accountability* and *control*. In order for an agent to attribute blame to someone for something he has to determine, first, if the other agent is accountable for (or has caused) the state of affairs, and second, if the other agent had control over it (or was able to prevent it).

Formally, in the language of DMAL-GA:

$$Control_i(\varphi) \overset{def}{=} \bigvee_{a \in PAct} [a]_i \neg\varphi$$

The formula $Control_i(\varphi)$ should be read as "agent i has control over φ". By definition, this is the case *if and only if* there exists an action $a \in PAct$, such that φ will be false after agent i executes it. 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.

$$Account_i(a, \varphi) \overset{def}{=} \varphi \wedge [-a]_i \neg\varphi$$

The formula $Account_i(a, \varphi)$ should be read as agent i is accountable for (caused) φ by doing a". By definition, this is the case *if and only if* φ is true now and was not true before i performed a [4]. A possible instance of this formula is $Account_{troll}(offComment, \neg discussNoOff)$.

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 (Lorini et al., 2013), here we only focus on their role in anger and contempt.

We can now define the appraisal of blame in the following manner:

$$Blame^{k}_{i,j}(a,\varphi) \overset{def}{=} [B^{k}_{i}](Account_{j}(a,\varphi) \wedge [-a]_{j}Control_{j}(\varphi))$$

The formula $Blame^{k}_{i,j}(a,\varphi)$ should be read as "agent i blames with strength k agent j for doing a and causing φ". By definition, this is the case *if and only if* agent i believes with strength k that agent j is accountable for φ by doing a, and that before doing a, j had control over φ.

It is important to stress here that we define blame without any negative or moral connotations. Instead, it is viewed as a belief about the accountability of an agent for a given state of affairs, and his control over the the situation. This is much in the spirit of how Lazarus talks about blame in his discussion about anger (Lazarus, 1991, pp. 219). The negative and moral aspects of blame will enter our model later when viewing blame as part of a negatively-valenced emotion and its associated goal-thwarting state of affairs.

Before defining anger, we need a way of talking about the practical possibility of an agent to make a formula true. For this we use $[a]_i\varphi$ over all actions:

$$Pos_{i}(\varphi) \overset{def}{=} \bigvee_{a \in PAct} [a]_{i}\varphi \tag{15}$$

The formula $Pos_i(\varphi)$ should be read as "there is a practical possibility for agent i to make φ true". By definition, this is the case *if and only if* there exists an action a, such that if performed by i, φ will be true. 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, leading to the removal of the offensive comment. Compare this definition and the one of $Control_i(\varphi)$. The content behind both concepts is similar, therefore their formalization looks similar. Note that defining only one of the two concepts and expressing the other through it, would have been sufficient, however, we prefer having them both. Firstly, because of clarity in subsequent definitions where they occur, and secondly, because it allows for redefining them without impacting other concepts.

Finally, we can define anger in DMAL-GA:

Definition 9 (Anger)**.** For $i, j \in Agt$; $a, b \in PAct$; $h, k, l \in Num$ and $\varphi \in Lit$:

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

$$Anger^{l}_{i,j}(a,\varphi,b) \overset{def}{=} \bigvee_{l=merge(h,k)} (AchG^{k}_{i}(\varphi) \wedge Int_i b \wedge \\ Blame^{h}_{i,j}(a,\neg[b]_i\varphi) \wedge [B_i]Pos_i(\varphi))$$

where *merge* is a monotonically increasing function of its two arguments, h and k.[5] Its domain being the set

$$EmoInt = \{y \ : \ \text{there are } x_1, x_2 \in Num \text{ s.t. } merge(x_1,x_2) = y\}$$

$\dashv$

The formula $Anger^{l}_{i,j}(a,\varphi,b)$ should be read as "agent i is angry with intensity l at agent j for doing a and preventing i from achieving φ by doing b". By definition, this is the case *if and only if* agent i has an achievement goal φ with desirability level k, intends to do b, and blames agent j with some strength h for performing the physical action a, thus preventing him from achieving φ by doing b". 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., $Anger_{obs,troll}(offComment, discussNoOff)$).

Let us dissect the above definition of anger and see how it matches the concepts discussed in Sect. 3.1. The first conjunct, $AchG^{k}_{i}(\varphi)$, captures the prototypical feature of any emotion: to be about a desired goal state (φ). The next two conjuncts, $Int_i b$ and $Blame^{h}_{i,j}(a,\neg[b]_i\varphi)$, represent the anger-specific appraisal of blaming someone else for a goal-thwarting state of affairs. Here the goal-thwarting state is represented as the agent's belief not to be able to achieve his goal by executing the intended plan b ($\neg[b]_i\varphi$), although he believes this was possible before action b was performed ($[-a]_j[b]_i\varphi$). This observation about the agent's attitudes is expressed as the following simple proposition:

Proposition 1. *Let $\mathfrak{M}$ be a DMAL-GA model, $w \in W$; a, $b \in PAct$; $i, j \in Agt$; $l \in Num$ and $\varphi \in Fml$.*

If $\mathfrak{M}, w \models Anger^{l}_{i,j}(a,\varphi,b)$, then $\mathfrak{M}, w \models [B^{h}_{i}](\neg[b]_i\varphi \wedge [-a]_j[b]_i\varphi)$ for some $h \in Num$.

Proof. From $\mathfrak{M}, w \models Anger^{l}_{i,j}(a,\varphi,b)$ follows by definition that $\mathfrak{M}, w \models AchG^{k}_{i}(\varphi) \wedge Blame^{h}_{i,j}(a,\neg[b]_i\varphi) \wedge Int_i b$ for some $k, h \in Num$ s.t. $l = merge(h,k)$. From the $\mathfrak{M}, w \models Blame^{h}_{i,j}(a,\neg[b]_i\varphi)$ conjunct follows by definition that $\mathfrak{M}, w \models [B^{h}_{i}]Account_j(a,\neg[b]_i\varphi)$. Finally, from $\mathfrak{M}, w \models [B^{h}_{i}]Account_j(a,\neg[b]_i\varphi)$ by the definition of $Account_i(a,\varphi)$ it follows that $\mathfrak{M}, w \models [B^{h}_{i}](\neg[b]_i\varphi \wedge [-a]_j[b]_i\varphi)$. □

Finally, $[B_i]Pos_i(\varphi)$, the last conjunct in the definition, highlights the positive evaluation by the agent of his coping potential – the type of secondary appraisal claimed to be an indispensable part of anger.

[5] As suggested by some appraisal theorists (Ortony et al., 1990; Lazarus, 1991), the function *merge* models the intensity of emotions by merging the strength of the negative belief behind blame and the desirability of φ. Possible instances of such a merging function are $\frac{h+k}{2}$ and $h \times k$.

At first glance the definition of anger might seem excessively complex. However, a case can be made that removing parts from it will deflate the concept, and make it indistinguishable from other emotions. Furthermore, the language of DMAL-GA is rich enough to be able to express more specific cases of anger (e.g. one can define a formula as concise as $Anger_i(\varphi)$) without the general case having to suffer in expressiveness. Therefore, concerns for excessive complexity seem to be unwarranted.

Moral 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 then reduced to exercise of individual will in the pursuit of personal preferences. We surmised 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 (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. In case of psychological harm, the personal preference can be viewed as about (not) having certain types of beliefs.

We represent now the emotion of moral anger, together with the concept of harm, in the language of DMAL-GA.

$$Harm_{i,j}(a,\varphi) \overset{def}{=} AchG_j\varphi \wedge Account_i(a, \neg Pos_j(\varphi)) \tag{16}$$

The formula $Harm_{i,j}(a,\varphi)$ should be read as "agent i harmed agent j by doing a and preventing him from achieving φ". By definition, agent i harmed agent j by doing a and preventing him from achieving φ *if and only if* j has an achievement goal φ, i is accountable for j not being able achieve its 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., $Harm_{troll,poster}(offComment, discussNoOff)$. Note also that this definition is quite similar to the one for anger, for we can view anger as triggered by harm to oneself.

Now to the definition of the moral flavor of anger:

Definition 10 (Moral anger)**.** For $i,j,k \in Agt$, $i \neq k$; $a \in PAct$; $l \in EmoInt$ and $\varphi \in Lit$:

$$MAnger^l_{i,j}(a,\varphi,\psi) \overset{def}{=} Anger^l_{i,j}(a,\varphi) \wedge [B_i](Harm_{j,k}(a,\psi) \wedge (\varphi \rightarrow \psi))$$

$\dashv$

The formula $MAnger^l_{i,j}(a,\varphi,\psi)$ should be read as "agent i is morally angry at j for harming k by doing a and preventing k from achieving ψ and preventing i from following his moral norm φ". By definition, $MAnger^l_{i,j}(a,\varphi,\psi)$ is true *if and only if* 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).

We can see how this definition captures our previous analysis of the concept of moral anger, namely, as a type of anger with content related to the autonomy of others. Note that here 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, not some intrinsically moral property. 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.

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 becomes the agent posting the original comment, j the troll and i the observing participant (feeling morally angry). Furthermore, for this scenario, ψ should be the original poster's wish to present and discuss his opinion without being offended, φ represents 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: $MAnger_{obs,troll,poster}(offComment, noOffLang, discussNoOff)$.

As with non-moral anger, we can write $MAnger_{i,j}(a, \varphi, \psi)$, $MAnger_{i,j}(a, \varphi)$ or $MAnger_{i}(a, \varphi)$, all defined as abbreviations on the general definition.

Disgust

As we saw in the discussion on disgust in Sect. 3.2, the crucial appraisal behind the elicitation of disgust, in addition to the goal-incongruance typical to all emotions, is a cause of contamination. We will apply more-or-less the same strategy as with anger: using primitive concepts such as goals, beliefs and actions together with some more complex ones such as the appraisal of accountability. The difference will be in the use the special atoms C_i for talking about contamination.

Formally, in the language of DMAL-GA:

Definition 11 (Disgust). For $i \in Agt$; $a \in PAct$; $h \in Num$ and $\varphi \in Lit$:

$$Disgust_i^h(a, \varphi) \overset{def}{=} AvdG_i^h(\varphi) \wedge [B_i](Account_i(a, \varphi) \wedge (\varphi \to C_i))$$

$\dashv$

The formula $Disgust_i^h(a, \varphi)$ should be read as "agent i is disgusted with intensity h from experiencing a which caused φ which then led to his contamination". By definition, this is the case *if and only if* agent i has an avoidance goal φ with desirability level h, believes to be accountable for φ by doing a, and believes that φ leads to the contamination C_i, which should be read as "agent i is contaminated".

Again, the first conjunct, $AvdG_i^k(\varphi)$, captures the prototypical feature of any emotion: to be about a (un)desired goal state (φ). The second conjunct captures the property of disgust of being about a kind of contamination of the agent.

Moral disgust

As was the case with anger, moral disgust is a type of disgust, but this time involving the actions of others.

Definition 12 (Moral disgust)**.** For $i, j \in Agt$; $a, b \in PAct$; $h \in Num$ and $\varphi \in Lit$:

$$MDisgust_{i,j}^h(a, \varphi, b) \overset{def}{=} Disgust_i^h(a, \varphi) \wedge [B_i](Account_j(b, C_j))$$

$\dashv$

The formula $MDisgust_{i,j}^h(a, \varphi, b)$, which should be read as "agent i is disgusted with intensity h from agent j doing b which caused i to experience a and cause φ", and define it as agent i has an avoidance goal φ, believes j to have caused φ by doing b, 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, b is the action of posting a comment containing foul language, and φ expresses i's exposure to dirty language after seeing (action a) the comment. 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 $MDisgust_{obs,troll}(foulComment, foulLang, seeComment)$. 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.

Contempt

After having defined the elicitation of moral anger and disgust we move on to contempt. The formalization is based on the discussion and definitions from Sect. 3.3. We will use primitive concepts such as goals, beliefs and actions together with some more complex ones such as the appraisal of blame. Furthermore, we will use the special atoms V_i and $Sig_{i,j}$ for talking about violations of duties by agent i, and social significance, 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.

Definition 13 (Contempt)**.** For $i, j \in Agt$; $a \in PAct$; $h, k \in Num$ and $\varphi \in Lit$:

$$Contempt^{l}_{i,j}(a, \varphi) \overset{def}{=} \bigvee_{l=merge(h,k)} (AvdG^{k}_{i}(\varphi) \wedge Blame^{h}_{i,j}(a, \varphi) \wedge [B_i](Sig_{i,j} \wedge (\varphi \to V_j)))$$

where *merge* is the same as in Definition 9. ⊣

The formula $Contempt^{l}_{i,j}(a, \varphi)$ should be read as "agent i is contemptuous with intensity l towards agent j for doing a and making φ true". By definition, this is the case *if and only if* agent i has an avoidance goal φ with desirability level k, blames agent j with some strength h for performing the physical action a, thus making φ true, believes j to be a significant other and that φ leads to a violating the social hierarchy".

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 Ortony et al. (1990) have considered them similar.

Similarly to anger we can intuitively define the following abbreviations: $Contempt_{i,j}(a, \varphi)$, $Contempt_{i,j}(\varphi)$, $Contempt_{i,j}(a)$ and $Contempt_{i,j}$.

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: $Contempt_{obs,troll}(offComment, offTopic)$.

5.2 Coping

In this section we move from elicitation of the moral emotions to mechanisms for coping with them. We provide definitions of some of the prototypical coping strategies involved.

Before we proceed, we first need to be able to define triggering conditions for coping strategies: a way of specifying when a given coping strategy is to be executed. Following Dastani and Lorini (2012), we address this requirement by introducing a function Trg:

$$Trg \ : \ Agt \times CStr \to Fml$$

It maps strategies and agents to formulae: for every coping strategy β and agent i, $Trg(i, \beta)$ captures the conditions under which the strategy is triggered for agent i. Instead of $Trg(i, \beta)$, we will write $Tr_i(\beta)$.

Following appraisal theorists (Lazarus, 1991; Ortony et al., 1990; Scherer et al., 2001), here we assume that coping is triggered by emotion, and in what follows the triggering conditions of coping strategies for the moral emotions will always include emotion elicitation.

Moral anger

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 (Lazarus and Folkman, 1984; Watkins, 2008).

Formally, in the language of DMAL-GA, this can be stated as follows:

$$Tr_i(at^+) = MAngry_{i,j,k}(a, \varphi, \psi) \wedge [K_i][at]_i \neg Harm_{j,k}(a, \psi) \tag{17}$$

where $at \in PAct$ and all the other variables are as in Definition 10. An immediate observation is the following:

Proposition 2. *Let $\mathfrak{M}$ be a DMAL-GA model, $w \in W$; $at, a \in PAct$; $i, j, k \in Agt$ and $\varphi, \psi \in Fml$.*

If $\mathfrak{M}, w \models [K_i][at]_i \neg Harm_{j,k}(a, \psi)$, then $\mathfrak{M}, w \models [at]_i \neg MAngry_{i,j,k}(a, \varphi, \psi)$.

Proof. Directly from validities 3, 4 and Definition 10. □

That is, successfully triggering the coping strategy at^+ from equation (17) for agent i, and executing the action at, removes the presence of moral anger – a property necessary for successful coping (Lazarus and Folkman, 1984; Watkins, 2008).

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.

Moral disgust

From our discussion in Sect. 3.2 it became clear that the prototypical coping strategy when dealing with disgust is an intention-affecting strategy to try and expel the source of contamination. Here we formalize such a coping strategy in the language of DMAL-GA as follows:

$$Tr_i(exp^+) = Disgust_i(a, \varphi) \wedge [K_i][exp]_i \neg Account_i(a, \varphi) \tag{18}$$

where $exp \in PAct$ and all the other variables are as in Definition 11.

It simply says that 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. 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:

Proposition 3. *Let $\mathfrak{M}$ be a DMAL-GA model, $w \in W$; $a \in PAct$; $i, j \in Agt$ and $\varphi \in Fml$.*

If $\mathfrak{M}, w \models [K_i][exp]_i \neg Account_i(a, \varphi)$, then $\mathfrak{M}, w \models [exp]_i \neg MDisgust_i(a, \varphi)$

Proof. Directly from validities 3, 4 and Definition 12. □

That is, successfully triggering the coping strategy exp^+ from equation (18) for agent i, eventually removes the presence of disgust – a property necessary for successful coping (Lazarus and Folkman, 1984; Watkins, 2008).

Contempt

As already stated in Sect. 3.3, contempt has the interesting characteristic of affecting one's appreciation of the other agent's social worthiness, without having direct influence on one's behavior. Here we formalize this prototypical coping strategy in the language of DMAL-GA as follows:

$$Tr_i(Sig_{i,j} \downarrow^B) = Contempt_{i,j} \tag{19}$$

where all variables are as in Definition 13.

It simply says that an agent i feeling contempt towards agent j will reduce his belief in the $Sig_{i,j}$ formula expressing the social worthiness of j to i.

Proposition 4. *Let $\mathfrak{M}$ be a DMAL-GA model, $w \in W$; $a \in PAct$; $i, j \in Agt$ and $\varphi \in Fml$.*

If $\mathfrak{M}, w \models Contempt_{i,j}(a, \varphi)$, then $\mathfrak{M}, w \models [Sig_{i,j} \downarrow^B]_i^n \neg Contempt_{i,j}(a, \varphi)$ for some $n \in \mathbb{N}$.

Proof. Directly from Validities 11 and 12 and Definition 13. □

That is, successfully triggering the coping strategy $Sig_{i,j} \downarrow^B$ from equation (19) for agent i, eventually removes the presence of contempt – a property necessary for successful coping (Lazarus and Folkman, 1984; Watkins, 2008).

5.3 Discussion on sanction-oriented behavior

In this section we discuss the explanatory power of our formal model when it comes to the rationality of cooperative, morally congruent, behavior.

By now we have seen how coping with moral emotions affects the goals and beliefs of agents, but is there something special about the different ways of successfully coping with moral emotions? We suggest a positive answer to this question by describing a possible mechanism for maintaining the social status of moral rules. With the help of couple of coherence assumptions on the beliefs and intentions of agents involved in a moral conduct, we can show that sanction-oriented behavior gets promoted in agents experiencing moral emotions.

We focus on a single type of sanctioning behavior: attack, understood broadly (including cases of both verbal and physical attacks), as a behavior that aims at harming (thwarting the goals of) another agents. However, we think that similar consideration can be made for behaviors not directly appreciated as a kind of attack, but having, nevertheless, a similar effect on the agent experiencing the results. Think of the act of withdrawal from the moral transgressor in case of disgust, and that of reducing one's appreciation of the social significance of the moral transgressor in case of contempt. Both types of behavior can have a negative effect to the goal satisfaction of the target agent (for a broader discussion on these topics see Elster (1994, 1999); Gewirth (1981); Prinz (2007), as well as the literature on shame and guilt).

Attack

Note that although we said that the type of behavior to be expected when coping with anger is attack, at from equation (17) can be any action from the set of possible actions, with the only requirement being that the agent knows it can preclude harm. However, as mentioned in the introduction, according to Frijda (1986) attack is at the core of aggression caused by anger. Furthermore, in a famous study, Conte and Castelfranchi (1995) showed, by means of computer simulation, how attack can serve as a sanction to agents violating a social norm. So, how does attacking behavior fit the current formalism? To answer this question, we will first say something about the nature of attack, and then formulate its relationship to coping.

What aggression and sanctioning have in common is that they are both harming behaviors. The difference being in their justification: a sanction usually obtains its justification from a previously established rule or a norm (Conte and Castelfranchi, 1995), whereas aggression can have purely idiosyncratic motivation (Anderson and Bushman, 2002; Buss, 1962). In our view, attack can then be seen as the action behind aggression and sanctioning. In the current formalism, attack towards someone is modeled as a type of action causing harm to that someone.

Now to the question of how attack becomes desired when coping with anger. We claim that the answer lies in the motivation of the blameworthy agent causing the anger. According to the BDI paradigm, and assumed here, intentions are aimed at desired states. It is therefore expected that agent j from equation (17) – the agent that caused harm and therefore anger – was actually motivated by a desired state of affairs. Furthermore, assume a conflict between his desire and the thwarted desire of agent k: that they can not be fulfilled together. In our theft example, this amounts to saying that the thief wanted to have the wallet, and that only one of him and the blind person can have it at the same time. In this case it is easy to show that an attack towards j (the thief) causing him harm by making his desire impossible will remove the harm done to k (the blind person), and alleviate as a consequence the anger felt by i (the observer). Formally this amounts to the following proposition:

Proposition 5. *Let $\mathfrak{M}$ be a DMAL-GA model, $w \in W$; at $\in PAct$; $i, j, k \in Agt$ and $\varphi, \psi \in Fml$. If $\mathfrak{M}, w \models [at]_i(Harm_{i,j}(a, \varphi) \wedge (\varphi \vee \psi) \wedge \neg(\varphi \wedge \psi))$, then $\bigwedge_{b \in PAct} \mathfrak{M}, w \models [at]_i \neg Harm_{j,k}(b, \psi)$.*

Proof. From $\mathfrak{M}, w \models [a]_i Harm_{i,j}(a, \varphi)$ and the definition of harm (equation (16)), it follows that $\mathfrak{M}, w \models [a]_i \neg Pos_j(\varphi)$. From this and the definition of Pos (equation (15)), it follows that $\mathfrak{M}, w \models [a]_i \neg [nop]_j \varphi$, which leads to $\mathfrak{M}, w \models [a]_i \neg \varphi$. Then from this and $\mathfrak{M}, w \models [a]_i(\varphi \vee \psi)$ in the assumption it follows that $\mathfrak{M}, w \models [a]_i \psi$. From this it follows that $\mathfrak{M}, w \models [a]_i [nop]_k \psi$ and, as a consequence that $\mathfrak{M}, w \models [a]_i Pos_k(\psi)$. Finally, from this we conclude $\mathfrak{M}, w \models \bigwedge_{b \in PAct} [a]_i \neg Harm_{j,k}(b, \psi)$. □

According to Proposition 5, attacking an agent by thwarting his desire that caused him to harm someone, can remove the harm done to that someone. Consequently, if a morally angry person knows about such an action, the coping mechanism in equation (17) will favor the adoption of this action as his intention. All this suggests a cognitive mechanism, grounded in emotion elicitation and coping, that leads to behavior aimed at enforcing a social norm (in our theft example, the do-not-steal social norm). Although here the case has been made for moral anger only, a similar observation can be made for the other two emotions from the other-condemning family.

With this we conclude our exposition of the formal framework, and proceed to a comparison of our model with similar work.

6 Related Work

As outlined previously, the most important features of the presented formal system are its multi-agent flavor, its inclusion of emotion intensity based on belief and goal strengths, and its representation of coping strategies. Although the importance of emotion intensity and coping mechanisms has been stressed by appraisal theorist, most of the formal models in the literature have ignored at least one of them. Examples include Adam et al. (2009); Lorini and Schwarzentruber (2010); Turrini et al. (2010) where the quantitative aspect of emotions have been ignored. Lorini (2011), on the other hand, is much in the spirit of the current approach, but does not talk about coping.

In Section 4.1, we stated that the inspiration of this work has been Dastani and Lorini (2012). However, although the syntax of the two systems is rather similar, there are significant semantic differences. This is especially true for the evaluation of formulae containing dynamic operators: On this point, we take the classical PDL (Propositional Dynamic Logic) approach of using relations on the set of worlds for modeling the effects of a physical actions, whereas Dastani and Lorini (2012) take a syntactic approach similar to that of Situation Calculus (Reiter, 2001). On the syntax level, in addition to the multi-agent flavor, the language has been extended with a new kind of actions: the converse of physical actions. They allow for reversal of the effects of physical actions, and as a consequence, for reasoning about the state of the world before the execution of an action: a feature of crucial importance to some of the already discussed cognitive components of the moral emotions, such as responsibility and blame. Furthermore, sensing actions as found in Dastani and Lorini (2012) have been removed from the formalism for ease of exposition and their irrelevance to the emotion elicitation conditions.

Another influencing work on the topic has been Steunebrink et al. (2009). Inspired by Frijda (1986), it provides a formal model of emotions extended with intensities and action tendencies. Unlike the present approach, 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 provide any decidability results or axiomatization, whereas the current work does provide axiomatization which in future work can be analyzed about decidability.

Finally, Gratch and Marsella (2004) propose a computational model of emotions which incorporates both emotion intensities and coping. 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 have formalized the elicitation conditions and coping strategies of a set of socially-grounded emotions, dubbed moral. The formalization is based on appraisal theories of emotion and the CAD Triad Hypothesis, and is grounded in the DMAL-GA (a multi-agent BDI logic) framework. In this system, emotions are defined based on agents' actions and attitudes (including graded beliefs, goals and intentions). An important feature of the framework is the quantitative aspect of emotions: intensities are function of belief and goal strengths. The moral aspect of the modeled emotions is based on Schweder's ethics, and is represented using concepts from the content of the agents' beliefs and goals (harm, contamination, social significance). 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, as well as providing decidability results. 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

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

Anderson, A. C. and Bushman, J. B. (2002). Human aggression. *Psychology*, 53.

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

Averill, R. J. (1982). Anger and aggression: An essay on emotion.

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

Blackburn, P., De Rijke, M., and Venema, Y. (2002). *Modal logic*, volume 53. Cambridge Univ Pr.

Blackburn, S. (1998). *Ruling passions*. Clarendon Press Oxford.

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

Buss, H. A. (1962). The psychology of aggression. *The Journal of Nervous and Mental Disease*, 135:180–181.

Cleckley, M. H. (1964). *The mask of sanity: An attempt to clarify some issues about the so called psychopathic personality.* Aware Journalism.

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

Conte, R. and Castelfranchi, C. (1995). Understanding the functions of norms in social groups through simulation. *Artificial societies: The computer simulation of social life*.

Damasio, A. (2005). Descartes'error: Emotion, reason, and the human brain.

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

Dastani, M. and Meyer, C. J. J. (2006). Programming agents with emotions. pages 215–219.

Dubreuil, B. and Grégoire, J.-F. (2013). Are moral norms distinct from social norms? a critical assessment of jon elster and cristina bicchieri. *Theory and Decision*, 75:137–152.

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

Elster, J. (1999). *Alchemies of the Mind*. Cambridge Univ Press.

Fischer, M. J. and Ladner, R. E. (1979). Propositional dynamic logic of regular programs. *Journal of computer and system sciences*, 18(2):194–211.

Frijda, H. N. (1986). *The emotions*. Cambridge Univ Pr.

Gewirth, A. (1981). *Reason and morality*. University of Chicago Press.

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

Haidt, J. (2003). The moral emotions. *Handbook of affective sciences*, pages 852–870.

Hare, D. R. and Hart, D. S. (1993). Psychopathy, mental disorder, and crime.

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

Hoek, W. (2001). Logical foundations of agent-based computing. 2086.

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

Laverny, N. and Lang, J. (2005). From knowledge-based programs to graded belief-based programs, part i: On-line reasoning*. *Synthese*, 147:277–321.

Lazarus, S. R. (1991). *Emotion and adaptation.* Oxford University Press, USA.

Lazarus, S. R. and Folkman, S. (1984). *Stress, appraisal, and coping.* Springer Publishing Company.

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

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

Lorini, E. and Schwarzentruber, F. (2010). A logic for reasoning about counterfactual emotions. *Artificial Intelligence.*

Oatley, K. and Johnson-Laird, N. P. (1996). 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*, pages 363–393.

Ohbuchi, K.-i., Kameda, M., and Agarie, N. (1989). Apology as aggression control: its role in mediating appraisal of and response to harm. *Journal of personality and social psychology*, 56(2):219.

Ortony, A., Clore, L. G., and Collins, A. (1990). *The cognitive structure of emotions.* Cambridge Univ Pr.

Parikh, R. (1978). *The completeness of propositional dynamic logic.* Springer.

Plutchik, R. (1980). *Emotion: A psychoevolutionary synthesis.* Harper & Row New York.

Prinz, J. (2007). *The emotional construction of morals.* Oxford University Press.

Rao, S. A. and Georgeff, P. M. (1991). Modeling rational agents within a bdi-architecture. *KR*, 91:473–484.

Redmond, S. (2014). *Celebrity and the media.* Palgrave Macmillan.

Reiter, R. (2001). *Knowledge in action: logical foundations for specifying and implementing dynamical systems.* MIT press.

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

Rozin, P., Haidt, J., and McCauley, R. C. (1999a). 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*, pages 429–445.

Rozin, P., Haidt, J., and McCauley, R. C. (2008). 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*, pages 757–776.

Rozin, P., Lowery, L., Imada, S., and Haidt, J. (1999b). 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.

Scherer, R. K. (2001). Appraisal considered as a process of multilevel sequential checking: A component process approach. *Appraisal processes in emotion: Theory, methods, research*, 92.

Scherer, R. K., Schorr, A., and Johnstone, T. (2001). *Appraisal processes in emotion: Theory, methods, research.* Oxford University Press, USA.

Schmidt, R. and Tishkovsky, D. (2002). Multi-agent logics of dynamic belief and knowledge. 2424.

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

Sloman, A. and Croucher, M. (1981). Why robots will have emotions. Proc 7th Int. Joint Conf. on AI.

Spohn, W. (1988). Ordinal conditional functions: A dynamic theory of epistemic states. *Causation in decision, belief change, and statistics*, 2:105–134.

Staller, A. and Petta, P. (2001). Introducing emotions into the computational study of social norms. *JOURNAL OF ARTIFICIAL SOCIETIES AND SOCIAL SIMULATION*, 4.

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

Tomasello, M. and Vaish, A. (2013). Origins of human cooperation and morality. *Annual review of psychology*, 64:231–255.

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

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

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