Other-Condemning Anger = Blaming Accountable Agents for Frustrated Intentions (PRIMA 2017)

Mehdi Dastani, Emiliano Lorini, John-Jules Meyer, Alexander Pankov

Other-Condemning Anger = Blaming Accountable Agents for Unattainable Desires

Mehdi Dastani[1(✉)], Emiliano Lorini[2], John-Jules Meyer[1], and Alexander Pankov[1]

[1] Utrecht University, Utrecht, The Netherlands
{m.m.dastani,j.j.c.meyer}@uu.nl, a.pankov@students.uu.nl
[2] IRIT-CNRS, Toulouse, France
emiliano.lorini@irit.fr

Abstract. This paper provides a formalization of the other-condemning anger emotion which is a social type of anger triggered by the behaviour of other agents. Other-condemning anger responds to frustration of committed goals by others, and motivates goal-congruent behavior towards the blameworthy agents. Understanding this type of anger is crucial for modelling human behavior in social settings as well as designing socially aware artificial systems. We utilize existing psychological theories on other-condemning anger and propose a logical framework to formally specify this emotion. The logical framework is based on dynamic multi-agent logic with graded cognitive attitudes.

1 Introduction

Other-condemning anger is a reaction to the frustration of goals to which agents are committed, and motivates goal congruent behavior towards the agents believed to be accountable for the goal frustration [7,10,15,21]. Imagine someone who has to upload his paper in a submission system just before the deadline, but notices that the Internet connection is broken for some maintenance operations without a notice. The frustration of not having submitted the paper makes the author angry and motivates him to write a letter of complaint to the Internet company. Situations like this may also occur for autonomous software agents where similar responses are desirable, not only because of the believability of the agents' behaviours, but also because of the efficiency and effectiveness of goal-congruent responses. Imagine a situation where autonomous robots commit themselves to transport containers from one place to another in an environment such as harbours. A robot R_1 that aims at picking up its container at a designated position may notice the container is removed by another robot R_2. A desirable response of the robot R_1 would be to send a request to the robot R_2, who is believed by R_1 to be accountable for the removal of the container, to make the container accessible to R_1 and/or to send a warning message to the manager of the environment to report this irregularity. We stress that it is the general function of anger, i.e., specific type of response to specific type of situation, that we aim at integrating in the model of autonomous agents, rather than

https://doi.org/10.1007/978-3-319-69131-2_2

the physiological aspects of anger that is characteristic to human body. In this sense, it is the coordination role of emotions in agents' behaviour that motivates our work. For autonomous software agents that interact in social settings, the other-condemning anger emotion can be considered as a behavioural pattern or a heuristic that steers their behaviours.

There are reasons to believe that emotions in general, and other-condemning anger in particular, play an important role in rational behavior and in maintaining social order within societies [4,6,8,23]. Although there have been some efforts in artificial intelligence to provide a precise specification of emotions in general [5,17,18,28], there has not been, to our knowledge, a precise and adequate specification dedicated to the other-condemning anger emotion. We follow psychological literature [7,10,15,20,25] that explain other-condemning emotions in terms of complex social constructs such as controllability, accountability and blameworthiness. These social concepts require an adequate formalization of notions such as actions, control, causality, and their relations with the agents' cognitive states. As the above robot example illustrates, the angry robot R_1 believes its transportation goal being frustrated and that this is due to the removal action of robot R_2 who had control over its removal action (in the sense that R_2 could have chosen not to remove the container) and thus accountable for the caused consequences (i.e., R_1 cannot accomplish its transportation goal). The overtly social nature (being concerned with other agents) of this type of anger and its potential to influence others' behavior, make them essential for modelling human-like social interaction and designing socially aware artificial systems, which can be used for example in entertainment and serious games, crowd simulations, and human-computer interaction.

This paper proposes a logical model of multi-agent systems (Sect. 3) in which agents are specified by means of their knowledge, beliefs, desires, intentions, and actions. The logical model allows us to formally specify agents' anger. We present a logical specification of the appraisal and coping processes involved in other-condemning anger (Sect. 4). The specification is provided gradually by first specifying the underlying social and cognitive concepts such as control, accountability, blameworthiness, beliefs, goals, intentions and actions. We distinguish two types of anger. The first type of anger, called *plain anger*, involves two agents and captures the setting where an agent's committed goals is frustrated by another agent. The second type of anger, called *social anger*, involves three agents and captures the situation where the first agent gets angry at the second agent because the second agent harms a third agent who is in some social relation with the first agent. For social anger, we assume some social rules the existence of which are due to (or depend on) some norms or organisation governing the multi-agent environment. These assumed social rules may relate the goals of the first and the third agents such that the frustration of the third agent's goals by the second agent indirectly frustrates the goals of the first agent. For example, consider an extension of the robot example with a new manager agent that is responsible for the distribution and accomplishment of the transportation goals of all transport robots, including robot R_1. In this setting, the manager agent

and robot R_1 are in an organisational setting where the achievement of the transportation goals of R_1 may contribute to the achievement of the manager goals. If R_2 frustrates the goal of R_1, then R_2 will indirectly and through the existence of the social rule frustrate the manager's goals and therefore make this agent angry. The theoretical and empirical supports for our formalization are derived from cognitive psychology [7,10,15,21,25,26]. We first provide an informal description of the other-condemning anger emotion, followed by a presentation of the syntax, semantics, axiomatization and decidability of a *dynamic multi-agent logic of graded attitudes*, which will be used to ground the informal description of the anger emotion.

2 Other-Condemning Anger

Other-condemning anger is commonly viewed as a negatively valenced reaction to the actions of other agents [21]. It is an instance of the *other-condemning* emotions [10], and triggered by frustration of a goal commitment [15,21]. In our robot example, the goal that the transport robot is committed to, i.e., the goal to have the container at its designated position, is frustrated. This broad view of other-condemning anger has been refined by emotion theories to distinguish it from other negative emotions such as sadness, guilt and remorse that also can arise from *goal incongruence*.

Most emotion theories distinguish other-condemning anger from other negative emotions by attributing *blame* for goal incongruence to other agents [7,15]. As a result, blame towards someone else becomes a necessary condition for other-condemning anger, for without the attribution of blame we can expect an emotion such as sadness. What does it mean, however, to blame someone for goal incongruence? According to [15], blame is an appraisal based on *accountability* and imputed *control*. To attribute accountability is to know who caused the relevant goal-frustrating event, and to attribute control is to believe that the accountable agent could have acted differently without causing the goal-incongruence. In our example, robot R_1 believes that robot R_2 is accountable for removing the container and that R_2 has the choice not to remove the container. According to Lazarus, anger is triggered if, in addition to above conditions, the *coping potential* (the evaluation of the possible responses) is viable. In our running example, the robot R_1 can send a request to R_2 to make the container accessible to R_1 and/or to report this irregularity to the environment manager. The prototypical *coping strategy* of other-condemning anger generally involves attack, or other means of getting back at the blameworthy agent, with the intention of restoring a goal congruent state of affairs [7,13,15].

The second type of other-condemning anger, i.e., social anger, is similar to what is often called *moral anger*, where a first agent is morally angry at a second agent because the second agent harms a third agent by violating some moral norm [23]. In such cases, an agent can rightfully be angry without any of his own goals being directly frustrated. In our extended example, the manager agent, which may be a software agent as well, may get angry at robot R_2, because R_2

has frustrated the goal of R_1. The actual reason for an agent to get angry at the third agent is the existence of a social rule that prescribes and promotes cooperation. For example, in case of human agents the reason for being angry can be the violation of a moral rule that prescribes agents not to harm the autonomy of each other. The typical coping strategy for social anger is similar to the coping strategy for the plain anger and promotes socially congruent behavior. Combining this aspect of social anger with the elicitation conditions of plain anger allows us to informally describe other-condemning anger in psychological terms as follows:

Displeasure from thwarting of a personal goal, or a social rule aimed at preserving the goal commitment of other 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 favouring attack towards *the blameworthy agent.*

3 The Logical Framework

In this section we define the logic DMAL-GA (*Dynamic Multi-Agent Logic of Graded Beliefs*). This logic serves as the basis for the formalization of the other-condemning anger emotion. The logic is a multi-agent extension of the DL-GA logic developed by Dastani and Lorini in [5]. However, there are substantial differences in the syntax and semantics of the system. Most importantly, here atomic actions are considered as special type of assignments, whereas Dastani and Lorini take an approach similar to that of Situation Calculus [24]. The consideration allows us to model the converse of actions in DMAL-GA and to define it as the reverse of the effects of atomic actions. The converse of actions is a prerequisite for formalizing concepts such as accountability and blame, which in turn play a central role in defining the other-condemning anger emotion. In particular, for the characterization of the accountability we need to refer to the action that has just occurred and look at the state from which the action is performed. The converse of the actions allows us to do this. We also need to refer to the actions that can possibly occur next.

Syntax. We assume a non-empty finite set of agents $Agt = \{1, \ldots, n\}$ and a non-empty finite set of atomic propositions $Atm = \{p, q, \ldots\}$ describing the environment in which the agents act. Because we aim at modelling the intensity of the anger emotion, we also assume a non-empty finite set of natural numbers $Num^+ = \{x \in \mathbb{N} : 0 \leq x \leq max\}$ with $max \in \mathbb{N}\backslash\{0\}$. Let also $Num^- = \{-x : x \in Num^+\backslash\{0\}\}$ and $Num = Num^+ \cup Num^-$. The set of literals is defined in the usual way as follows: $Lit = Atm \cup \{\neg p : p \in Atm\}$. Let $Act = \{toggle(p) : p \in Atm\}$ be the set of atomic actions. Specifically, $toggle(p)$ should be read as "toggle the truth value of p", and understood as changing the truth value of p. This construct represents a simple notion of atomic action consisting in changing the truth value of a specific atomic proposition. We assume that changing the truth value of an atomic proposition in not always feasible such that a specific toggle action may not be available/executable at every state. For notational convenience, elements of Act are denoted by $a, b, \ldots$ For every agent

$i \in Agt$, agent i's set of events is defined to be $Evt_i = \{(i, a) : a \in Act\}$ and the set of all agents' events is defined to be $Evt = \bigcup_{i \in Agt} Evt_i$. An event (i, a) indicates that action a is performed by agent i. For notational convenience, elements of Evt are denoted by $e, e', \ldots$ Following [22], we use $-e$ to denote the converse of event $e \in Evt$, which allows us to describe properties of states before an atomic action of type $toggle(p)$ has been performed by an agent. We use $\alpha, \beta, \ldots$ to denote an event or its converse, i.e., $\alpha, \beta, \ldots$ denote the elements of $Evt \cup \{-e : e \in Evt\}$. We define $SeqEvt$ to be the set of all possible finite sequences of events or their converse. Elements of $SeqEvt$ are denoted by $\epsilon, \epsilon', \ldots$ The empty sequence of events is denoted by nil.

The language $\mathcal{L}$ of the logic DMAL-GA is defined by the following grammar:

$$\varphi, \psi ::= p \mid \neg\varphi \mid \varphi \wedge \psi \mid exc_i^h \mid Des_i^k l \mid Int_i(\epsilon, a) \mid Fut(\epsilon, e) \mid Past(\epsilon, e) \mid K_i\varphi \mid [\alpha]\varphi$$

where p ranges over Atm, i ranges over Agt, h ranges over Num^+, k ranges over Num, a ranges over Act, e ranges over Evt, ϵ ranges over $SeqEvt$ and α ranges over $Evt \cup \{-e : e \in Evt\}$. The other Boolean constructions on formulae ($\vee$, $\rightarrow$, $\leftrightarrow$, $\top$ and $\bot$) are defined in the standard way using $\neg$ and $\wedge$.

The set of formulae contains special constructions exc_i^h, $Des_i^k l$ and $Int_i(\epsilon, a)$ which are used to represent agents' mental states. Formulae exc_i^h is used to identify the degree of exceptionality of a given world for a given agent i. Following [27], the worlds that are assigned the smallest numbers are the least exceptional and therefore the most plausible ones. 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 following we will use exc_i^h to define graded beliefs of agent i. 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, the following abbreviations are used in the rest of the paper: $AchG_i^k l \stackrel{def}{=} Des_i^k l$ and $AvdG_i^k l \stackrel{def}{=} Des_i^{-k} l$ for $k > 0$, where $AchG$ and $AvdG$ respectively stand for *achievement goal* and *avoidance goal*. Formulae $Int_i(\epsilon, a)$ represent the agents' intentions or commitments about atomic actions. Specifically, $Int_i(\epsilon, a)$ has to be read "after the sequence of events ϵ, agent i intends to perform action a".

The formulae $Fut(\epsilon, e)$ and $Past(\epsilon, e)$ represent the dynamics of the system by means of its action structure. They are introduced to refer to respectively the event (i.e., an action of an agent) that can possibly occur next in a state and the event that has just occurred in a state. In particular, $Fut(\epsilon, e)$ denotes the fact that event e is an option in the state reached after the execution of event sequence ϵ, and $Past(\epsilon, e)$ denotes the fact that event e has just occurred in the state reached after the execution of event sequence ϵ. These two formulae allow us to reason about options and performed actions after the execution of arbitrary sequence of events. The formula $Fut(\epsilon, e)$ has to be read as "the event e can possibly occur in the state reached by the sequence of events ϵ", while the formula $Past(\epsilon, e)$ has to be read as "the event e has just occurred in the state reached after the sequence of events ϵ". Note the use of nil in formulae $Fut(nil, e)$

and $Past(nil, e)$ which have the interpretation that the event e possibly occurs next to the "current state" and the event e has just occurred prior to the "current state", respectively.

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 of knowledge. Finally, the formula $[\alpha]\varphi$ covers the dynamic nature of the formalism by referring to the state of the world after the occurrence of an event or its converse. It should be read as "the occurrence of event α leads to φ" or "the occurrence of event α results in φ". For notational convenience, we use special dynamic operators of the form $\langle\langle e\rangle\rangle$ and $\langle\langle -e\rangle\rangle$ where $\langle\langle e\rangle\rangle\,\varphi$ and $\langle\langle -e\rangle\rangle\,\varphi$ have to be read as, respectively, "the event e is going to possibly occur next and φ will be true afterwards" and "the event e has just occurred and φ was true before":

$$\langle\langle e\rangle\rangle\,\varphi \stackrel{def}{=} Fut(nil, e) \wedge [e]\varphi \qquad \& \qquad \langle\langle -e\rangle\rangle\,\varphi \stackrel{def}{=} Past(nil, e) \wedge [-e]\varphi$$

We also define the concept of present-directed intention, denoted by $Int_i a$, that is, the intention to do the action a now:

$$Int_i a \stackrel{def}{=} Int_i(nil, a)$$

An important 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: $exc_i^{\leq k} \stackrel{def}{=} \bigvee_{0\leq l\leq k} exc_i^l$ for all $i \in Agt$ and for all $k \in Num^+$. Now, following [14,27], we define the following concept of belief:

$$B_i\varphi \stackrel{def}{=} K_i(exc_i^0 \rightarrow \varphi)$$

The formula $B_i\varphi$ says that agent i believes a formula φ *if and only if* φ is true in all worlds that are maximally plausible (or minimally exceptional) for the agent. We moreover define the following concept of graded belief, for $h > 0$:

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

The formula $B_i^{\geq h}\varphi$ says that agent i believes a formula φ with strength at least h *if and only if* φ is true in all worlds with exceptionality degree for the agent of less than h. Finally, we define the following concept of *exact* degree of belief, for $h > 0$:

$$B_i^h\varphi \stackrel{def}{=} \begin{cases} B_i^{\geq h}\varphi \wedge \neg B_i^{\geq h+1}\varphi & \text{if } 0 < h < max \\ B_i^{\geq max}\varphi & \text{if } h = max \end{cases}$$

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

Models. The language $\mathcal{L}$ is interpreted relative to a possible world semantics with special functions that represent the dynamic structure of the model. These

functions are defined to ensure models with linear past and branching future, which are tree-like structures. Given a state of the model, these structures allow us to refer to the event that has just occurred and the events that can possible occur next. Specifically, the language $\mathcal{L}$ is interpreted on structures called DMAL-GA models.

Definition 1. *The tuple* $\mathfrak{M} = \Big(W, (\sim_i)_{i\in Agt}, (\mathcal{E}_i)_{i\in Agt}, (\mathcal{D}_i)_{i\in Agt}, (\mathcal{I}_i)_{i\in Agt}, \mathcal{F}, \mathcal{P}, \mathcal{V}\Big)$ *is a* DMAL-GA model *where:*

- W *is a nonempty set of worlds or states;*
- $\sim_i \subseteq W \times W$ *is an equivalence relation representing knowledge*
- $\mathcal{E}_i : W \longrightarrow Num^+$ *is a total function representing exceptionality degrees of states*
- $\mathcal{D}_i : W \times Lit \longrightarrow Num$ *is a total function representing desirability of facts*
- $\mathcal{I}_i : W \times SeqEvt \longrightarrow 2^{Act}$ *is a total function representing agents' intentions*
- $\mathcal{F} : W \times SeqEvt \longrightarrow 2^{Evt}$ *is a total function indicating future events;*
- $\mathcal{P} : W \times SeqEvt \longrightarrow Evt$ *is a partial function indicating past events;*
- $\mathcal{V} : W \longrightarrow 2^{Atm}$ *is a valuation function*

$\sim_i$ 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 $\mathcal{E}_i$ is the plausibility grading of the possible worlds for agent i, and is used to interpret the atomic formulae exc_i^h. $\mathcal{E}_i(w) = h$ means that, according to agent i, the world w has a degree of exceptionality h, or alternatively, degree of plausibility $max - h$. The function $\mathcal{E}_i$, 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. We assume that DMAL-GA models satisfy the following *normality* condition with respect to the $\mathcal{E}_i$ functions:

(Norm) for all $i \in Agt$ and for all $w \in W$, there is $v \in W$ s.t. $w \sim_i v$ and $\mathcal{E}_i(v) = 0$.

This condition ensures that the real world, the world with exceptionality zero, is among possible worlds. The function $\mathcal{D}_i$ is the desirability 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.

$\mathcal{I}_i(w, \epsilon)$ represents the set of actions that agent i intends to perform in the state that is reached after the sequence of events ϵ performed at world w. In other words, for every possible sequence of events ϵ and for every agent i, we describe the set of intentions that agent i will have in the state that is reached after this sequence. $\mathcal{I}_i(w, \epsilon) = \emptyset$ means that agent i will have no intention in the state that is reached after the sequence of events ϵ performed at world w.

$\mathcal{F}(w, \epsilon)$ and $\mathcal{P}(w, \epsilon)$ represent, respectively, the events which *can possibly* occur in the state reached after the sequence of events ϵ is performed at world w and the event which has just occurred in the state that is reached after the sequence of events ϵ is performed at world w. We call $\mathcal{F}$ and $\mathcal{P}$ *agenda* functions, given their similarity with the agenda function in [19]. $\mathcal{F}(w, \epsilon) = \emptyset$ means that no event can possibly occur in the state reached after the execution of the sequence ϵ at world w. If $\mathcal{P}(w, \epsilon)$ is undefined (since $\mathcal{P}$ is assumed to be a partial function), then it means that no event has just occurred in the state reached after the sequence of events ϵ is performed at world w. When ϵ is the empty sequence *nil*, then $\mathcal{F}(w, nil)$ and $\mathcal{P}(w, nil)$ denote, respectively, the events which *can possibly* occur at w (i.e., the options available at w) and the event which has just occurred at w (i.e., the event leads to w). Figure 1 illustrates how the dynamic structure of a *DMAL-GA* model can be specified by means of these two functions.

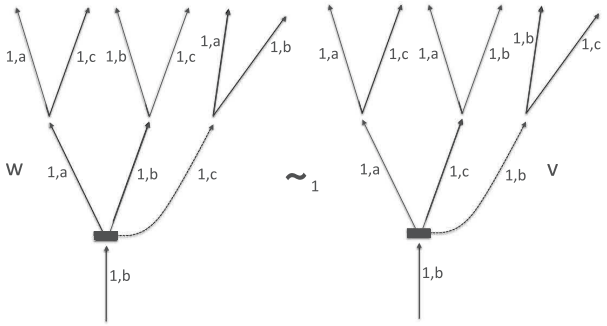

Fig. 1. Representation of the epistemic-temporal aspects of a *DMAL-GA* model

Figure 1 represents two worlds w and v that are indistinguishable from agent 1's perspective. Each world is associated with a particular evolution of the system that is specified by the functions $\mathcal{F}$ and $\mathcal{P}$. The small black rectangle in each world represents the reference point, which corresponds to the empty event sequence *nil*. Full arrows represent *possible* transitions (i.e., transitions corresponding to the execution of available actions) while dotted arrows represent counterfactual *impossible* transitions (i.e., transitions corresponding to the execution of non-available/non-executable actions). The following is a partial presentation of $\mathcal{F}$ and $\mathcal{P}$ applied to world w.

$$\mathcal{F}(w, nil) = \{(1, a), (1, b)\} \qquad \mathcal{P}(w, nil) = \{(1, b)\}$$
$$\mathcal{F}(w, (1, b)) = \{(1, b), (1, c)\} \qquad \mathcal{F}(w, (1, c)) = \{(1, a), (1, b)\}$$

It should be emphasized that Fig. 1 is not complete as it does not show all possible and impossible transitions. This is done to keep the figure simple and clear. Assuming the set of actions $Act = \{a, b, c\}$, other possible and impossible transitions should be drawn at each choice point. In particular, a complete figure should have all events at each choice point, either as a possible or as an impossible transition.

We assume that DMAL-GA models satisfy the following *equivalence* condition for intention and agenda functions:

(Equiv) for all $i \in Agt$, $\epsilon, \epsilon' \in SeqEvt$, $e \in Evt$, and $w \in W$, $\mathcal{I}_i(w, \epsilon;e;-e;\epsilon') = \mathcal{I}_i(w, \epsilon;\epsilon') = \mathcal{I}_i(w, \epsilon;-e;e;\epsilon')$, $\mathcal{F}(w, \epsilon;e;-e;\epsilon') = \mathcal{F}(w, \epsilon;\epsilon') = \mathcal{F}(w, \epsilon;-e;e;\epsilon')$ and $\mathcal{P}(w, \epsilon;e;-e;\epsilon') = \mathcal{P}(w, \epsilon;\epsilon') = \mathcal{P}(w, \epsilon;-e;e;\epsilon')$.

The previous constraint just means that the consecutive occurrences of a event e and its corresponding converse event $-e$ is ineffective. We also assume that DMAL-GA models satisfy the following *temporal coherence* condition between events and their converse counterparts:

(Coh) for all $\epsilon \in SeqEvt$, for all $e \in Evt$, and for all $w \in W$, $e \in \mathcal{F}(w, \epsilon;-e)$, and $\mathcal{P}(w, \epsilon;e) = e$.

For instance, suppose that $\epsilon = nil$. Then, the previous condition says that (i) before e has occurred, it was possible that e occurs, and (ii) after e occurs, it is the case that e has just occurred.

Given the structures for interpreting the DMAL-GA language, we specify truth conditions of formulae.

Definition 2. *Given a model* $\mathfrak{M} = \Big(W, (\sim_i)_{i\in Agt}, (\mathcal{E}_i)_{i\in Agt}(\mathcal{D}_i)_{i\in Agt}, (\mathcal{I}_i)_{i\in Agt}, \mathcal{F}, \mathcal{P}, \mathcal{V}\Big)$ *The* truth conditions *of formulae are defined as follows:*

- $\mathfrak{M}, w \models p$ iff $p \in \mathcal{V}(w)$;
- $\mathfrak{M}, w \models Des_i^k l$ iff $\mathcal{D}_i(w, l) = k$;
- $\mathfrak{M}, w \models exc_i^h$ iff $\mathcal{E}_i(w) = h$;
- $\mathfrak{M}, w \models Int_i(\epsilon, a)$ iff $a \in \mathcal{I}_i(w, \epsilon)$;
- $\mathfrak{M}, w \models Fut(\epsilon, e)$ iff $e \in \mathcal{F}(w, \epsilon)$;
- $\mathfrak{M}, w \models Past(\epsilon, e)$ iff $\mathcal{P}(w, \epsilon) = e$;
- $\mathfrak{M}, w \models \neg\varphi$ 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* $v \in W$ *s.t.* $v \sim_i w$;
- $\mathfrak{M}, w \models [\alpha]\varphi$ iff $\mathfrak{M}^\alpha, w \models \varphi$;

where $\mathfrak{M}^\alpha$ *is defined according to Definition 3.*

We write $\models \varphi$ to say that φ is valid and say that φ is satisfiable if $\neg\varphi$ is not valid. Before defining the updated model $\mathfrak{M}^\alpha$ let us briefly illustrate the interpretation epistemic formulas by means of the model given in Fig. 1. We have the following:

$$\mathfrak{M}, w \models K_1 Fut(nil, (1, a))$$
$$\mathfrak{M}, w \models K_1 Past(nil, (1, b))$$
$$\mathfrak{M}, w \models \neg K_1 Fut(nil, (1, b)) \ \wedge \ \neg K_1 \neg Fut(nil, (1, b))$$
$$\mathfrak{M}, w \models K_1 Fut((1, b), (1, b))$$

For instance, $\mathfrak{M}, w \models \neg K_1 Fut(nil, (1, b)) \wedge \neg K_1 \neg Fut(nil, (1, b))$ means that at w in the model of Fig. 1, agent 1 is uncertain whether she can possibly perform action b. Moreover, at w in the model of Fig. 1, agent 1 knows that, after having performed action b, she can possibly perform it again.

Definition 3. *Given a model* $\mathfrak{M} = \Big(W, (\sim_i)_{i\in Agt}, (\mathcal{E}_i)_{i\in Agt}(\mathcal{D}_i)_{i\in Agt}, (\mathcal{I}_i)_{i\in Agt}, \mathcal{F}, \mathcal{P}, \mathcal{V}\Big)$ *and an event* $\alpha \in \{(i, toggle(p)), -(i, toggle(p)) \ : \ i \in Agt \text{ and } p \in Atm\}$*, the update of* $\mathfrak{M}$ *by* α*, is* $\mathfrak{M}^\alpha = \Big(W, (\sim_i)_{i\in Agt}, (\mathcal{E}_i)_{i\in Agt}, (\mathcal{D}_i)_{i\in Agt}, (\mathcal{I}_i^\alpha)_{i\in Agt}, \mathcal{F}^\alpha, \mathcal{P}^\alpha, \mathcal{V}^\alpha\Big)$ *where for all* $w \in W$*,* $i \in Agt$ *and* $\epsilon \in SeqEvt$*:*

$$\mathcal{I}_i^\alpha(w,\epsilon) = \mathcal{I}_i(w,\alpha;\epsilon) \qquad \mathcal{F}^\alpha(w,\epsilon) = \mathcal{F}(w,\alpha;\epsilon) \qquad \mathcal{P}^\alpha(w,\epsilon) = \mathcal{P}(w,\alpha;\epsilon)$$

$$\mathcal{V}^\alpha(w) = \begin{cases} \mathcal{V}(w) \cup \{p\} & \text{if } p \notin \mathcal{V}(w) \\ \mathcal{V}(w) \setminus \{p\} & \text{if } p \in \mathcal{V}(w) \end{cases}$$

The update of model $\mathfrak{M}$ by the event α just consists in: (i) updating the intention functions $\mathcal{I}_i$ as well as the agenda functions $\mathcal{F}$ and $\mathcal{P}$, and (ii) modifying the valuation function $\mathcal{V}$. In particular, if $a = toggle(p)$, then the truth value of p should be toggled. To illustrate the update of $\mathcal{F}$ let us consider an example. Suppose $F(w, \alpha; \beta) = \{e, e'\}$. This means that after performing the sequence of events $\alpha; \beta$ in state w we arrive in a state in which two events can possibly occur, namely, e and e'. The event α makes us to move one step right along the sequence $\alpha; \beta$ and to eliminate the first element (α) from it. Therefore, $F^\alpha(w, \beta) = F(w, \alpha; \beta) = \{e, e'\}$. The updates of $\mathcal{I}_i$ and $\mathcal{P}$ can be illustrated in a similar way. The following proposition guarantees that $\mathfrak{M}^\alpha$ is indeed a DMAL-GA model.

Proposition 1. *Let* $\mathfrak{M}$ *be a DMAL-GA model and* $\alpha \in Evt$*. Then,* $\mathfrak{M}^\alpha$ *is a DMAL-GA model too.*

We now present some interesting validities of the logic DMAL-GA.

Proposition 2. *For all* $i \in Agt$ *and for all* $e \in Evt$*, we have:*

$$\models \varphi \leftrightarrow [e][-e]\varphi \tag{1}$$
$$\models \varphi \leftrightarrow [-e][e]\varphi \tag{2}$$

Validities (1) and (2) in the preceding proposition capture the dependence between events and their converse counterparts. Similarly to [5], we have the following set of validities related to beliefs.

Proposition 3. *For all* $i \in Agt$*, and for all* $h, k \in Num^+$ *such that* $h \geq 1$ *and* $k \geq 1$*:*

$$\models K_i\varphi \rightarrow B_i^{\geq h}\varphi \tag{3}$$
$$\models B_i\varphi \leftrightarrow B_i^{\geq 1}\varphi \tag{4}$$
$$\models \neg(B_i\varphi \wedge B_i\neg\varphi) \tag{5}$$
$$\models (B_i^{\geq h}\varphi \wedge B_i^{\geq k}\psi) \rightarrow B_i^{\geq min[h,k]}(\varphi \wedge \psi) \tag{6}$$
$$\models (B_i^{\geq h}\varphi \wedge B_i^{\geq k}\psi) \rightarrow B_i^{\geq max[h,k]}(\varphi \vee \psi) \tag{7}$$

3.1 Axiomatization and Decidability

In this section we present an axiomatics and a decidability result for the logic DMAL-GA. The following theorem establish the axiomatization of the logic.

Theorem 1. *The logic DMAL-GA is axiomatized as an extension of the proposition multimodal logic $S5^n$ for the epistemic operators K_i with: (i) a theory describing the constraints imposed on agents' mental states and actions given in Fig. 2, (ii) reduction axioms of the dynamic operators $[\alpha]$ given in Fig. 3, and (iii) the following rule of replacement of equivalents:*

$$\frac{\psi_1 \leftrightarrow \psi_2}{\varphi \leftrightarrow \varphi[\psi_1/\psi_2]}$$

Proof (*Sketch*). To prove soundness of the principles in Figs. 2 and 3 is just a routine exercise. The completeness proof proceeds as follows. By standard canonical model argument, it is routine to show that the axioms and rules of inference of the multimodal logic $S5^n$ for every epistemic operator K_i together with the principles in Fig. 2 and all principles of classical propositional logic provide a complete axiomatization for the fragment of DMAL-GA with no dynamic operators. Let us call DMAL-GA$^-$ this fragment and $\mathcal{L}^-$ its corresponding language. Call *red* the mapping which iteratively applies the equivalences in Figure 3 from the left to the right, starting from one of the innermost modal operators. *red* pushes the dynamic operators inside the formula, and finally eliminates them when facing an atomic formula. By the rule of replacement of equivalents, it is routine to prove that $red(\varphi) \leftrightarrow \varphi$ is DMAL-GA valid. Now, suppose φ is DMAL-GA valid. Hence, $red(\varphi)$ is valid in DMAL-GA. By the completeness of DMAL-GA$^-$, $red(\varphi)$ is also provable there. DMAL-GA being a conservative extension of DMAL-GA$^-$, $red(\varphi)$ is provable in DMAL-GA, too. As the reduction axioms and the rule of replacement of equivalents are part of our axiomatics, the formula φ must also be provable in DMAL-GA. □

$\bigvee_{h \in Num^+} exc_i^h$	$\bigvee_{k \in Num} Des_i^k l$
$\neg K_i \neg exc_i^0$	$Past(\epsilon; e, e)$
$exc_i^h \rightarrow \neg exc_i^{h'}$ if $h \neq h'$	$Des_i^k l \rightarrow \neg Des_i^{k'} l$ if $k \neq k'$
$Past(\epsilon, e) \rightarrow \neg Past(\epsilon, e')$ if $e \neq e'$	$Int_i(\epsilon; e'; -e'; \epsilon', a) \leftrightarrow Int_i(\epsilon; \epsilon', a)$
$Fut(\epsilon; e'; -e'; \epsilon', e) \leftrightarrow Fut(\epsilon; \epsilon', e)$	$Past(\epsilon; e'; -e'; \epsilon', e) \leftrightarrow Past(\epsilon; \epsilon', e)$
$Int_i(\epsilon; -e'; e'; \epsilon', a) \leftrightarrow Int_i(\epsilon; \epsilon', a)$	$Fut(\epsilon; -e'; e'; \epsilon', e) \leftrightarrow Fut(\epsilon; \epsilon', e)$
$Past(\epsilon; -e'; e'; \epsilon', e) \leftrightarrow Past(\epsilon; \epsilon', e)$	$Fut(\epsilon; -e, e)$

Fig. 2. Theory of the agents' mental states and actions

$$[\alpha]p \leftrightarrow \begin{cases} \neg p & \text{if } \alpha \in \{(i, toggle(p)), -(i, toggle(p))\} \\ & \text{for some } i \in Agt \\ p & \text{otherwise} \end{cases}$$

$$[\alpha]Int_i(\epsilon, a) \leftrightarrow Int_i(\alpha;\epsilon, a) \qquad [\alpha]Fut(\epsilon, e) \leftrightarrow Fut(\alpha;\epsilon, e)$$

$$[\alpha]Past(\epsilon, e) \leftrightarrow Past(\alpha;\epsilon, e) \qquad [\alpha]exc_i^h \leftrightarrow exc_i^h$$

$$[\alpha]Des_i^k l \leftrightarrow Des_i^k l \qquad [\alpha]\neg\varphi \leftrightarrow \neg[\alpha]\varphi$$

$$[\alpha](\varphi_1 \wedge \varphi_2) \leftrightarrow ([\alpha]\varphi_1 \wedge [\alpha]\varphi_2) \qquad [\alpha]K_i\varphi \leftrightarrow K_i[\alpha]\varphi$$

Fig. 3. Reduction axiom schemas for the operators $[\alpha]$

Theorem 2. *The satisfiability problem of the logic DMAL-GA is decidable.*

Proof. (*Sketch*) Hardness just follows from the fact that the satisfiability problem of the multimodal logic S5^n is PSPACE-hard [11] and that DMAL-GA extends the multimodal logic S5^n. As in the proof of Theorem 1 above let us call DMAL-GA$^-$ the fragment of DMAL-GA with no dynamic operators and let red be the mapping which allows us to eliminate the dynamic operators. The problem of checking the validity of a DMAL-GA$^-$ formula φ is reducible to the problem of *global* logical consequence in S5^n with a finite set of global axioms Γ, Γ includes all principles in Fig. 2 which are relevant for $Sub(\varphi)$, the set of subformulas of φ. That is, we have $\models_{DMAL-GA^-} \varphi$ if and only if $\Gamma \models_{S5^n} \varphi$. The problem of global logical consequence in S5^n with a finite set of global axioms is reducible to the problem of validity checking in S5^n and these two problems are decidable. Thus, it follows that the problem of validity checking in the logic DMAL-GA$^-$ is decidable too. From the fact that $red(\varphi) \leftrightarrow \varphi$ is DMAL-GA valid and the fact that DMAL-GA is a conservative extension of DMAL-GA$^-$, it follows that red provides an effective procedure for reducing a DMAL-GA formula φ into an equivalent DMAL-GA$^-$ formula $red(\varphi)$. Thus, since the problem of validity checking in DMAL-GA$^-$ is decidable, it follows that the problem of validity checking in DMAL-GA is decidable too. □

4 Formalizing Anger

We are now well-equipped to formalize the other-condemning anger emotion. This requires translating our informal definitions into the language of DMAL-GA. The appraisal behind the elicitation of other-condemning anger is *blame*. According to the appraisal theories of emotion, there are two more basic concepts behind blame: *accountability* and *control*. For an agent to attribute blame to someone for something, he has to determine, first, if the other agent is accountable for (or having 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, for agent i to attribute blame to agent j for a state of affairs, i must believe

that j is (1) accountable for the state of affairs, and (2) had control over it (or was able to prevent it).

We first define what it means to have control over a state of affairs φ, denoted as $Control_i(\varphi)$ and read as "agent i has control over state of affairs φ". We say that agent i has control over state of affairs φ *if and only if* there exists an event $e \in Evt_i$ such that e can possibly occur next (i.e., if e is an option) and the occurrence of e maintains the truth value of φ. In other words, "agent i has control over the state of affairs φ if i is able to maintain its truth value". Formally,

$$Control_i(\varphi) \stackrel{def}{=} (\varphi \wedge \bigvee_{e \in Evt_i} \langle\langle e \rangle\rangle \varphi) \vee (\neg\varphi \wedge \bigvee_{e \in Evt_i} \langle\langle e \rangle\rangle \neg\varphi)$$

An instance of the $Control_i(\varphi)$ formula is $Control_{R_2}(XatY)$, where R_2 is one of the robots from our example, and $XatY$ denotes the state of affairs where container X is at some spacial location Y. This formula states that R_2 has control over the position of container X because it can ensure to maintain the position of X, i.e., if X is currently at Y then I can ensure that X is at Y in the next state and if X is currently not at Y then X will not be at Y at the next state.

For the notion of accountability, we assume an agent being accountable for a state of affair if and only if the state of affair is realized because of the action of the agent. In order to express accountability, we use the formula $Account_i(a, \varphi)$ which should be read as "agent i is accountable for (has caused) φ by doing a". By definition, this is the case *if and only if* φ is true now and was not true before event (i, a) occurred,[1] i.e.,

$$Account_i(a, \varphi) \stackrel{def}{=} \varphi \wedge \langle\langle -(i, a) \rangle\rangle \neg\varphi$$

An instance of this formula is $Account_{R_2}(pickXatY, \neg XatY)$, where $pickXatY$ denotes an action (or a complete plan) for picking up container X from location Y. This formula states that R_2 can be held accountable for container X not being at position Y because R_2 has picked up X from Y. The appraisal of blame can now be defined.

$$Blame^k_{i,j}(a, \varphi) \stackrel{def}{=} B^k_i(Account_j(a, \varphi) \wedge [-(j, a)]Control_j(\varphi))$$

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 agent j is accountable for φ by doing a, and that before the event (j, a), j had control over φ. Going back to our robot example, we can speak of R_1 blaming R_2 for picking up the container from its location. Formally expressed as $Blame^k_{R_1,R_2}(pickXatY, XatY)$, for some $k > 0$. It is important to stress that we define blame without any negative connotations. Instead, it is viewed as a belief about the accountability of an agent for, and his control over,

[1] We assume that only one agent acts at each moment.

a given state of affairs. This is much in the spirit of how Lazarus talks about blame in his discussion on anger [15, p. 219].

Before defining other-condemning anger, we need a way of talking about the practical possibility of an agent to make a formula true. For this we use:

$$Pos_i(\varphi) \stackrel{def}{=} \bigvee_{e \in Evt_i} \langle\langle e \rangle\rangle \varphi$$

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 event $e \in Evt_i$ such that e can possibly occur next and φ will be true after its occurrence. In our example, this can be understood as robot R_1 being able to obtain the removed container, by say, sending a message to R_2 requesting the container to be returned, and thus making the formula $Pos_{R_1}(R_1 holds X)$ true.

4.1 Plain Anger

We can now define plain anger in the logic DMAL-GA as follows. $Anger^l_{i,j}(a, \varphi, b) \stackrel{def}{=}$

$\bigvee_{l=merge(h,k)}(AchG^k_i(\varphi) \wedge Int_i b \wedge Blame^h_{i,j}(a, \neg \langle\langle (i,b) \rangle\rangle \varphi) \wedge B_i Pos_i(\varphi))$ where *merge* is a monotonically increasing function of its two arguments, h and k.[2] Its range being the set $EmoInt = \{y : \exists x_1, x_2 \in Num^+ \text{ s.t. } merge(x_1, x_2) = y\}$. 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, intends to do b, and blames agent j for performing the action a, thus preventing him from achieving φ by doing b".

Let us dissect the definition of plain anger and see how it matches our informal definition. The first conjunct, $AchG^k_i(\varphi)$, captures the prototypical feature of any emotion, i.e., to be about goal state φ. The next two conjuncts, $Int_i b$ and $Blame^h_{i,j}(a, \neg \langle\langle (i,b) \rangle\rangle \varphi)$, 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 belief of agent i not to be able to achieve his goal by executing the intended plan b, which is expressed by $\neg \langle\langle (i,b) \rangle\rangle \varphi$, although i believes this was possible before action a was performed by agent j, which is expressed by $[-(j,a)] \langle\langle (i,b) \rangle\rangle \varphi$. This observation about the agent's attitudes is expressed as the following simple proposition:

Proposition 4. *Let $\mathfrak{M}$ be a DMAL-GA model, $w \in W$; $a, b \in Act$; $i, j \in Agt$; $l \in EmoInt$ and $\varphi \in Lit$. If $\mathfrak{M}, w \models Anger^l_{i,j}(a, \varphi, b)$, then*

$\mathfrak{M}, w \models B^h_i(\neg \langle\langle (i,b) \rangle\rangle \varphi \wedge [-(j,a)] \langle\langle (i,b) \rangle\rangle \varphi)$ *for some $h \in Num^+$*

[2] As suggested by some appraisal theorists [15,21], 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 $h \times k$.

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. Note that this practical possibility of achieving φ does not involve performing b, for agent i believes, according to Proposition 4, of not being able to achieve φ by means of b, i.e., $B_i^h \neg \langle\langle (i,b) \rangle\rangle \varphi$. For our robot example we can assume the following facts to hold:

- $AchG^k_{R_1}(R_1 holds X)$: robot R_1 wants with strength k to obtain container X;
- $Int_{R_1}(pickXatY)$: robot R_1 intends to pick up container X from its location Y;
- $B_{R_1} Pos_{R_1}(R_1 holds X))$: robot R_1 believes it has the practical possibility to achieve its goal of obtaining container X;
- $B^h_{R_1} Account_{R_2}(pickXatY, \neg \langle\langle (R_1, pickXatY) \rangle\rangle R_1 holds X)$: robot R_1 believes with strength h that robot R_2 is accountable for R_1 not being able to obtain container X by picking it up from location Y;
- $B^h_{R_1}[-(R_2, pickXatY)] Control_{R_2}(\langle\langle (R_1, pickXatY) \rangle\rangle R_1 holds X)$: robot R_1 believes, with strength h, that before R_2 obtained container X from location Y, R_2 had control over R_1 obtaining container X by picking it up from location Y; in other words, R_2 could have done something else.

Combining these assumptions with our definitions above one can conclude that $Anger^l_{R_1,R_2}(pickXatY, R_1 holds X, pickXatY)$, where $l = merge(h,k)$. That is, robot R_1 is angry with intensity l at robot R_2 for picking up container X from location Y, and thus preventing R_1 to pick it up instead in order to hold it (presumably with the intention of transporting it somewhere else).

4.2 Social Anger

Social settings are often governed by specific social rules or norms, which causes agents to become related to each other. For example, in a social setting governed by the norm to respect the autonomy of each other, one agent can get angry at a second one, not because of the negative consequence of the action of the second agent for the first agent, but because the second agent has violated the norm by restricting the autonomy of a third agent. Similarly, in an organisational setting, a manager agent can get angry at one agent because the agent ignores an organisational rule with respect to a third agent. In our robot example, the manager agent gets angry at R_2 because R_2 has frustrated the goals of R_1.

Proceeding to the social anger, we reassert that it is a flavor of other-condemning anger with its content related to the harm done to other agents. Although there are different types of harm distinguished in the literature [12,20], what they all have in common is the violation of personal preferences by others. We represent now the emotion of social anger, together with the concept of harm, in the language of DMAL-GA.

$$Harm^k_{i,j}(a,\varphi) \stackrel{def}{=} AchG^k_j \varphi \wedge Account_i(a, \neg Pos_j(\varphi))$$

The formula $Harm^k_{i,j}(a,\varphi)$ should be read as "agent i harmed with strength k agent j by doing a and preventing him from achieving φ". By definition, this is the case *if and only if* j has an achievement goal φ and i is accountable for j not having the practical possibility to achieve its goal φ. In our robot example, R_2 harmed R_1 preventing R_1 from achieving its goal of obtaining the container: $Harm^k_{R_2,R_2}(pickXatY, R_1 holdsX)$. Social anger can now be defined as follows:

$$SAnger^l_{i,j,k}(a,\varphi,\psi) \stackrel{def}{=} \bigvee_{l=merge(m,n)} (\bigvee_{b\in Act} Anger^m_{i,j}(a,\varphi,b) \wedge B_i(Harm^n_{j,k}(a,\psi) \wedge (\varphi \rightarrow \psi)))$$

The formula $SAnger^l_{i,j,k}(a,\varphi,\psi)$ should be read as "agent i is socially angry with intensity l at agent j for harming agent k preventing k from achieving ψ by doing a and preventing i from following his social concern φ". By definition, $SAnger^l_{i,j,k}(a,\varphi,\psi)$ is true *if and only if* (1) $Anger^m_{i,j}(a,\varphi,b)$ for some $b \in Act$, i.e., agent i is angry at agent j for doing a and thereby preventing him from achieving φ by some action b, (2) agent i believes $Harm^n_{j,k}(a,\psi)$, i.e., agent i believes agent j has harmed agent k by preventing the achievement of k's goal ψ, and (3) agent i believes ψ holds if φ holds.

To illustrate, let us consider again our robot example. For social anger, agent k from the definition above translates to robot R_1, j translates to robot R_2, and i to the manager agent M, who is socially angry. Furthermore, ψ is R_1's wish to obtain the container, φ is the wish of the manager that the autonomy of other agents should be respected, and a the act of picking up the container.

4.3 Anger Related Validities

Our formalization of anger and other concepts respects the following intuitive validities.

- After the occurrence of an event an agent is accountable for a state of affairs iff the state of affair does currently not hold, the state of affair is the case after the event, and the event creates the history.
  $\models [(i,a)]Account_i(a,\phi) \leftrightarrow \neg\phi \wedge [(i,a)]\phi \wedge [(i,a)]Past(nil,(i,a))$
- Blame requires choices in the direct past.
  $\models Blame^k_{i,j}(a,\phi) \rightarrow B^k_i([-(j,a)] \bigvee_{e\in Evt_j}[e]\neg\phi)$
- No blame for unavoidable.
  $\models B^k_i(\phi \wedge \bigwedge_{e\in Evt_j}[e]\neg\phi) \rightarrow [(j,b)]\neg Blame^k_{i,j}(b,\phi)$ for $(j,b) \in Evt_j$
- No blame for trivialities and impossibilities.
  $\models \neg Blame^k_{i,j}(a,\top)$
  $\models \neg Blame^k_{i,j}(a,\bot)$
- Decomposition of accountability.
  $\models Account_i(a,\neg\phi) \leftrightarrow \neg Account_i(a,\phi)$
  $\models Account_i(a,\phi\vee\psi) \leftrightarrow Account_i(a,\phi) \vee Account_i(a,\psi)$
  $\models Account_i(a,\phi\wedge\psi) \rightarrow Account_i(a,\phi) \vee Account_i(a,\psi)$
  $\models Account_i(a,\phi) \wedge Account_i(a,\psi) \rightarrow Account_i(a,\phi\wedge\psi)$

– No anger at those who are not accountable for your disability or desired outcome of your choice.
$\models \quad (\neg B_i^k Account_j(a, \neg Fut(nil, (i,b)) \quad \wedge \neg B_i^k Account_j(a, \neg[(i,b)]\phi)) \quad \rightarrow \neg Anger_{i,j}^l(a, \phi, b)$
for $l = merge(h,k)$ and $h \in Num^+$
– Social Anger with respect to oneself implies Anger, but not vice versa.
$\models SAnger_{i,j,i}^l(a, \varphi, \varphi) \rightarrow Anger_{i,j}^l(a, \varphi, b)$ for some $(i,b) \in Evt_i$
$\not\models Anger_{i,j}^l(a, \varphi, b) \rightarrow SAnger_{i,j,i}^l(a, \varphi, \varphi)$ for any $(i,b) \in Evt_i$

4.4 Coping with (Social) Anger

Most psychologists agree that the innate coping strategy in anger is *aggression* towards the blameworthy agent [2,3], including *attack* and *threat* with the goal being the removal of the obstruction that caused anger. 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 example, the participant's decision to report the irregularity to an environment administrator is based on the evaluation of his inability to ensure robot R_2 makes the container accessible to robot R_1: an estimate of his coping potential.

Following [5], coping is specified in terms of a function $Trg : Agt \times CStr \rightarrow \mathcal{L}$ that maps agents Agt and strategies $CStr$ to formulae from $\mathcal{L}$: for every agent i and coping strategy β, $Trg(i, \beta)$ denotes the conditions for i that triggers the strategy. We consider coping strategies $CStr$ for social anger as *intention-affecting* strategies a^+ (adopting intention a) and a^- (removing intention a). As social anger is elicited when an agent is harmed, we specify coping with social anger as adopting the intention a for which it is known to lead to $Harm_{j,k}(a, \psi)$ being false, i.e.,

$$Tr(i, b^+) = SAnger_{i,j,k}^l(a, \varphi, \psi) \wedge K_i[(i,b)]\neg Harm_{j,k}^n(a, \psi)$$

where $b \in Act$ and all the other variables as used for social anger. An immediate observation is the following:

Proposition 5. *Let $\mathfrak{M}$ be a model, $w \in W$, a, $b \in Act$, $i, j, k \in Agt$ and $\varphi \in \mathcal{L}$. If $\mathfrak{M}, w \models K_i[(i,b)]\neg Harm_{j,k}(a, \psi)$, then $\mathfrak{M}, w \models [(i,b)]\neg SAnger_{i,j,k}^l(a, \varphi, \psi)$ for $l \in EmoInt$.*

That is, successfully triggering the coping strategy b^+ for agent i, and executing the action b, removes the presence of social anger – a property necessary for successful coping [16]. In our example, this amounts to saying that in case of social anger one should expect attacking behavior (banning, warning) towards the violating robot R_2. This way the problem of harming robot R_1 will be mitigated by repairing the transportation task of R_1 or banning robot R_2 from operating in the transportation environment. It is important to note that the triggering condition is not the same as the selecting/executing the strategy. The selection/execution of a strategy is a separate issue which should take the intensity of the involved social anger emotion and its corresponding harm into account. This issue is not discussed as it is outside the scope of this paper.

5 Concluding Remarks

Although the focus of this paper is other-condemning anger, the presented logical framework is powerful enough to model various other-condemning social emotions such as disgust and contempt. We left out a formalization of other social emotions due to space limitation. The characteristic features of the presented framework are its multi-agent flavor and the inclusion of emotion intensity. Although the importance of emotion intensity has been stressed by appraisal theorist, most of the formal models in the literature have ignored at least one of them. For example, [1,18] ignores emotion intensity and [5] does not have multi-agent flavor. Our proposed model is inspired by [5], but we consider other-condemning and socially oriented anger, which requires extending the single agent framework proposed in [5] to a multi-agent framework with the converse of actions to reason about the state of the world before action execution. This feature is of crucial importance to some components of anger, e.g., responsibility and blame. Another influencing work on the topic has been [28]. Unlike our approach, [28] take emotion intensity as primitive, without explaining how it depends on belief and goal strengths. Furthermore, [28] does not provide any decidability results or axiomatization, whereas the current work does provide axiomatization and a decidability result. Finally, [9] proposes a formal 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.

We intend to extend the set of other-condemning emotions in future work and provide an analysis on the relation between various moral emotions. We also aim at extending the dynamic nature of our proposed logic by allowing more complex actions and extend the accountability not to actions that have been performed in previous state, but to some state in the past.

References

1. Adam, C., Herzig, A., Longin, D.: A logical formalization of the occ theory of emotions. Synthese **168**, 201–248 (2009)
2. Averill, J.R.: Studies on anger and aggression: Implications for theories of emotion. Am. Psychol. **38**(1), 1145–1160 (1983)
3. Averill, J.R.: Anger and Aggression: An Essay on Emotion. Springer, New York (1982)
4. Blackburn, S.: Ruling Passions. Clarendon Press, Oxford (1998)
5. Dastani, M., Lorini, E.: A logic of emotions: from appraisal to coping. In: Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems, vol. 2, pp. 1133–1140 (2012)
6. Elster, J.: Rationality, emotions, and social norms. Synthese **98**(1), 21–49 (1994)
7. Frijda, N.H.: The Emotions. Cambridge Univ. Pr., Cambridge (1986)
8. Gewirth, A.: Reason and Morality. University of Chicago Press, Chicago (1981)
9. Gratch, J., Marsella, S.: A domain-independent framework for modeling emotion. Cogn. Syst. Res. **5**(4), 269–306 (2004)
10. Haidt, J.: The moral emotions. Handbook Affect. Sci. **11**, 852–870 (2003)

11. Halpern, J., Moses, Y.: A guide to completeness and complexity for modal logics of knowledge and belief. Artif. Intell. **54**, 319–379 (1992)
12. Helwig, C.C., Zelazo, P.D., Wilson, M.: Children's judgments of psychological harm in normal and noncanonical situations. Child Dev. **72**(1), 66–81 (2001)
13. Izard, C.E.: Human Emotions. Plenum, New York (1977)
14. Laverny, N., Lang, J.: From knowledge-based programs to graded belief-based programs, part i: On-line reasoning*. Synthese **147**, 277–321 (2005)
15. Lazarus, R.S.: Emotion and Adaptation. Oxford University Press, USA (1991)
16. Lazarus, R.S., Folkman, S.: Stress, Appraisal, and Coping. Springer, New York (1984)
17. Lorini, E.: A dynamic logic of knowledge, graded beliefs and graded goals and its application to emotion modelling. In: van Ditmarsch, H., Lang, J., Ju, S. (eds.) LORI 2011. LNCS, vol. 6953, pp. 165–178. Springer, Heidelberg (2011). doi:10.1007/978-3-642-24130-7_12
18. Lorini, E., Schwarzentruber, F.: A logic for reasoning about counterfactual emotions. Artif. Intell. **175**, 814–847 (2010)
19. Meyer, J.-J.C., van der Hoek, W., van Linder, B.: A logical approach to the dynamics of commitments. Artif. Intell. **113**, 1–40 (1999)
20. Ohbuchi, K., Kameda, M., Agarie, N.: Apology as aggression control: its role in mediating appraisal of and response to harm. J. Pers. Soc. Psychol. **56**(2), 219 (1989)
21. Ortony, A., Clore, G., Collins, A.: The Cognitive Structure of Emotions. Camb. Uni. Pr., Cambridge (1990)
22. Parikh, R.: The completeness of propositional dynamic logic. In: Winkowski, J. (ed.) MFCS 1978. LNCS, vol. 64, pp. 403–415. Springer, Heidelberg (1978). doi:10.1007/3-540-08921-7_88
23. Prinz, J.: The Emotional Construction of Morals. Oxford University Press, Oxford (2007)
24. Reiter, R.: Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems. MIT Press, Cambridge (2001)
25. Rozin, P., Lowery, L., Imada, S., Haidt, J., et al.: The cad triad hypothesis: a mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). J. Personal. Soc. Psychol. **76**, 574–586 (1999)
26. Scherer, K.R.: Appraisal considered as a process of multilevel sequential checking: a component process approach. Apprais. Process. Emot. Theory Methods Res. **92**, 120 (2001)
27. Spohn, W.: Ordinal conditional functions: a dynamic theory of epistemic states. Causation Dec. Belief Change Stat. **2**, 105–134 (1988)
28. Steunebrink, B.R., Dastani, M., Meyer, J.-J.C.: A formal model of emotion-based action tendency for intelligent agents. In: Lopes, L.S., Lau, N., Mariano, P., Rocha, L.M. (eds.) EPIA 2009. LNCS, vol. 5816, pp. 174–186. Springer, Heidelberg (2009). doi:10.1007/978-3-642-04686-5_15