A Formal Model of Emotion Triggers

An Approach for BDI Agents with Achievement Goals

Bas R. Steunebrink (bas@idsia.ch)
Dalle Molle Institute for Artificial Intelligence (IDSIA),
Lugano, Switzerland

Mehdi Dastani (mehdi@cs.uu.nl)
John-Jules Ch. Meyer (jj@cs.uu.nl)
Intelligent Systems Group,
Institute of Information and Computing Sciences,
Utrecht University,
Utrecht, The Netherlands

Abstract. This paper formalizes part of a well-known psychological model of emotions. In particular, the logical structure underlying the conditions that trigger emotions are studied and then hierarchically organized. The insights gained therefrom are used to guide a formalization of emotion triggers, which proceeds in three stages. The first stage captures the conditions that trigger emotions in a semiformal way, i.e., without committing to an underlying formalism and semantics. The second stage captures the main psychological notions used in the emotion model in dynamic doxastic logic. The third stage introduces a BDI-based framework (belief–desire–intention) with achievement goals, which is used to firmly ground the preceding stages. The result is a formalization of emotion triggers for BDI agents with achievement goals. The idea of proceeding in these stages is to provide different levels of commitment to formalisms, so that it remains relatively easy to extend or replace the used formalisms without having to start from scratch. Finally, we show that the formalization renders properties of emotions that are in line with the psychological model on which it is based.

Keywords: Logic of Emotions, Intelligent Agents, Cognitive Modeling

1. Introduction

Recently, there has been a lot of interest in bringing emotions to Artificial Intelligence, in particular to model rational agents [28, 18, 31, 2, 13, 21, 22, 6, 33, 34, 35, 1, 32]. There are (at least) three important reasons for this. First, an obvious application of emotions is to make artificial agents and robots more believable to human users. In particular, their behaviors are expected to appear increasingly convincing, social, and intuitive if they seem to have an emotional state matching that of a human in the same situation [28, 2, 13, 21]. Second, from a more theoretical perspective, it is investigated what the role of emotions is in models of human decision-making and how they may be employed to make these models more accurate and effective [8, 18, 4]. Third, there exists psychological [10, 20, 7, 24] and neurological [5] evidence that emotions are not only relevant but even necessary for rational behavior. Particularly, it has been shown that persons who do not experience emotions (e.g., due to specific brain damage) have trouble distinguishing between important and irrelevant details, consistently make bad decisions, and do not display adequate social behavior necessary to function normally in society. A related, more philosophical argument posits that emotions are an inevitable consequence of mechanisms that allow for intelligent and rational behavior in complex environments with limited resources [31, 12, 8, 17].

There is little consensus among psychologists as to what exactly constitutes an emotion and how it differs from related affective processes such as moods and impulses. However, this does not mean that making broad classifications is impossible or useless. According to a classification by Gross [14], *emotions* typically have specific objects and give rise to action tendencies relevant to these objects. Moreover, emotions can be both positive and negative. Emotions are often distinguished from *moods*, which are more diffuse and last longer than emotions. Other affective processes include *stress*, which arises in taxing circumstances and produces only negative responses; and *impulses*, which are related to hunger, sex, and pain and give rise to responses with limited flexibility. Of these four types of affective processes, we will focus on *emotions* in this paper.

With respect to emotions, usually three phases are distinguished. First, the perceived situation is *appraised* by an individual based on what he or she thinks is relevant and important. For example, Alice, who likes receiving presents, is given a necklace by Bob. Alice then judges receiving the necklace as desirable and Bob's action as praiseworthy. Consequently, the appraisal of this action and its outcome causes gratitude towards Bob to be triggered for Alice. Note that different types of emotions may be triggered simultaneously by the same situation, some of which may even be seen as conflicting. For example, Alice may at the same time be disappointed because it was not the necklace she had hoped to receive. Emotion theories dealing with appraisal are for example [10, 27, 20, 24, 30]. Second, the appraisal of some situation can cause the triggered emotions, if exceeding some threshold, to create a conscious awareness of emotional feelings, leading to the *experience* of having emotions. For example, Alice's gratitude towards Bob will have a certain intensity and will probably decrease over a certain amount of time. All this may depend on, e.g., the degree of desirability of receiving a necklace and Alice's previous attitude towards Bob. Emotion theories dealing with these quantitative aspects of emotions are for example [27, 10, 7]. Third, emotional feelings need to be *regulated*. For example, Alice may want to organize her behavior such that positive emotions are triggered as often as possible and negative emotions are avoided or drowned by positive ones. She could do this by being nice to Bob so that he will give her more presents, or avoiding him altogether so that she will never again be confronted with his bad taste in jewelry. In fact, some emotion theories posit that the main purpose of emotions is to function as a heuristical mechanism for selecting behaviors [5, 24, 20]. Emotion theories dealing with coping and behavioral consequences of emotions are for example [10, 7, 19, 24].

With respect to models of rational agency, important topics include tracking of goal achievements (i.e. rate of success), revision of plans, and where to focus attention. Emotion theories can provide solutions to these issues by treating emotions as heuristics in the deliberation and decision making of agents. Given the aim to integrate emotions in the models of artificial rational agents on the one hand, and the existence of psychological theories of emotions on the other hand, a question that thus arises is how we can adopt, formalize, and use psychological theories of emotions in models of rational agency. Following psychological theories, at least three main issues (appraisal, experience, regulation) need to be addressed in this effort, but of course this cannot be done properly and comprehensively in one paper. The central question addressed in this paper is how to model and integrate the appraisal part of emotions in agent models. In particular, a formal framework must be built that is suitable for modeling agents and for investigating how appraisal can be integrated in this formal framework. In order to integrate emotions into an agent model, we must map psychological concepts onto agent concepts so that they can be appropriately formalized.

In this paper, we will present a formalization of the eliciting conditions of emotions as described in the psychological model of Ortony, Clore & Collins [27].[1] We have chosen the OCC model because it provides a clear classification of a broad range of emotion types, it lists concise descriptions of the conditions that elicit emotions, and for this it uses concepts that are well studied and relatively straightforward to formalize. The presented formalization constitutes a formal model for the appraisal part of the OCC emotion theory. This formal model is an extension of an agent specification framework (i.e., KARO [23, 22]) that specifies agents in terms of cognitive concepts that are also used in the OCC model. For each emotion type from the OCC model we translate the eliciting conditions, which define the appraisal process corresponding to that emotion type, into concepts from the formal agent framework. The contribution of this paper is to provide a formal agent model in which the appraisal part of the complete set of emotion types from the OCC model is integrated. We show that the formalization is adequate by providing a logical analysis and proving intuitive properties of the model.

This paper is organized as follows. In section 2 we give an overview of the psychological OCC model of emotions. We will particularly study its logical structure in great detail, because it is on this structure that our formalization will be based. The formalization will proceed in three stages, spread over sections 3, 4, and 5, respectively. The first stage will be a semiformal specification of the logical structure of eliciting conditions. The second stage will

[1] Henceforth to be referred to as “the OCC model” or simply “OCC.”

capture the main notions used in the OCC model in dynamic doxastic logic. The third stage will formalize the main appraisal notions used in the OCC model in the KARO framework, which is an extension of dynamic doxastic logic, and thus firmly grounds the preceding stages. In section 6 we will discuss and compare related work, and section 7 will conclude this paper.

2. The OCC Model

In their book "The Cognitive Structure of Emotions" [27], Ortony, Clore & Collins have proposed a very interesting model of emotions that provides specifications of the eliciting conditions of emotions and the variables that affect their intensities. This psychological model of emotions is popular among computer scientists that are trying to build systems that reason about emotions or incorporate emotions in artificial characters. This popularity is due to the model's clear and convincing structure.

2.1. Overview of the OCC Model

The OCC model classifies 22 types of emotions. This is done by considering on which kinds of aspects of a situation one can focus his or her attention. OCC consider a human can either focus on consequences of events,[2] actions of agents, or aspects of objects. If one focuses on a consequence of an event, one can appraise this consequence as *desirable* or *undesirable* (or both, or neither) with respect to one's *goals*. For example, joy about winning a lottery is an event-based emotion, because the satisfaction of the goal to become rich is a desirable consequence of the event of winning the lottery. If one focuses on an action of an agent, one can appraise this action as *praiseworthy* or *blameworthy* (or both, or neither) with respect to one's *standards*. For example, pride about saving a child from drowning is an action-based emotion, because it is praiseworthy to perform an action which satisfies the standard that one should save a person's life whenever (reasonably) possible. If one focuses on an aspect of an object, one can appraise this aspect as *appealing* or *unappealing* (or both, or neither) with respect to one's *attitudes*. For example, love for an old car is an object-based emotion, because the car may have appealing aspects according to one's attitudes.

Within these three main categories of emotion types, the OCC model makes further differentiations based on, e.g., whether prospects are relevant (as in hope and fear), whether events apply to others (as in pity and gloating), or whether an action was performed by the self or someone else (to

[2] On page 18 [27], OCC say that "events are simply people's construals about things that happen." From a computational perspective, however, we would say that this is far from *simple*!

distinguish e.g. pride from admiration). Additionally, some event-based and action-based emotion types are combined to form a group of emotions concerning consequences of events *caused* by actions of agents. For example, anger can arise when one focuses on both the blameworthy action of another agent and an undesirable event which has been (presumed to be) caused by it. It should be emphasized that in the OCC model, emotions are never used to describe the entire cognitive state of an agent (as in "Alice is happy"); rather, emotions are always relative to individual events, actions, and objects. So Alice can be joyous about receiving her new furniture and at the same time be distressed about the height of the accompanying bill.

2.2. The Logical Structure of the OCC Model

Although many ad hoc or simplified implementations of the OCC model have been made, there have been fewer attempts at formalizing the complete, logical structure of the proposed emotion model (e.g., [13, 33, 1, 32]). Here we will attempt to do so, using a formal logic containing constructs to reason about agents, their beliefs and actions, objects, and events. We will be formalizing the eliciting conditions of emotions in this logic, trying to stay as close as possible to the OCC model.

On page 19 of their book [27], OCC present a diagram which structures their emotion types based on *focus of attention*. This diagram is often reproduced when an overview of the OCC model is to be given. In this section we give an overview of the OCC model, but we will illustrate the OCC model with a slightly different diagram (see figure 2.1). The purpose of this paper is to provide a logical account of emotion triggers. However, OCC's diagram as it appeared in [27], which is based on focus of attention, is not very well suited to guide our formalization, because it is not compositional (see [32] for an detailed discussion on this issue). Therefore, we have created the new diagram illustrating the structure of the emotion types based on their *eliciting conditions*. It should be noted that in personal communication, Ortony and Clore have confirmed figure 2.1 to be an accurate compositional illustration of the logical structure underlying the eliciting conditions of their emotion types [26]. The following paragraphs serve to explain figure 2.1, which, importantly, serves as a guide for the formalization that will be presented in the next sections.

Figure 2.1 can be seen as an *inheritance* structure. This means that the depicted emotion types are *specializations* of those above them and *generalizations* of those below them. This inheritance-based perspective results in a compositional formulation of the eliciting conditions. At the most general level, all emotions are valenced reactions (to something). Although valenced reactions can have different magnitudes, each one is at least either positive or negative. Therefore 'positive' and 'negative' have been placed at the top of the

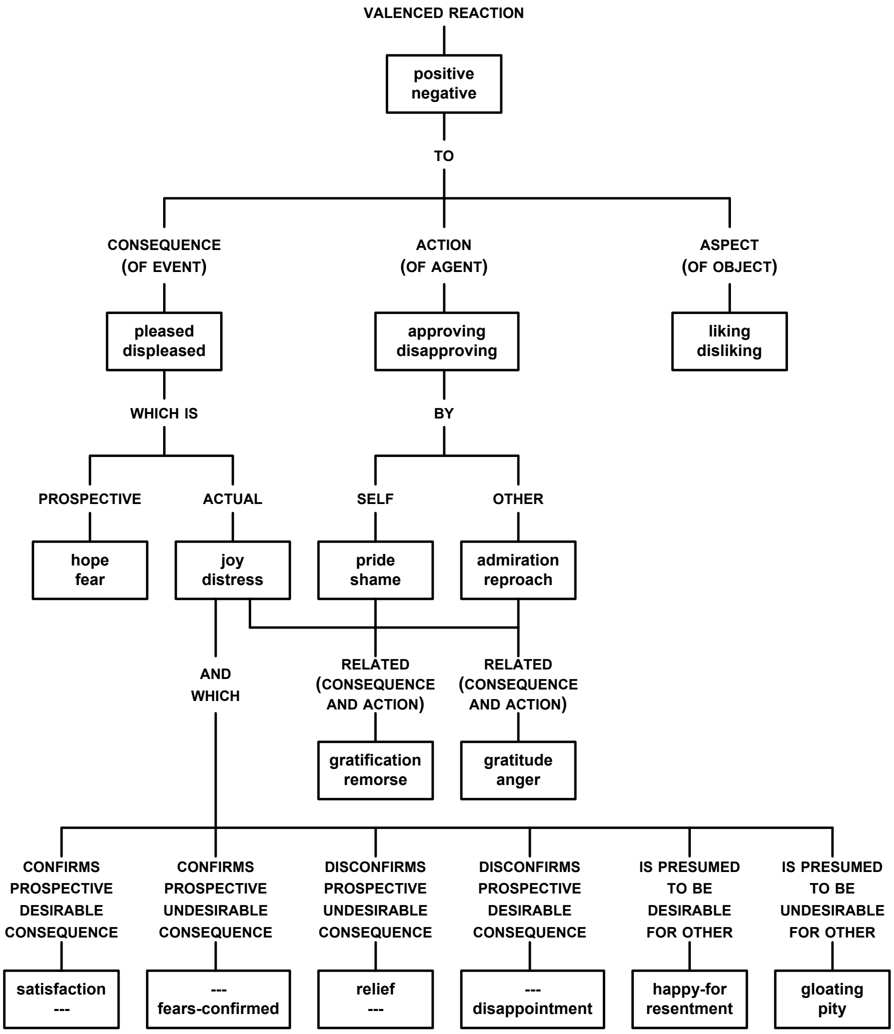


Figure 2.1. An inheritance-based view of the eliciting conditions of emotions of the OCC model.

hierarchy. At the next level, the OCC model specifies that valenced reactions can be directed at either consequences of events,[3] actions of agents, or aspects

[3] Although the phrase "(un)desirable event" is used many times by OCC, events are actually always appraised with respect to their consequences. For example, an earthquake in itself does not have a valence; only the consequences of this event (e.g., valuable lessons for seismologists, property damage, loss of life) are appraised as being desirable or undesirable. Because desirability only applies to consequences of events, every instance of the phrase "(un)desirable event" should actually be read as a shorthand for "(un)desirable consequence of an event" [26].

of objects. ‘Pleased’ and ‘displeased’ have been chosen by OCC to function as labels for the most general type of valenced reactions to consequences of events, because they are very neutral sounding words with respect to intensity of experience, focus of attention, motivational and behavioral effects, etc. For the same reasons, ‘approving’ and ‘disapproving’ are used as labels for the most general type of valenced reactions to actions of agents, and ‘liking’ and ‘disliking’ are used as labels for the most general type of valenced reactions to aspects of objects.

With respect to valenced reactions to consequences of events, a distinction is made based on whether the consequence in question is prospective[4] or not. For example, learning that tomorrow it will rain is an event, but it has an undesirable consequence (e.g., the undermining of the goal to have a dry picnic) that is prospective and not actual. But this event can also have a consequence which is actual (e.g., the achievement of the goal to know the weather forecast). This differentiation on prospects then results in distinguishing between the ‘hope’ and ‘fear’ types on the one hand (e.g., Alice fears tomorrow her picnic will get wet) and the ‘joy’ and ‘distress’ types on the other hand (e.g., Alice is joyous about having learned the weather forecast).

With respect to valenced reactions to actions of agents, a distinction is made on whether the action in question has been performed by the self or by someone else. The distinction between “one’s own action” and “someone else’s action” is, however, not as simple as it may seem. A mother may be proud of the achievements of her son, even though the actions of her son are, strictly speaking, not her own. To resolve this, the OCC model uses the concept of a *cognitive unit*: the mother can consider herself and her son as part of a single cognitive unit and then, when appraising her son’s actions as praiseworthy, feel proud of the actions performed by (an agent in) the cognitive unit. This differentiation on cognitive unit then results in distinguishing between the ‘pride’ and ‘shame’ types on the one hand, and the ‘admiration’ and ‘reproach’ types on the other hand.

At this point the reader may expect there to be a branch below liking and disliking, after seeing branches being added below pleased/displeased and approving/disapproving. Indeed, in the original diagram of the OCC model ([27], page 19), a single branch appears below liking/disliking with the emotion types ‘love’ and ‘hate’. The idea of OCC was that ‘love’ and ‘hate’ are examples of emotions of the type ‘liking’ and ‘disliking’, respectively

[4] The term “prospect” (used in, e.g., hope and fear) is intentionally ambiguous: it is used to refer to both *future* events and *uncertain* (past or current) events. For example, hoping that tomorrow will be a sunny day is future-directed, whereas hoping that a mailed package has safely reached its intended recipient is uncertainty-directed. Many formalizations appear to use OCC’s notion of prospect in only one of these senses. For example, Adam [1] and Gratch & Marsella [13] only used uncertain prospects when formalizing hope and fear, whereas Steunebrink, Dastani & Meyer [33] only used future prospects.

[26]. However, this means that the distinction between love/hate and liking/disliking does not constitute a differentiation in terms of eliciting conditions, but merely that ‘love’ is a *token* for the type of emotions labeled ‘liking’ and that ‘hate’ is a *token* for the type of emotions labeled ‘disliking’. So in our inheritance-based perspective, no branch has to be added below liking/disliking, because ‘love’ and ‘hate’ are not specializations of ‘liking’ and ‘disliking’ with respect to eliciting conditions.

In addition to valenced reaction to either consequences of events or actions of agents, the OCC model also considers several types of emotions arising from observing relations between the two. Specifically, these emotion types correspond to valenced reactions to consequences of events *caused* by actions of agents. The eliciting conditions of these so-called compound emotion types are conjunctions of their constituent emotion types: ‘joy’ plus ‘pride’ is ‘gratification’, ‘joy’ plus ‘admiration’ is ‘gratitude’, ‘distress’ plus ‘shame’ is ‘remorse’, and ‘distress’ plus ‘reproach’ is ‘anger’. Note, however, that this “plus” contains an implicit assertion about their (presumably causal) relation. For example, ‘anger’ is specified as “(disapproving of) someone else’s blameworthy action and (being displeased about) the *related* undesirable event”.

At the bottom of the hierarchy (of figure 2.1) are several emotions that are special types of ‘joy’ and ‘distress’. Consider first ‘satisfaction’, ‘fears-confirmed’, ‘relief’, and ‘disappointment’. Together with ‘hope’ and ‘fear’, these six emotion types are called prospect-based in the OCC model. However, the eliciting conditions of the former four emotion types are not extensions of ‘hope’ and ‘fear’, but of ‘joy’ and ‘distress’. Specifically, ‘satisfaction’, ‘fears-confirmed’, ‘relief’, and ‘disappointment’ are emotions in response to *actual* consequences of events, namely consequences signaling the confirmation or disconfirmation of a *previously prospective* consequence. The relation between, e.g., hope and disappointment is thus more of a temporal kind. For example, first Bob hopes Alice will show up for their date, but when she does not, his hope turns into disappointment. Thus ‘satisfaction’, ‘fears-confirmed’, ‘relief’, and ‘disappointment’ are not special kinds of ‘hope’ or ‘fear’, but more like continuations of ‘hope’ or ‘fear’, counting from the point when an event has been perceived that signals the confirmation or disconfirmation of the thing hoped for or feared.[5]

Next consider ‘happy-for’, ‘resentment’, ‘gloating’, and ‘pity’, called the fortunes-of-others emotion types. These are valenced reactions arising from

[5] The dashes accompanying ‘satisfaction’, ‘fears-confirmed’, ‘relief’, and ‘disappointment’ in figure 2.1 are intentional and do not indicate a problem. For example, ‘satisfaction’ is a label for a positively valenced reaction to the confirmation of a prospective desirable consequence, but the dashes below ‘satisfaction’ are a placeholder for a *negatively* valenced reaction to the confirmation of a prospective desirable consequence. Of course, in practice such a negative reaction to something positive never occurs, and thus does not have to be labeled.

presuming that events have consequences for others. However, for there to be, e.g., a ‘happy-for’ emotion, the consequence that is desirable for the other must also be desirable for oneself to some degree (probably because it satisfies an interest goal to wish other people well, as suggested on page 94 [27]). But if a consequence of an event is appraised as being desirable for oneself, the conditions for ‘joy’ are satisfied. So logically speaking, the eliciting conditions of ‘happy-for’ and ‘gloating’ entail those of ‘joy’, and the eliciting conditions of ‘resentment’ and ‘pity’ entail those of ‘distress’. Therefore, in an inheritance-based hierarchy of eliciting conditions, the fortunes-of-others emotion types must be placed below ‘joy’ and ‘distress’, because the latter two emotion types are generalizations of them.

The specifications of eliciting conditions resulting from reading figure 2.1 are summed up in table 2.1. It is crucial to take note of table 2.1 because it will serve as our guide in the formalization of emotion triggers.

3. Capturing the Logical Structure of the OCC Model

In this section, we will make a start with the formalization of the eliciting conditions of emotions according to the OCC model. It is important to note that a distinction is made between what triggers an emotion and how an emotion is experienced. The intensity at which an emotion is felt is influenced by many factors. Emotional experience is probably multidimensional, but it is also assumed that an estimate can be made of the “overall felt intensity” of any emotion [11]. For example, a questionnaire about emotional feelings may include a question like “indicate how angry you were when hearing about the political murder on a scale from 1 to 10” and such questions are usually not difficult to answer.

In the rest of this paper, it is assumed the following relation between emotion triggering and emotion experience exists:

emotion_type is experienced if and only if
(1) *emotion_type* has been triggered sometime in the past and
(2) overall felt intensity of *emotion_type* is positive

With “positive” we mean having a value strictly greater than zero. Emotional intensity does not take on negative values. An emotion type for which the triggering conditions hold is not necessarily experienced, because its intensity may be too low (i.e., zero). And an emotion type which is being experienced does not have to have its triggering conditions to hold, because it may have been triggered some time in the past. In this section and the next two, we will focus on the triggering conditions of the emotion types of the OCC model. The treatment of emotional experience, as expressed in terms of triggering and intensity as above, can be found in [32].

Table 2.1. These emotion type specifications correspond directly to figure 2.1.

positive and *negative*	are valenced reactions (to “something”)
pleased	is being *positive* about a consequence (of an event)
displeased	is being *negative* about a consequence (of an event)
hope	is being *pleased* about a prospective consequence (of an event)
fear	is being *displeased* about a prospective consequence (of an event)
joy	is being *pleased* about an actual consequence (of an event)
distress	is being *displeased* about an actual consequence (of an event)
satisfaction	is *joy* about the confirmation of a prospective desirable consequence
fears-confirmed	is *distress* about the confirmation of a prospective undesirable consequence
relief	is *joy* about the disconfirmation of a prospective undesirable consequence
disappointment	is *distress* about the disconfirmation of a prospective desirable consequence
happy-for	is *joy* about a consequence (of an event) presumed to be desirable for someone else
resentment	is *distress* about a consequence (of an event) presumed to be desirable for someone else
gloating	is *joy* about a consequence (of an event) presumed to be undesirable for someone else
pity	is *distress* about a consequence (of an event) presumed to be undesirable for someone else
approving	is being *positive* about an action (of an agent)
disapproving	is being *negative* about an action (of an agent)
pride	is *approving* of one’s own action
shame	is *disapproving* of one’s own action
admiration	is *approving* of someone else’s action
reproach	is *disapproving* of someone else’s action
gratification	is *pride* about an action and *joy* about a related consequence
remorse	is *shame* about an action and *distress* about a related consequence
gratitude	is *admiration* about an action and *joy* about a related consequence
anger	is *reproach* about an action and *distress* about a related consequence
liking	is being *positive* about an aspect (of an object)
disliking	is being *negative* about an aspect (of an object)

Note that in this section, all formulas are only semiformal, since no semantics are yet given, only intuitive readings. We use logical connectives with their usual interpretation and some operators with suggestive names. The idea is that the presented formulas are formal enough to adequately capture the logical structure of the psychological model, while remaining free from a commitment to an underlying formalism. In later sections we will commit to one formalism and provide a firm grounding, but it is our intention that it remains possible to plug in another formalism if desired. This way one can provide a different interpretation of the operators, without having to start from scratch regarding the overall logical structure of the model.

The structure of our formalization of emotion triggers was illustrated in figure 2.1; we will be following this figure (from top to bottom) and its textual version table 2.1 in the next subsections.

3.1. General Emotion Types

At the most abstract level, the OCC model considers an emotion as a valenced reaction, which can either be positive or negative. So 'positive' and 'negative' are regarded by OCC as the most general, undifferentiated emotion types. In order to know which one of 'positive' or 'negative' (or both, or neither) is triggered at some point, something must be perceived and valued, which is called appraisal. We should note that the term 'appraisal' can be used to mean perception as well as valuation. In order to avoid confusion, we explicitly mention perception (of an event, action, or object) as a precondition for appraisal, and use the term appraisal strictly for valuation.

We can thus (trivially) specify the triggering conditions of 'positive' and 'negative' as the perception of something good and bad, respectively:

$$\mathbf{Positive}_i^{\mathbf{T}}(X) \overset{\text{def}}{=} \mathbf{Perceive}_i(X) \land \mathbf{Good}_i(X) \tag{3.1}$$

$$\mathbf{Negative}_i^{\mathbf{T}}(X) \overset{\text{def}}{=} \mathbf{Perceive}_i(X) \land \mathbf{Bad}_i(X) \tag{3.2}$$

Emotions are always relative to something, and here X stands for that "something." $\mathbf{Positive}_i^{\mathbf{T}}(X)$ is read as "a positively valenced reaction to X is triggered for agent i." The superscript "$\mathbf{T}$" (for trigger) indicates that we are talking about eliciting conditions, in order to avoid confusion with actual experience. It is crucial to note that $\mathbf{Positive}_i^{\mathbf{T}}(X)$ is *not* the same as "agent i is positive about X." The feeling of being positive about X may manifest itself gradually over time, if at all, and may not coincide with the satisfaction of its triggering conditions, which is what $\mathbf{Positive}_i^{\mathbf{T}}(X)$ expresses.

$\mathbf{Perceive}_i(X)$ is read as "agent i perceives X." $\mathbf{Good}_i(X)$ is read as "agent i appraises X as good," and similarly for $\mathbf{Bad}_i(X)$.

With 'positive' and 'negative' at the top of the hierarchy, the first differentiation is with respect to the object of the emotion. As previously described, the OCC model considers three types: *consequences of events*, *actions of agents*, and *aspects of objects*. We can thus define 'perceive' as a disjunction of perceiving either of these three:

$$\begin{aligned}\mathbf{Perceive}_i(X) \overset{\text{def}}{=} {} & \mathbf{PerceiveConseq}_i(X) \lor \\ & \mathbf{PerceiveAction}_i(X) \lor \\ & \mathbf{PerceiveObject}_i(X)\end{aligned} \tag{3.3}$$

These three perception constructs will be clarified in the next three subsections, respectively.

If a consequence of an event is appraised as being good or bad, it is said to be ‘desirable’ or ‘undesirable’, respectively. If an action of an agent is appraised as being good or bad, it is said to be ‘praiseworthy’ or ‘blameworthy’, respectively. If an aspect of an object is appraised as being good or bad, it is said to be ‘appealing’ or ‘unappealing’, respectively. ‘Good’ and ‘bad’ can thus be defined in terms of these six notions as follows:

$$\mathbf{Good}_i(X) \overset{\text{def}}{=} \mathbf{Des}_i(X) \vee \mathbf{Praisew}_i(X) \vee \mathbf{Appeal}_i(X) \tag{3.4}$$

$$\mathbf{Bad}_i(X) \overset{\text{def}}{=} \mathbf{Undes}_i(X) \vee \mathbf{Blamew}_i(X) \vee \mathbf{Unappeal}_i(X) \tag{3.5}$$

A note about the types of arguments is in order here. As noted previously, desirability is only applicable to consequences of events, praiseworthiness is only applicable to actions of agents, etc. However, the current construction says that, e.g., $\mathbf{PerceiveConseq}_i(X) \wedge \mathbf{Praisew}_i(X) \rightarrow \mathbf{Positive}_i^{\mathbf{T}}(X)$ is valid. Of course, in this example, either $\mathbf{PerceiveConseq}_i(X)$ or $\mathbf{Praisew}_i(X)$ must be applied to the wrong type of X. Therefore it is assumed that all these constructs[6] evaluate to false if they are applied to an argument of the wrong type. This way $\mathbf{PerceiveConseq}_i(X) \wedge \mathbf{Praisew}_i(X)$ is always false and the implication is still true.

It should also be noted that none of the desirable–undesirable, praiseworthy–blameworthy, and appealing–unappealing pairs are considered to be opposites, nor are they considered to be mutually exclusive. For example, a consequence which is not desirable is not necessarily undesirable; a lack of appeal does not make something unappealing; and the exact same action can be appraised as being both praiseworthy and blameworthy. Therefore, we really need six distinct appraisal constructs here. These three pairs of appraisal operators will be clarified in the next three subsections, respectively.

3.1.1. *General Event-based Emotion Types*

Here we consider the emotion types concerning consequences of events. At the top of this branch are placed the labels ‘pleased’ and ‘displeased’ (see figure 2.1). OCC consider *desirability* as the central variable measuring how positive an event is for an individual. An event that is valued negatively is called *undesirable*. As noted previously, undesirability is not the same as the absence of desirability, nor are desirability and undesirability assumed to exclude each other; they are seen as separate variables.[7] In the following,

[6] That is, $\mathbf{PerceiveConseq}$, $\mathbf{PerceiveAction}$, $\mathbf{PerceiveObject}$, $\mathbf{Des}$, $\mathbf{Undes}$, $\mathbf{Praisew}$, $\mathbf{Blamew}$, $\mathbf{Appeal}$, and $\mathbf{Unappeal}$.

[7] OCC use undesirability as a kind of negative desirability, i.e., a consequence is undesirable if its desirability is strictly less than zero. Here we prefer to keep desirability and undesirability as separate measures, each ranging over non-negative values. But it will be clear that the two approaches do not conflict.

we will treat desirability in a qualitative manner, i.e., something is either desirable or not. Of course, there can be degrees of desirability, so when we say that something is desirable, this may be read as "having strictly positive desirability," and when we say "not desirable," this may be read as "having zero desirability." Analogous readings apply to undesirability.

Below is then a (semiformal) logical description of the eliciting conditions of 'pleased' and 'displeased'. $\mathbf{Pleased}_i^{\mathbf{T}}(c)$ should be read as "pleased about consequence c of an event is triggered for agent i," and similarly for 'displeased'.

$$\mathbf{Pleased}_i^{\mathbf{T}}(c) \overset{\text{def}}{=} \mathbf{PerceiveConseq}_i(c) \land \mathbf{Des}_i(c) \tag{3.6}$$

$$\mathbf{Displeased}_i^{\mathbf{T}}(c) \overset{\text{def}}{=} \mathbf{PerceiveConseq}_i(c) \land \mathbf{Undes}_i(c) \tag{3.7}$$

These formulas express that the eliciting conditions of being 'pleased' and being 'displeased' have two components; namely, the perception of a consequence of an event and the appraisal of that consequence as being (un)desirable. $\mathbf{PerceiveConseq}_i(c)$ is read as "agent i perceives consequence c (of an event)" and $\mathbf{Des}_i(c)$ is read as "agent i appraises consequence c as desirable (with respect to its goals)," or, less precisely, "i desires c." It will be clear that 'pleased' and 'displeased' are *undifferentiated* event-based emotions, because nothing is assumed about what kind of consequence we are dealing with, nor anything about who or what caused the event, nor to whom (other than the appraising agent) the consequence applies.

3.1.2. *General Action-based Emotion Types*

At the top of the branch of emotion types concerning actions of agents are placed the labels 'approving' and 'disapproving', which are regarded by OCC as the most general action-based emotion types. The OCC model considers *praiseworthiness* and *blameworthiness* to be the central variables for valuating actions of agents. Analogously to 'pleased' and 'displeased' (see above), 'approving' and 'disapproving' can be specified as perceiving an action of an agent and appraising that action as praiseworthy or blameworthy (or both, if one has conflicting standards). $\mathbf{Approving}_i^{\mathbf{T}}(j{:}a)$ should be read as "approving of action a by agent j is triggered for agent i," and similarly for 'disapproving'.

$$\mathbf{Approving}_i^{\mathbf{T}}(j{:}a) \overset{\text{def}}{=} \mathbf{PerceiveAction}_i(j{:}a) \land \mathbf{Praisew}_i(j{:}a) \tag{3.8}$$

$$\mathbf{Disapproving}_i^{\mathbf{T}}(j{:}a) \overset{\text{def}}{=} \mathbf{PerceiveAction}_i(j{:}a) \land \mathbf{Blamew}_i(j{:}a) \tag{3.9}$$

$\mathbf{PerceiveAction}_i(j{:}a)$ is read as "agent i perceives agent j has performed action a." $\mathbf{Praisew}_i(j{:}a)$ is read as "agent i appraises action a by agent j as praiseworthy (with respect to its standards)," and similarly for $\mathbf{Blamew}_i(j{:}a)$ and 'blameworthy'.

3.1.3. *General Object-based Emotion Types*

At the top of the branch of emotion types concerning aspects of objects are placed the labels 'liking' and 'disliking', which are regarded by OCC as the most general object-based emotion types. The OCC model considers *appealingness* and *unappealingness* to be the central variables for valuating aspects of objects. Just as desirability only applies to *consequences* of events, the OCC model considers appealingness to apply to *aspects* of objects. In the rest of this paper, however, we will simplify slightly by not representing aspects explicitly. This is usually not problematic, because different aspects of an object can often be regarded as objects themselves. For example, different aspects of a car, (e.g., headlights, doors, wheels) are objects themselves that can be liked or disliked. The appraisal of aspects that are not objects (e.g., the car's color) is simply assumed to be handled implicitly by the constructs for appealingness and unappealingness when applied to the object in question.

Analogously to the event-based and action-based emotion types above, 'liking' and 'disliking' can be specified as perceiving an object and appraising that object as appealing or unappealing. $\mathbf{Liking}^{\mathbf{T}}_{i}(x)$ should be read as "liking of object x is triggered for agent i," and similarly for 'disliking'.

$$\mathbf{Liking}^{\mathbf{T}}_{i}(x) \overset{\text{def}}{=} \mathbf{PerceiveObject}_{i}(x) \wedge \mathbf{Appeal}_{i}(x) \tag{3.10}$$

$$\mathbf{Disliking}^{\mathbf{T}}_{i}(x) \overset{\text{def}}{=} \mathbf{PerceiveObject}_{i}(x) \wedge \mathbf{Unappeal}_{i}(x) \tag{3.11}$$

$\mathbf{PerceiveObject}_{i}(x)$ is read as "agent i perceives object x." $\mathbf{Appeal}_{i}(x)$ is read as "agent i appraises object x as appealing (with respect to its attitudes)," and similarly for $\mathbf{Unappeal}_{i}(x)$ and 'unappealing'.

3.2. Concrete Emotion Types

In this section we will provide semiformal descriptions of the eliciting conditions of the third layer (of figure 2.1) of emotion types.

3.2.1. *Event-based Emotion Types*

The first differentiation with respect to event-based emotion types is on whether the consequence in question is prospective or actual. First, we will treat the case of prospective consequences of events, leading to the emotion types labeled as 'hope' and 'fear'.

$$\mathbf{Hope}^{\mathbf{T}}_{i}(c) \overset{\text{def}}{=} \mathbf{Pleased}^{\mathbf{T}}_{i}(c) \wedge \mathbf{Prospective}_{i}(c) \tag{3.12}$$

$$\mathbf{Fear}^{\mathbf{T}}_{i}(c) \overset{\text{def}}{=} \mathbf{Displeased}^{\mathbf{T}}_{i}(c) \wedge \mathbf{Prospective}_{i}(c) \tag{3.13}$$

$\mathbf{Prospective}_{i}(c)$ is read as "agent i considers c to be a prospective consequence (of an event)." $\mathbf{Hope}^{\mathbf{T}}_{i}(c)$ is then read as "hope about consequence c (of an event) is triggered for agent i," and similarly for 'fear'.

Next, we will treat the case of actual consequences, leading to the emotion types labeled as 'joy' and 'distress'.

$$\mathbf{Joy}_i^{\mathbf{T}}(c) \overset{\text{def}}{=} \mathbf{Pleased}_i^{\mathbf{T}}(c) \wedge \mathbf{Actual}_i(c) \tag{3.14}$$

$$\mathbf{Distress}_i^{\mathbf{T}}(c) \overset{\text{def}}{=} \mathbf{Displeased}_i^{\mathbf{T}}(c) \wedge \mathbf{Actual}_i(c) \tag{3.15}$$

$\mathbf{Actual}_i(c)$ is read as "agent i considers c to be an actual consequence (of an event)." $\mathbf{Joy}_i^{\mathbf{T}}(c)$ is then read as "joy about consequence c (of an event) is triggered for agent i," and similarly for 'distress'.

We emphasize again that 'joy' and 'distress' are considered as nothing more than convenient labels for these emotion types. Other labels are perfectly possible as well; for example, emotions of the type labeled as 'joy' include contentment, delight, being glad, happiness, cheerfulness, being ecstatic, and so on and so forth. Similarly, emotions of the type labeled as 'distress' include sadness, upset, being distraught, shock, etc. If one further differentiates the type of event towards which one is distressed, even more specific labels can be chosen. For example, being distressed about the loss of a loved one can be labeled as 'grief' and being distressed about the loss of an opportunity can be labeled as 'regret'. The OCC model does not pursue further differentiation of 'joy' and 'distress' besides the emotion types shown at the bottom of figure 2.1, but this is certainly an interesting direction for future research.

3.2.2. *Attribution Emotion Types*

The OCC model considers one differentiation in the action-based emotion types, namely in the actor. By differentiating with respect to the concept of *cognitive unit* (see page 7), the action-based emotion types can be captured as follows:

$$\mathbf{Pride}_i^{\mathbf{T}}(j{:}a) \overset{\text{def}}{=} \mathbf{Approving}_i^{\mathbf{T}}(j{:}a) \wedge \mathbf{CogUnit}_i(j) \tag{3.16}$$

$$\mathbf{Shame}_i^{\mathbf{T}}(j{:}a) \overset{\text{def}}{=} \mathbf{Disapproving}_i^{\mathbf{T}}(j{:}a) \wedge \mathbf{CogUnit}_i(j) \tag{3.17}$$

$$\mathbf{Admiration}_i^{\mathbf{T}}(j{:}a) \overset{\text{def}}{=} \mathbf{Approving}_i^{\mathbf{T}}(j{:}a) \wedge \neg\mathbf{CogUnit}_i(j) \tag{3.18}$$

$$\mathbf{Reproach}_i^{\mathbf{T}}(j{:}a) \overset{\text{def}}{=} \mathbf{Disapproving}_i^{\mathbf{T}}(j{:}a) \wedge \neg\mathbf{CogUnit}_i(j) \tag{3.19}$$

$\mathbf{CogUnit}_i(j)$ is read as "agent i views agent j as being in a cognitive unit with itself." $\mathbf{Pride}_i^{\mathbf{T}}(j{:}a)$ is then read as "pride about action a of agent j is triggered for agent i," and similarly for 'shame', 'admiration', and 'reproach'.

3.2.3. *Attraction Emotion Types*

The OCC model does not structure the valenced reactions to aspects of objects, even though OCC admit that momentary reactions of liking and disliking are among the most salient experiences for humans. Interestingly, they do consider one variable affecting the intensity of liking and disliking reactions (besides appealingness), namely *familiarity*, but they have chosen not to differentiate based on this variable. Differentiating based on familiarity would not be correct because the relation between familiarity and overall liking or disliking is not monotonic [27, 26, 25]. As is also suggested by the proverb "familiarity breeds contempt," liking of an object can decrease when one is very familiar with it, even though initially, liking usually increases with familiarity. Indeed, in the OCC model it is suggested that the relation between familiarity and overall liking probably follows a bell shape.

3.3. Compounds

Two branches of figure 2.1 combine to form the so-called compound emotion types. These emotions arise when one focuses on both the praiseworthiness of an action and the desirability of the *related* consequences. According to OCC, the eliciting conditions of the compound emotion types are a conjunction of the eliciting conditions of an event-based emotion ('joy' or 'distress') and an action-based emotion ('pride', 'shame', 'admiration', or 'reproach'), together with an assertion about their relatedness. However, realizing that an action and a consequence of an event are related may come at a later time than perceiving either the action or the consequence. Therefore, we need to be able to look back in time when describing the eliciting conditions of the compound emotion types. To this end, we use the construct $\mathbf{Past}\,\varphi$, which asserts that φ was true sometime in the past, where, importantly, the past is understood to include the present. The compound emotion types can then be captured as follows:

$$\begin{aligned}
\mathbf{Gratification}^{\mathbf{T}}_i(j{:}a, c) &\overset{\text{def}}{=} \mathbf{Past}\,\mathbf{Pride}^{\mathbf{T}}_i(j{:}a) \wedge \mathbf{Past}\,\mathbf{Joy}^{\mathbf{T}}_i(c) \\
&\qquad\qquad \wedge\, \mathbf{PerceiveRelated}_i(j{:}a, c) \qquad (3.20) \\
\mathbf{Remorse}^{\mathbf{T}}_i(j{:}a, c) &\overset{\text{def}}{=} \mathbf{Past}\,\mathbf{Shame}^{\mathbf{T}}_i(j{:}a) \wedge \mathbf{Past}\,\mathbf{Distress}^{\mathbf{T}}_i(c) \\
&\qquad\qquad \wedge\, \mathbf{PerceiveRelated}_i(j{:}a, c) \qquad (3.21) \\
\mathbf{Gratitude}^{\mathbf{T}}_i(j{:}a, c) &\overset{\text{def}}{=} \mathbf{Past}\,\mathbf{Admiration}^{\mathbf{T}}_i(j{:}a) \wedge \mathbf{Past}\,\mathbf{Joy}^{\mathbf{T}}_i(c) \\
&\qquad\qquad \wedge\, \mathbf{PerceiveRelated}_i(j{:}a, c) \qquad (3.22) \\
\mathbf{Anger}^{\mathbf{T}}_i(j{:}a, c) &\overset{\text{def}}{=} \mathbf{Past}\,\mathbf{Reproach}^{\mathbf{T}}_i(j{:}a) \wedge \mathbf{Past}\,\mathbf{Distress}^{\mathbf{T}}_i(c) \\
&\qquad\qquad \wedge\, \mathbf{PerceiveRelated}_i(j{:}a, c) \qquad (3.23)
\end{aligned}$$

$\mathbf{PerceiveRelated}_i(j{:}a, c)$ is read as "agent i perceives action a of agent j as being related to consequence c." $\mathbf{Gratification}^{\mathbf{T}}_i(j{:}a, c)$ is then read as "gratification about action a of agent j and the related consequence c is triggered for agent i," and similarly for 'remorse', 'gratitude', and 'anger'.

In order to ensure that the action appearing twice in each of these definitions is really the same action, it is assumed that all actions are unique. This can be seen as each performed action being a unique *instance* of an action. For example, $j{:}a$ in $\mathbf{Gratification}_i^{\mathbf{T}}(j{:}a, c)$ appears in both $\mathbf{Pride}_i^{\mathbf{T}}(j{:}a)$ and $\mathbf{PerceiveRelated}_i(j{:}a, c)$. With this assumption of uniqueness, the action that is the object of the perceived relation (i.e., $j{:}a$ in $\mathbf{PerceiveRelated}_i(j{:}a, c)$) must really be the same action as the one which is the object of the earlier action-based emotion (e.g., $j{:}a$ in $\mathbf{Pride}_i^{\mathbf{T}}(j{:}a)$).

3.4. Derived Emotion Types

In this section we will provide semiformal descriptions of the eliciting conditions of the bottom layer (of figure 2.1) of emotion types.

3.4.1. *Prospect-based Emotion Types*

Whereas 'hope' and 'fear' concern *unconfirmed* prospects of events, the OCC model also distinguishes emotion types concerning *confirmed* and *disconfirmed* prospects, namely 'satisfaction', 'fears-confirmed', 'relief', and 'disappointment'. As explained in section 2.2, a confirmation or disconfirmation is regarded as an actual consequence of an event, and therefore these emotion types are specializations of 'joy' and 'distress'. However, they do depend on an earlier instance of 'hope' or 'fear', so the formalizations below use the $\mathbf{Past}$ operator to capture this temporal link. These four prospect-based emotion types can then be captured as follows:

$$\mathbf{Satisfaction}_i^{\mathbf{T}}(c, c') \overset{\text{def}}{=} \mathbf{Joy}_i^{\mathbf{T}}(c) \land \mathbf{Past}\,\mathbf{Hope}_i^{\mathbf{T}}(c') \land \mathbf{Confirms}_i(c, c') \tag{3.24}$$

$$\mathbf{Fears\text{-}confirmed}_i^{\mathbf{T}}(c, c') \overset{\text{def}}{=} \mathbf{Distress}_i^{\mathbf{T}}(c) \land \mathbf{Past}\,\mathbf{Fear}_i^{\mathbf{T}}(c') \land \mathbf{Confirms}_i(c, c') \tag{3.25}$$

$$\mathbf{Relief}_i^{\mathbf{T}}(c, c') \overset{\text{def}}{=} \mathbf{Joy}_i^{\mathbf{T}}(c) \land \mathbf{Past}\,\mathbf{Fear}_i^{\mathbf{T}}(c') \land \mathbf{Disconfirms}_i(c, c') \tag{3.26}$$

$$\mathbf{Disappointment}_i^{\mathbf{T}}(c, c') \overset{\text{def}}{=} \mathbf{Distress}_i^{\mathbf{T}}(c) \land \mathbf{Past}\,\mathbf{Hope}_i^{\mathbf{T}}(c') \land \mathbf{Disconfirms}_i(c, c') \tag{3.27}$$

$\mathbf{Confirms}_i(c, c')$ is read as "agent i considers consequence c as (partially) confirming consequence c'", and likewise for 'disconfirm'. $\mathbf{Satisfaction}_i^{\mathbf{T}}(c, c')$ is then read as "satisfaction about consequence c confirming consequence c' is triggered for agent i," and similarly for 'fears-confirmed', 'relief', and 'disappointment'.

3.4.2. *Fortunes-of-others Emotion Types*

Finally, the four so-called fortunes-of-others emotion types are also specializations of 'joy' and 'distress', as explained in section 2.2. These emotion types concern consequences of events *presumed* to be desirable or undesirable for someone else. In order to capture presumptions, we introduce the $\mathbf{Presume}$ operator; $\mathbf{Presume}_i\varphi$ is read as "agent i presumes φ (to be true)." When grounding these semiformal specifications in a BDI-based logic (as we will do later), the presume operator can easily be conflated with belief. However, in order to remain independent of any underlying formalism, we stick to OCC's phrasing at this point and use 'presume' as the name for the operator. The fortunes-of-others emotion types can then be captured as follows:

$$\mathbf{Happy\text{-}for}_i^{\mathbf{T}}(c,j) \overset{\text{def}}{=} \mathbf{Joy}_i^{\mathbf{T}}(c) \land \mathbf{Presume}_i\mathbf{Des}_j(c) \tag{3.28}$$

$$\mathbf{Pity}_i^{\mathbf{T}}(c,j) \overset{\text{def}}{=} \mathbf{Distress}_i^{\mathbf{T}}(c) \land \mathbf{Presume}_i\mathbf{Undes}_j(c) \tag{3.29}$$

$$\mathbf{Gloating}_i^{\mathbf{T}}(c,j) \overset{\text{def}}{=} \mathbf{Joy}_i^{\mathbf{T}}(c) \land \mathbf{Presume}_i\mathbf{Undes}_j(c) \tag{3.30}$$

$$\mathbf{Resentment}_i^{\mathbf{T}}(c,j) \overset{\text{def}}{=} \mathbf{Distress}_i^{\mathbf{T}}(c) \land \mathbf{Presume}_i\mathbf{Des}_j(c) \tag{3.31}$$

$\mathbf{Happy\text{-}for}_i^{\mathbf{T}}(c,j)$ is read as "happy-for about consequence c (of an event) for agent j is triggered for agent i," and similarly for 'pity', 'gloating', and 'resentment'.

It may be interesting to note that it is not required for agent i to presume that agent j is aware of the event in question as well. For example, suppose Alice has just learned that she has won a magnificent cruise for two. She may feel very happy for her husband (who she intends to take the cruise with) without him being aware of the prize yet. Of course, Alice may feel inclined to tell her husband about the prize as soon as possible, but it would be unreasonable to argue that she cannot feel happy for him before having informed him.

3.5. Properties

To finish this section, we will show some properties of the (semiformal) specifications presented in this section. So far we have 'reduced' the eliciting conditions of the emotion types of the OCC model to formulas involving some standard logical connectives and the following seventeen constructs:

PerceiveConseq
PerceiveAction
PerceiveObject
PerceiveRelated
Past
Des
Undes
Praisew
Blamew
Appeal
Unappeal
Prospective
Actual
CogUnit
Confirms
Disconfirms
Presume

If the specifications presented in this section are accurate, then the eliciting conditions of the emotion types of the OCC model are constructed around no more than seventeen[8] notions, represented by the seventeen constructs above. About half of these constructs will be grounded in dynamic doxastic logic in the next section, whereas the remaining constructs will be grounded in KARO (which is a BDI-based extension of dynamic doxastic logic) in the section after that.

In the following properties, let $\vdash_{\mathrm{T}}$ (where T stands for trigger) be a classical propositional entailment relation with formulas (3.1)–(3.31) as axioms. Furthermore, $\Gamma \vdash_{\mathrm{T}} \varphi$ denotes that φ is derivable assuming Γ. Although formal proofs of propositions appearing later in this paper are provided in Appendix A, the derivations of the propositions below only involve manipulation of regular propositional connectives and therefore we deem it as unnecessary to spell out these derivations.

The following properties read exactly as the type specifications for pleased, displeased, approving, disapproving, liking, and disliking given in table 2.1.

$$\vdash_{\mathrm{T}} \mathbf{Pleased}_i^{\mathbf{T}}(c) \leftrightarrow \mathbf{Positive}_i^{\mathbf{T}}(c) \tag{3.32}$$

$$\vdash_{\mathrm{T}} \mathbf{Displeased}_i^{\mathbf{T}}(c) \leftrightarrow \mathbf{Negative}_i^{\mathbf{T}}(c) \tag{3.33}$$

$$\vdash_{\mathrm{T}} \mathbf{Approving}_i^{\mathbf{T}}(i{:}a) \leftrightarrow \mathbf{Positive}_i^{\mathbf{T}}(i{:}a) \tag{3.34}$$

$$\vdash_{\mathrm{T}} \mathbf{Disapproving}_i^{\mathbf{T}}(i{:}a) \leftrightarrow \mathbf{Negative}_i^{\mathbf{T}}(i{:}a) \tag{3.35}$$

$$\vdash_{\mathrm{T}} \mathbf{Liking}_i^{\mathbf{T}}(x) \leftrightarrow \mathbf{Positive}_i^{\mathbf{T}}(x) \tag{3.36}$$

$$\vdash_{\mathrm{T}} \mathbf{Disliking}_i^{\mathbf{T}}(x) \leftrightarrow \mathbf{Negative}_i^{\mathbf{T}}(x) \tag{3.37}$$

For example, table 2.1 states that “*pleased* is being *positive* about a consequence (of an event).” So if we put a consequence c into ‘positive’ (i.e., $\mathbf{Positive}_i^{\mathbf{T}}(c)$), we should get ‘pleased’. And indeed, formula (3.32) states that $\mathbf{Positive}_i^{\mathbf{T}}(c)$ is equivalent to $\mathbf{Pleased}_i^{\mathbf{T}}(c)$. The other properties follow the same pattern.

The following properties state that each pair of ‘siblings’ in the third layer of figure 2.1 completely subdivide their ‘parent’.

$$\Gamma \vdash_{\mathrm{T}} \mathbf{Pleased}_i^{\mathbf{T}}(c) \leftrightarrow (\mathbf{Hope}_i^{\mathbf{T}}(c) \lor \mathbf{Joy}_i^{\mathbf{T}}(c)) \tag{3.38}$$

$$\Gamma \vdash_{\mathrm{T}} \mathbf{Displeased}_i^{\mathbf{T}}(c) \leftrightarrow (\mathbf{Fear}_i^{\mathbf{T}}(c) \lor \mathbf{Distress}_i^{\mathbf{T}}(c)) \tag{3.39}$$

$$\vdash_{\mathrm{T}} \mathbf{Approving}_i^{\mathbf{T}}(i{:}a) \leftrightarrow (\mathbf{Pride}_i^{\mathbf{T}}(j{:}a) \lor \mathbf{Admiration}_i^{\mathbf{T}}(j{:}a)) \tag{3.40}$$

$$\vdash_{\mathrm{T}} \mathbf{Disapproving}_i^{\mathbf{T}}(i{:}a) \leftrightarrow (\mathbf{Shame}_i^{\mathbf{T}}(j{:}a) \lor \mathbf{Reproach}_i^{\mathbf{T}}(j{:}a)) \tag{3.41}$$

where $\Gamma = \mathbf{PerceiveConseq}_i(c) \to (\mathbf{Actual}_i(c) \lor \mathbf{Prospective}_i(c))$.

[8] Not counting propositional connectives.

The following properties state that ‘siblings’ in the third layer of figure 2.1 exclude each other.

$$\Gamma \vdash_{\mathrm{T}} \neg(\mathbf{Hope}_i^{\mathbf{T}}(c) \wedge \mathbf{Joy}_i^{\mathbf{T}}(c)) \tag{3.42}$$

$$\Gamma \vdash_{\mathrm{T}} \neg(\mathbf{Fear}_i^{\mathbf{T}}(c) \wedge \mathbf{Distress}_i^{\mathbf{T}}(c)) \tag{3.43}$$

$$\vdash_{\mathrm{T}} \neg(\mathbf{Pride}_i^{\mathbf{T}}(j{:}a) \wedge \mathbf{Admiration}_i^{\mathbf{T}}(j{:}a)) \tag{3.44}$$

$$\vdash_{\mathrm{T}} \neg(\mathbf{Shame}_i^{\mathbf{T}}(j{:}a) \wedge \mathbf{Reproach}_i^{\mathbf{T}}(j{:}a)) \tag{3.45}$$

where $\Gamma = \neg(\mathbf{Actual}_i(c) \wedge \mathbf{Prospective}_i(c))$. Together with the previous set of properties, this means that the differentiations directly below pleased/displeased, approving/disapproving, and liking/disliking are strict and complete. It should be noted that these properties do *not* express that, e.g., an agent cannot experience fear and distress at the same time. The fact that $\mathbf{Fear}_i^{\mathbf{T}}(c) \wedge \mathbf{Distress}_i^{\mathbf{T}}(c)$ is a contradiction means that the perception of one consequence c cannot *trigger* both fear and distress (because the triggering conditions for the ‘fear’ and ‘distress’ emotion types exclude each other).

The following properties state that ‘awareness’ of what one finds desirable and undesirable leads to ‘joy’ being equivalent to “happy-for-self” and ‘distress’ being equivalent to “self-pity.”

$$\Gamma_1 \vdash_{\mathrm{T}} \mathbf{Joy}_i^{\mathbf{T}}(c) \leftrightarrow \mathbf{Happy\text{-}for}_i^{\mathbf{T}}(c, i) \tag{3.46}$$

$$\Gamma_2 \vdash_{\mathrm{T}} \mathbf{Distress}_i^{\mathbf{T}}(c) \leftrightarrow \mathbf{Pity}_i^{\mathbf{T}}(c, i) \tag{3.47}$$

where $\Gamma_1 = \mathbf{Des}_i(c) \rightarrow \mathbf{Presume}_i\mathbf{Des}_i(c)$
and $\Gamma_2 = \mathbf{Undes}_i(c) \rightarrow \mathbf{Presume}_i\mathbf{Undes}_i(c)$.

The following properties state that proper ‘pride’ and ‘shame’ (in the sense that the agent of the praiseworthy/blameworty action in question is exactly the self) are equivalent to “self-approving” and “self-disapproving,” respectively.

$$\Gamma \vdash_{\mathrm{T}} \mathbf{Pride}_i^{\mathbf{T}}(i{:}a) \leftrightarrow \mathbf{Approving}_i^{\mathbf{T}}(i{:}a) \tag{3.48}$$

$$\Gamma \vdash_{\mathrm{T}} \mathbf{Shame}_i^{\mathbf{T}}(i{:}a) \leftrightarrow \mathbf{Disapproving}_i^{\mathbf{T}}(i{:}a) \tag{3.49}$$

where $\Gamma = \mathbf{CogUnit}_i(i)$. Note that these properties read exactly as the specifications for ‘pride’ and ‘shame’ in table 2.1, e.g., “pride is approving of one’s own action.”

The inheritance-based view of the eliciting conditions of emotions, as illustrated in figure 2.1, raises the expectation that each depicted emotion type implies its parent. Indeed, chains of implications such as the one below can be made for all emotion types (except for the compound emotion types[9]). For

[9] Each of the compound emotion types does inherit the eliciting conditions of its parents (e.g., ‘remorse’ inherits from ‘distress’ and ‘shame’), but they are preceded by a $\mathbf{Past}$ operator because the two inherited sets of conditions do not have to be satisfied at the same time. Still, a chain of implications can be made if one allows it to be “contaminated” by a $\mathbf{Past}$ operator.

example:

$$\vdash_{\mathrm{T}} \mathbf{Gloating}^{\mathbf{T}}_i(c,j) \to \mathbf{Joy}^{\mathbf{T}}_i(c)$$
$$\vdash_{\mathrm{T}} \mathbf{Joy}^{\mathbf{T}}_i(c) \to \mathbf{Pleased}^{\mathbf{T}}_i(c)$$
$$\vdash_{\mathrm{T}} \mathbf{Pleased}^{\mathbf{T}}_i(c) \to \mathbf{Positive}^{\mathbf{T}}_i(c)$$

Finally, it is worth emphasizing that the OCC model does not require appraisal to be consistent. Indeed, the following propositions are *not* derivable.

$$\nvdash_{\mathrm{T}} \neg(\mathbf{Des}_i(c) \wedge \mathbf{Undes}_i(c)) \tag{3.50}$$
$$\nvdash_{\mathrm{T}} \neg(\mathbf{Praisew}_i(j{:}a) \wedge \mathbf{Blamew}_i(j{:}a)) \tag{3.51}$$
$$\nvdash_{\mathrm{T}} \neg(\mathbf{Appeal}_i(x) \wedge \mathbf{Unappeal}_i(x)) \tag{3.52}$$

So it is *not* assumed that an agent's goals, standards, and attitudes are consistent. This implies that 'mixed feelings' are possible, i.e., formulas such as $\mathbf{Admiration}^{\mathbf{T}}_i(j{:}a) \wedge \mathbf{Reproach}^{\mathbf{T}}_i(j{:}a)$ are satisfiable. In fact, for each pair of 'opposing' emotion types (i.e., those sharing a box in figure 2.1), we have that their eliciting conditions do not exclude each other. With slight abuse of notation, this can be expressed as follows.

$$\nvdash_{\mathrm{T}} \neg(\mathbf{Emotion}^{+\mathbf{T}}_i(X) \wedge \mathbf{Emotion}^{-\mathbf{T}}_i(X)) \tag{3.53}$$

where, e.g., $\mathbf{Emotion}^{+\mathbf{T}}_i(X)$ stands for $\mathbf{Hope}^{\mathbf{T}}_i(c)$ and $\mathbf{Emotion}^{-\mathbf{T}}_i(X)$ stands for $\mathbf{Fear}^{\mathbf{T}}_i(c)$.

4. Grounding in Dynamic Doxastic Logic

We will now introduce a formalism that will ground many of the constructs used in the previous section to (semiformally) specify the eliciting conditions of the emotion types of the OCC model. We have chosen to use dynamic doxastic logic, because this is a well-understood formalism which readily provides ways for reasoning about agents and their actions (because it is dynamic) and beliefs (because it is doxastic). Furthermore, this dynamic perspective allows for a straightforward representation of events and their consequences. Although dynamic doxastic logic is not concerned with objects, we will introduce a reasonable way of representing them as well.

We emphasize that it is perfectly possible to choose another formalism than we did, as long as it supports reasoning about events and desirability, actions and praiseworthiness, and objects and appealingness, as well as some temporal constructs for the prospect-based emotions. We do not define the appraisal constructs (desirability, praiseworthiness, appealingness) in pure dynamic doxastic logic in this section, because it lacks ways of representing goals, standards, and attitudes. In section 5, then, we add BDI-based constructs and finish the grounding of the specification of eliciting conditions of emotions. Also, we will not be concerned with formal semantics until section 5.

4.1. Basic Operators

It is assumed there exists a set ATM of atomic propositions, with typical element p. Furthermore, we use the propositional connectives $\neg$, $\wedge$, $\vee$, $\rightarrow$, and $\leftrightarrow$ with their usual interpretation, as well as $\bot$ for falsum and $\top$ for verum. We then typically use φ and ψ to denote arbitrary formulas.

In dynamic doxastic logic there are of course two modal operators, namely for belief and action. They are expressed and read as follows.

$\mathbf{B}_i\varphi$: Agent i believes φ (to be true).
$[i{:}\alpha]\varphi$: After the execution of action α by agent i, φ holds.

In the following we will also use converse actions (denoted as α^{-}), which are useful for expressing what was true before the execution of an action. For example, $[i{:}\alpha^{-}]\varphi$ expresses that, if it is the case that agent i has just performed action α, then φ was true before that action. Because we will often need to express that some agent has just performed some action, we define a convenient shorthand for this, as follows:

$$\mathbf{Done}(i{:}\alpha) \overset{\text{def}}{=} \langle i{:}\alpha^{-}\rangle\top \tag{4.1}$$

$\mathbf{Done}(i{:}\alpha)$ can thus be read as "agent i has just performed action α." Note that angled brackets are used to denote the dual of the action modality, as usual.

In dynamic logic, actions are used as an abstraction of time, which means that temporal operators can be interpreted over actions. We will be using the following three basic temporal and action-based operators:

$\mathbf{Prev}\,\varphi$: In the previous state, φ was true. We say "*the* previous state," because we assume a linear history (and a branching future).
$\mathbf{Past}\,\varphi$: Some time in the past (*including* the present), φ was/is true. Intuitively, this comes down to $\varphi \vee \mathbf{Prev}\,\varphi \vee \mathbf{Prev}\,\mathbf{Prev}\,\varphi \vee \ldots$, but since this is an infinite formula, the past operator cannot be defined as an abbreviation.
$\mathbf{Fut}\,\varphi$: Some time in the future (*including* the present), φ may be true. Intuitively, this can be seen as an existential quantification over agents and actions (*cf.* formula (4.29) on page 28).

Although the $\mathbf{Past}$ and $\mathbf{Fut}$ operators are—as usual—understood to include the present, in the following we will mostly be using the future operator in situations that exclude the present. For convenience, then, we define a strict version of the future operator as follows.

$$\mathbf{Fut}^{+}\varphi \overset{\text{def}}{=} \neg\varphi \land \mathbf{Fut}\,\varphi \tag{4.2}$$

$\mathbf{Fut}^{+}\varphi$ is then read as "some time in the future, but not presently, φ may be true."

In the following subsections, we will show how we define many of the constructs used in section 3 to model the emotion triggers. We will define these constructs as abbreviations (i.e. using $\overset{\text{def}}{=}$) of the operators just introduced. Some constructs, however, will be left undefined even here. These are the appraisal operators (i.e., $\mathbf{Des}$, $\mathbf{Undes}$, $\mathbf{Praisew}$, $\mathbf{Blamew}$, $\mathbf{Appeal}$, and $\mathbf{Unappeal}$), confirmation operators ($\mathbf{Confirms}$, $\mathbf{Disconfirms}$), and the cognitive unit operator ($\mathbf{CogUnit}$). There are separate reasons for this, which will be explained in section 4.5.

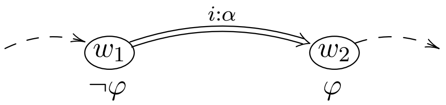
The next subsections are structured as follows. First we will define the constructs used for the event-based emotion types, then those for the action-based emotion types, and then those for the object-based emotion types. Finally, several properties of the presented definitions will be discussed in section 4.5.

4.2. Events and their Consequences

To start simple, we conflate presuming with believing:

$$\mathbf{Presume}_i\varphi \overset{\text{def}}{=} \mathbf{B}_i\varphi \tag{4.3}$$

The largest branch of the OCC model is concerned with valenced reactions to events; however, events are said to always be appraised with respect to their consequences. Now let us consider the distinction between consequences and events in more detail. The usual view in dynamic logic is that the execution of an action is regarded as an event. This makes sense because in dynamic logic, time passes only through the execution of actions, i.e., through a succession of events. Here we will follow this view and only regard executions of actions as events. We then consider a consequence of an event to be anything that was not true directly before the event, but is true directly after the event. This idea can be illustrated as follows.



This figure illustrates that state w_2 is the result of event $i{:}\alpha$, i.e., the execution of action α by agent i. Now any formula φ that is true in state w_2 (i.e.,

$w_2 \models \varphi$) but was not true in the previous state called w_1 (i.e., $w_1 \models \neg\varphi$) is considered to be a *consequence* of the event $i{:}\alpha$. It will be clear that an event can also have multiple or no consequences.

For convenience, then, we introduce the following construct to capture consequences of events:

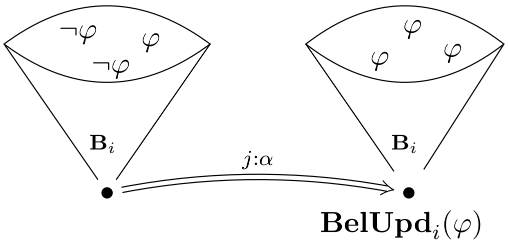
$$\mathbf{New}\,\varphi \overset{\text{def}}{=} \varphi \wedge \neg\mathbf{Prev}\,\varphi \tag{4.4}$$

$\mathbf{New}\,\varphi$ reads as: φ was not true in the previous state but φ is true in the current state. Note that $\mathbf{Prev}\,\psi$ expresses that ψ was true before the execution of the latest action, i.e., before the latest event. Therefore, if $\mathbf{New}\,\varphi$ holds for some formula φ, then φ can be regarded as a consequence of an event. Indeed, we have that $w_2 \models \mathbf{New}\,\varphi$ in the illustration above.

Even though $\mathbf{New}\,\varphi$ expresses that φ is a consequence of an event, it may very well be that no agent is aware of this consequence. There must be a change in an agent's beliefs before we can say that it perceives a consequence of an event (or anything in general). Belief changes (updates) can easily be defined using the $\mathbf{New}$ operator, as follows:

$$\mathbf{BelUpd}_i(\varphi) \overset{\text{def}}{=} \mathbf{New}\,\mathbf{B}_i\varphi \tag{4.5}$$

$\mathbf{BelUpd}_i(\varphi)$ reads as: the beliefs of agent i have just been updated with φ. A situation where $\mathbf{BelUpd}_i(\varphi)$ holds can be illustrated as follows:



In the state before the event $j{:}\alpha$, agent i does not believe φ, i.e., it envisages worlds where $\neg\varphi$ holds. In the state after the event $j{:}\alpha$, agent i believes φ, i.e., in all worlds it holds as possible, φ is true.[10] In that state, then, $\mathbf{BelUpd}_i(\varphi)$ is true. It should be noted that $\mathbf{BelUpd}_i(\varphi)$ says nothing about the event (e.g., the action $j{:}\alpha$ in the illustration above) that actually 'caused' the belief update; all it expresses is that *something* happened and as a consequence, agent i believes φ to be true. In other words, from the perspective of agent i, φ is a consequence of an event.

[10] With the danger of getting ahead of ourselves, we use possible world semantics to illustrate these definitions. Indeed, we will ground the belief and action modalities using possible world semantics in section 5. The definitions given in the present section do not really depend on such semantics; these illustration, then, only serve to get a feeling for what the defined constructs express.

With respect to emotions concerning consequences of events, the OCC model distinguishes between the types 'hope' and 'fear' on the one hand, and 'joy' and 'distress' on the other hand, based on whether the consequence in question is prospective or actual, respectively. As noted in section 2.2, the notion of "prospect" is intentionally ambiguous; it is used to describe both future consequences and uncertain consequences. In section 3 we used $\mathbf{Prospective}_i(\varphi)$ and $\mathbf{Actual}_i(\varphi)$ to express that agent i considers φ to be a prospective or actual consequence of an event, respectively. Using definitions similar to $\mathbf{BelUpd}$ above, we define $\mathbf{Prospective}$ and $\mathbf{Actual}$ as follows:

$$\mathbf{Prospective}_i(\varphi) \overset{\text{def}}{=} \mathbf{FutUpd}_i(\varphi) \vee \mathbf{UncUpd}_i(\varphi) \tag{4.6}$$

$$\mathbf{Actual}_i(\varphi) \overset{\text{def}}{=} \mathbf{BelUpd}_i(\varphi) \tag{4.7}$$

where

$$\mathbf{FutUpd}_i(\varphi) \overset{\text{def}}{=} \mathbf{New}\,\mathbf{B}_i\mathbf{Fut}^+\varphi \tag{4.8}$$

$$\mathbf{UncUpd}_i(\varphi) \overset{\text{def}}{=} \mathbf{New}\,(\neg\mathbf{B}_i\varphi \wedge \neg\mathbf{B}_i\neg\varphi) \tag{4.9}$$

We thus split the definition of $\mathbf{Prospective}$ into two cases in order to capture future as well as uncertain prospects. The definition of $\mathbf{Actual}$ is the same as $\mathbf{BelUpd}$ (see also the illustration above). The definition of $\mathbf{FutUpd}$ is also like $\mathbf{BelUpd}$ but then with φ replaced by $\neg\varphi \wedge \mathbf{Fut}\,\varphi$, i.e., agent i comes to believe that φ is not true but that there exists a future in which φ will be true. The definition of $\mathbf{UncUpd}$ ("uncertainty update") also resembles $\mathbf{BelUpd}$. If, in the illustration for $\mathbf{BelUpd}$ above, the left 'cloud' would contain either only φ's or only $\neg\varphi$'s, and the right 'cloud' would contain a mixture of φ's and $\neg\varphi$'s, then $\mathbf{UncUpd}_i(\varphi)$ would be true in the bottom right state. $\mathbf{UncUpd}_i(\varphi)$ thus expresses that agent i has just become uncertain about whether or not φ holds.

$\mathbf{Prospective}_i(\varphi)$ and $\mathbf{Actual}_i(\varphi)$ are now defined such that they cover both cases for perceiving consequences of events that are distinguished in the OCC model. This means that $\mathbf{PerceiveConseq}_i(\varphi)$, which we used to express that agent i perceives consequence φ of an event, can be (trivially) defined as follows:

$$\mathbf{PerceiveConseq}_i(\varphi) \overset{\text{def}}{=} \mathbf{Prospective}_i(\varphi) \vee \mathbf{Actual}_i(\varphi) \tag{4.10}$$

In effect, this definition specifies that perceiving a consequence of an event means either perceiving a prospective consequence or an actual consequence. Indeed, these are exactly the two cases distinguished in the OCC model. If one wishes to distinguish more kinds of consequences of events, more disjuncts covering those cases could be added.

4.3. Agents and their Actions

It is assumed there exists a set AGT of agent names with typical elements i and j, and a set ACT of actions with typical element α. Furthermore, it is assumed each action is unique, i.e., it can be performed only once. This can be seen as each performed action being a unique *instance* of an action.

In the previous section, we used $\mathbf{PerceiveAction}_i(j{:}\alpha)$ to express that agent i perceives that agent j has performed action α. Making use of the $\mathbf{BelUpd}$ construct introduced above, we define $\mathbf{PerceiveAction}$ as follows:

$$\mathbf{PerceiveAction}_i(j{:}\alpha) \overset{\text{def}}{=} \mathbf{BelUpd}_i(\mathbf{Past}\,\mathbf{Done}(j{:}\alpha)) \tag{4.11}$$

Note the use of the $\mathbf{Past}$ operator here. Because $\mathbf{Done}(j{:}\alpha)$ only expresses that agent j has *just* performed action α, we need the $\mathbf{Past}$ operator to express perceptions of actions that have been performed at some arbitrary time in the past. Thus $\mathbf{PerceiveAction}_i(j{:}\alpha)$ does not specify when exactly agent j performed action α, just that agent i now believes it did and that i did not believe so before.

For the compound emotion types (gratification, remorse, gratitude, anger) it was necessary to express a (presumed) relation between an action of an agent and a consequence. For this we used $\mathbf{PerceiveRelated}$, which we define here using $\mathbf{BelUpd}$ again:

$$\mathbf{PerceiveRelated}_i(j{:}\alpha, \varphi) \overset{\text{def}}{=} \mathbf{BelUpd}_i(\mathbf{Related}(j{:}\alpha, \varphi)) \tag{4.12}$$

$$\mathbf{Related}(i{:}\alpha, \varphi) \overset{\text{def}}{=} \mathbf{Past}\,(\mathbf{Done}(i{:}\alpha) \land \mathbf{New}\,\varphi) \tag{4.13}$$

For convenience we define a construct for relatedness separately (as it will be useful later on). $\mathbf{Related}(i{:}\alpha, \varphi)$ expresses that some time in the past, φ became true just when agent i had performed action α. We do not suggest this establishes a causal relationship between the action and the formula; indeed, the relation merely exists in their co-occurrence. Note that this definition correctly expresses a relation because of the assumption of uniqueness of actions. It should also be noted that the figure on page 23 illustrates this construct; in particular, $w_2 \models \mathbf{Related}(i{:}\alpha, \varphi)$. With these definitions, then, $\mathbf{PerceiveRelated}_i(j{:}\alpha, \varphi)$ expresses that agent i perceives action α of agent j to be related to consequence φ if and only if agent i comes to believe that φ became true exactly when action α was performed by agent j.

4.4. Objects and their Aspects

It is assumed there exists a set OBJ of object names with typical element x. The OCC model requires us to be able to view agents as objects, so it is required that AGT $\subseteq$ OBJ. Furthermore, it is assumed that there exists an atomic proposition for each object that identifies that object, i.e, $\{\, object_x \mid x \in \textsc{obj}\,\} \subseteq \textsc{atm}$. The notation $object_x$ is used to refer to the proposition identifying x as an object. For example, if $x = mona\_lisa$ ($\in \textsc{obj}$), then $object_x$ may be the proposition $mona\_lisa\_is\_an\_object$ ($\in \textsc{atm}$). Using this notation, the construct $\mathbf{PerceiveObject}$ used to capture the perception of objects can simply be defined as follows:

$$\mathbf{PerceiveObject}_i(x) \overset{\text{def}}{=} \mathbf{BelUpd}_i(object_x) \tag{4.14}$$

So perceiving an object is equated with a description of the object being added to the agent's beliefs.

4.5. Properties

We have so far 'reduced' the eliciting conditions of the emotion types of the OCC model to formulas involving propositional connectives and operators from dynamic doxastic logic (in particular, $\mathbf{B}$, $\mathbf{Prev}$, $\mathbf{Past}$, and $\mathbf{Fut}$). The following nine constructs are still undefined:

$\mathbf{Des}$	$\mathbf{Praisew}$	$\mathbf{Appeal}$	$\mathbf{Confirms}$	$\mathbf{CogUnit}$
$\mathbf{Undes}$	$\mathbf{Blamew}$	$\mathbf{Unappeal}$	$\mathbf{Disconfirms}$	

These constructs will be grounded in KARO (which extends dynamic doxastic logic) in the next section. There are several reasons why they have not been defined in this section. First, the six appraisal constructs ($\mathbf{Des}$, $\mathbf{Undes}$, $\mathbf{Praisew}$, $\mathbf{Blamew}$, $\mathbf{Appeal}$, $\mathbf{Unappeal}$) require the notions of goals, standards, and attitudes, which are absent in pure dynamic doxastic logic. Second, for the confirmation constructs ($\mathbf{Confirms}$, $\mathbf{Disconfirms}$) we want to be more precise about what is being compared first. From definitions (3.24)–(3.27) it can be observed that the things being compared for (dis)confirmation are objects of event-based emotion types, which means that they must be related to goals. Although it may be possible to define what it means for one consequence to confirm or disconfirm another consequence just in propositional logic, we can be more precise if we know what goals look like. Therefore, we postpone defining $\mathbf{Confirms}$ and $\mathbf{Disconfirms}$ until we have formalized goals. Third, a proper formalization of the notion of cognitive unit (as expressed by $\mathbf{CogUnit}$) would require substantially more (psychological) research, so we will leave this aspect mostly open in this paper.

To finish this section, we will show some properties of the formalization so far. In the following properties, let $\vdash_{\mathrm{DD}}$ (where DD stands for dynamic doxastic logic) be a normal modal entailment relation with formulas (3.1)–(3.31) and (4.1)–(4.14) as axioms. Then the following propositions are derivable.

$$\vdash_{\mathrm{DD}} \neg(\mathbf{FutUpd}_i(\varphi) \wedge \mathbf{UncUpd}_i(\varphi)) \tag{4.15}$$
$$\vdash_{\mathrm{DD}} \neg(\mathbf{BelUpd}_i(\varphi) \wedge \mathbf{UncUpd}_i(\varphi)) \tag{4.16}$$
$$\Gamma \vdash_{\mathrm{DD}} \neg(\mathbf{BelUpd}_i(\varphi) \wedge \mathbf{FutUpd}_i(\varphi)) \tag{4.17}$$
$$\Gamma \vdash_{\mathrm{DD}} \neg(\mathbf{Actual}_i(\varphi) \wedge \mathbf{Prospective}_i(\varphi)) \tag{4.18}$$
$$\Gamma \vdash_{\mathrm{DD}} \neg(\mathbf{Hope}_i^{\mathbf{T}}(\varphi) \wedge \mathbf{Joy}_i^{\mathbf{T}}(\varphi)) \tag{4.19}$$
$$\Gamma \vdash_{\mathrm{DD}} \neg(\mathbf{Fear}_i^{\mathbf{T}}(\varphi) \wedge \mathbf{Distress}_i^{\mathbf{T}}(\varphi)) \tag{4.20}$$
$$\vdash_{\mathrm{DD}} \mathbf{Pleased}_i^{\mathbf{T}}(c) \leftrightarrow (\mathbf{Hope}_i^{\mathbf{T}}(c) \vee \mathbf{Joy}_i^{\mathbf{T}}(c)) \tag{4.21}$$
$$\vdash_{\mathrm{DD}} \mathbf{Displeased}_i^{\mathbf{T}}(c) \leftrightarrow (\mathbf{Fear}_i^{\mathbf{T}}(c) \vee \mathbf{Distress}_i^{\mathbf{T}}(c)) \tag{4.22}$$

where $\Gamma = \neg(\mathbf{B}_i\varphi \wedge \mathbf{B}_i\neg\varphi)$. The first three properties state that $\mathbf{FutUpd}$, $\mathbf{UncUpd}$, and $\mathbf{BelUpd}$ are mutually exclusive. Because $\mathbf{Prospective}$ is defined in terms of $\mathbf{FutUpd}$ and $\mathbf{UncUpd}$, this immediately results in the fourth property. The fourth property then leads to the fifth and sixth properties, because (4.18) was exactly the assumption needed for properties (3.42) and (3.43). The seventh and eighth properties are the same as properties (3.38) and (3.39), except without needing assumptions due to the way $\mathbf{PerceiveConseq}$ (4.10) has been defined.

Below are several propositions showing how the dynamic and temporal operators introduced in this section interact. Because we have not yet introduced formal semantics, they cannot be called properties yet. The appendix provides formal proofs using the semantics introduced in section 5.

$$[i{:}\alpha]\mathbf{Done}(i{:}\alpha) \tag{4.23}$$
$$\mathbf{Done}(i{:}\alpha) \rightarrow \mathbf{Prev}\,\top \tag{4.24}$$
$$\mathbf{Prev}\,\top \rightarrow (\mathbf{Prev}\,\varphi \leftrightarrow \neg\mathbf{Prev}\,\neg\varphi) \tag{4.25}$$
$$\neg\mathbf{Prev}\,\top \wedge \varphi \rightarrow \mathbf{New}\,\varphi \tag{4.26}$$
$$\mathbf{Done}(i{:}\alpha) \wedge \varphi \rightarrow \mathbf{Prev}\,\langle i{:}\alpha\rangle\varphi \tag{4.27}$$
$$\mathbf{Prev}\,[i{:}\alpha]\varphi \wedge \mathbf{Done}(i{:}\alpha) \rightarrow \varphi \tag{4.28}$$
$$\mathbf{Fut}\,\varphi \leftrightarrow \varphi \vee \langle i_1{:}\alpha_1\rangle \cdots \langle i_n{:}\alpha_n\rangle\varphi \qquad (\exists i_1,\ldots,i_n, \exists\alpha_1,\ldots,\alpha_n) \tag{4.29}$$

The first proposition reads rather tautologically: after the execution of action α by agent i, i has done α. The second proposition states that, if an action has just been done, there must exist a previous state. The third propositions states that $\mathbf{Prev}$ is its own dual, provided that there exists a previous state. The fourth proposition states that if no previous state exists, all formulas that are true now are also 'new'. The fifth proposition states that everything that is true now must previously have been a possible result of the last performed action. As usual, $\langle\cdot\rangle$ abbreviates $\neg[\cdot]\neg$. The sixth proposition states that all necessary results of the last performed action must be true now. Finally, the seventh proposition states that $\mathbf{Fut}$ is like an existential quantification over agents and actions. Note that we slightly abuse notation here, because we do not have quantification in our object language.

5. Grounding in a BDI-based Logic

In this section we use KARO [22, 23] as a framework for grounding the formalization of the emotions of the OCC model. The KARO framework is a mixture of dynamic logic, epistemic / doxastic logic, and several additional (modal) operators for dealing with the motivational aspects of artificial agents. KARO was originally proposed as a specification logic for rational agents. It was thus designed to serve a purpose similar to that of the logics of Cohen & Levesque [3] and Rao & Georgeff [29]. A crucial difference with these approaches is that KARO is primarily based on dynamic logic [15] rather than temporal logic [9]. So one could view KARO as a kind of BDI (belief, desire, intention) logic based on dynamic logic. Although the specification of informational and motivational attitudes (such as knowledge and beliefs, and desires, goals, and commitments, respectively) had been the main aim for devising KARO ([16, 23]), the logic has also proven to be applicable for the description of agent behavior, more in general. For example, in [22] it has been used to specify four basic emotion types. Here we present a modest extension of KARO, including operators concerning past states, such that all emotion types of the OCC model can be specified in this framework.

5.1. Adding Motivational Constructs

In section 4.3 we introduced the set ACT of actions. From ACT we derive the set PLANS, consisting of sequential compositions of actions, with typical element π. PLANS is the smallest set such that ACT $\subseteq$ PLANS and if $\alpha \in$ ACT and $\pi \in$ PLANS, then $(\alpha\,;\pi) \in$ PLANS.

The notation of the dynamic operator is extended to sequential compositions of actions and its dual, as follows:

$$[i{:}(\alpha\,;\pi)]\varphi \overset{\text{def}}{=} [i{:}\alpha][i{:}\pi]\varphi \tag{5.1}$$

$$\langle i{:}\pi\rangle\varphi \overset{\text{def}}{=} \neg[i{:}\pi]\neg\varphi \tag{5.2}$$

In KARO, there are quite a few operators for expressing motivational attitudes of agents (*cf.* [16, 23]). Here we use three of them, namely for (achievement) goals, abilities, and commitments:

$\mathbf{G}_i\varphi$: Agent i has (achievement) goal φ. Here φ represents a state of affairs which agent i wants to achieve. A goal φ is said to have been *achieved* when agent i believes φ holds, i.e., when $\mathbf{B}_i\varphi$ is true.

$\mathbf{A}_i\pi$: Agent i has the ability to do π.
$\mathbf{Com}_i(\pi)$: Agent i is committed to doing π.

Using these operators, several constructs are defined expressing (possible) motivations of an agent:

$$\mathbf{PracPoss}_i(\pi, \varphi) \overset{\text{def}}{=} \mathbf{A}_i\pi \wedge \langle i{:}\pi\rangle\varphi \tag{5.3}$$

$$\mathbf{Can}_i(\pi, \varphi) \overset{\text{def}}{=} \mathbf{B}_i\mathbf{PracPoss}_i(\pi, \varphi) \tag{5.4}$$

$$\mathbf{PossIntend}_i(\pi, \varphi) \overset{\text{def}}{=} \mathbf{Can}_i(\pi, \varphi) \wedge \mathbf{B}_i(\mathbf{G}_i\varphi \wedge \neg\varphi) \tag{5.5}$$

An agent has the *practical possibility* to perform an action/plan π to bring about φ iff it has the ability to perform π and doing so may bring about φ. An agent *can* perform π to bring about φ iff it believes it has the practical possibility to do so. An agent has the *possible intention* to perform π to accomplish φ iff it *can* do so and it believes φ is an unachieved goal.

5.2. Semantics

KARO is a dynamic doxastic logic, so we will introduce belief models and action models. Belief models are of the form $\mathrm{M} = \langle S, R, V\rangle$, where

S is a non-empty set of states (or 'possible worlds').

$R = \{\, R_i \mid i \in \text{AGT}\,\}$ is a set of accessibility relations on S, one for each agent name, hence the notation R_i. So $R_i \subseteq S \times S$ for each $R_i \in R$.

$V : S \to 2^{\text{ATM}}$ is a valuation function, indicating which atomic propositions hold per state.

As is common in doxastic logic, each belief-accessibility relation R_i is required to be serial, transitive, and euclidean, i.e., the modal logic KD45 is used for belief models.

The semantics of actions are defined *over* the Kripke models of belief, as actions may change the mental states of agents. Action models are of the form $\mathcal{M} = \langle \mathcal{S}, \mathcal{R}, \mathrm{Aux}, \mathrm{Emo}\rangle$, where

$\mathcal{S}$ is a non-empty set of possible model–state pairs, where a model is of the form M as above and a state is from S therein. That is, if $(\mathrm{M}, s) \in \mathcal{S}$ and $\mathrm{M} = \langle S, R, V\rangle$ then it must be that $s \in S$.

$\mathcal{R} = \{\, \mathcal{R}_{i:\alpha} \mid i \in \text{AGT},\ \alpha \in \text{ACT}\,\}$ is a set of accessibility relations on $\mathcal{S}$. Each transition is labeled with an agent name and an action, hence the notation $\mathcal{R}_{i:\alpha}$. So $\mathcal{R}_{i:\alpha} \subseteq \mathcal{S} \times \mathcal{S}$ for each $\mathcal{R}_{i:\alpha} \in \mathcal{R}$.

$\mathrm{Aux} = \langle \mathit{Goals}, \mathit{Caps}, \mathit{Agd} \rangle$ is a structure of auxiliary functions, indicating per agent and model–state pair which goals (Goals), capabilities (Caps), and commitments (Agd) the agent has.

$\mathrm{Emo} = \langle \mathit{Des}, \mathit{Undes}, \mathit{Praisew}, \mathit{Blamew}, \mathit{Appeal}, \mathit{Unappeal}, \mathit{CogUnit} \rangle$ is a structure of appraisal and judgment functions, indicating per agent and model–state pair how that agent appraises consequences (Des, Undes), actions of agents ($\mathit{Praisew}$, Blamew), and objects (Appeal, $\mathit{Unappeal}$), and how it judges cognitive units ($\mathit{CogUnit}$).

In order to have a branching future and a single history, it is required that $\bigcup \mathcal{R}$ is injective. This ensures that any model–state pair can be reached from at most one other model–state pair. Note however that this does not exclude parallel actions. For example, if $((\mathrm{M}, s), (\mathrm{M}', s')) \in \mathcal{R}_{i:\alpha}$ and $((\mathrm{M}, s), (\mathrm{M}', s')) \in \mathcal{R}_{j:\beta}$, then state (M', s') is a result of the parallel execution of $i{:}\alpha$ and $j{:}\beta$ in state (M, s). With respect to the semantics of converse actions, let $\mathcal{R}_{i:\alpha^-} = (\mathcal{R}_{i:\alpha})^-$, as usual.

In the previous sections it was assumed that actions are unique. We are now in a position to formalize this assumption as a constraint on $\mathcal{R}$, as follows:

$$\forall \mathcal{R}_{i:\alpha} \in \mathcal{R} : \forall((\mathrm{M}, s), (\mathrm{M}', s')) \in \mathcal{R}_{i:\alpha} : \mathcal{R}_{i:\alpha} \cap (\mathcal{S}' \times \mathcal{S}') = \varnothing \tag{5.6}$$

where $\mathcal{S}' = \{\, (\mathrm{M}'', s'') \mid ((\mathrm{M}', s'), (\mathrm{M}'', s'')) \in (\bigcup \mathcal{R})^* \,\}$.[11] This constraint can be read as follows. If state (M', s') is a result of action α by agent i ($((\mathrm{M}, s), (\mathrm{M}', s')) \in \mathcal{R}_{i:\alpha}$), then no possible future state after (M', s') (collected in $\mathcal{S}'$) can be reachable by i doing α again ($\mathcal{R}_{i:\alpha} \cap (\mathcal{S}' \times \mathcal{S}') = \varnothing$). It may be interesting to note that this constraint implies that $\bigcup \mathcal{R}$ must be free of circles.

In line with previous work [33, 34], we define goal formulas as non-empty, consistent conjunctions of literals. This way goals can easily be broken up in subgoals; in particular, every non-empty 'subconjunction' of a goal formula is considered to be a subgoal. For example, if $p \wedge \neg q \wedge r$ is a goal, then p, $\neg q$, r, $p \wedge \neg q$, $p \wedge r$, $\neg q \wedge r$, and $p \wedge \neg q \wedge r$ are subgoals. Goal formulas are drawn from the set CCL which is defined as follows.

$$\mathrm{LIT} = \mathrm{ATM} \cup \{\, \neg p \mid p \in \mathrm{ATM} \,\} \tag{5.7}$$

$$\mathrm{CSL} = \{\, \Phi \mid \varnothing \subset \Phi \subseteq \mathrm{LIT},\ \Phi \nvdash_{PC} \bot \,\} \tag{5.8}$$

$$\mathrm{CCL} = \{\, \bigwedge \Phi \mid \Phi \in \mathrm{CSL} \,\} \tag{5.9}$$

where PC stands for Propositional Calculus (so each conjunction in CCL is consistent). So LIT is the set of literals, CSL is the set of consistent sets of literals, and CCL is the set of consistent conjunctions of literals.

[11] A^* denotes the reflexive transitive closure of relation A.

The mappings of the three auxiliary functions are now as follows. $Goals : \text{AGT} \times \mathcal{S} \to 2^{\text{CCL}}$ is a function returning the set of goals an agent has per model–state pair; $Caps : \text{AGT} \times \mathcal{S} \to 2^{\text{PLANS}}$ is a function that returns the set of actions that an agent is capable of performing per model–state pair; and $Agd : \text{AGT} \times \mathcal{S} \to 2^{\text{PLANS}}$ is a function that returns the set of actions that an agent is committed to (are on its 'agenda') per model–state pair. Note that it is *not* assumed that goals are mutually consistent.

Finally, the mappings of the appraisal and judgment functions are as follows. $Des, Undes : \text{AGT} \times \mathcal{S} \to 2^{\mathcal{L}}$ for desirability and undesirability (where $\mathcal{L}$ is the set of all well-formed formulas); $Praisew, Blamew : \text{AGT} \times \mathcal{S} \to 2^{\text{AGT} \times \text{ACT}}$ for praiseworthiness and blameworthiness; $Appeal, Unappeal : \text{AGT} \times \mathcal{S} \to 2^{\text{OBJ}}$ for appealingness and unappealingness; and $CogUnit : \text{AGT} \times \mathcal{S} \to 2^{\text{AGT}}$ for cognitive unit.

5.3. Interpretation in KARO

We now have all ingredients necessary for the interpretation of formulas, presented below. Formulas are interpreted in state s of model M, where $(\mathrm{M}, s) \in \mathcal{S}$. It should be noted that the pair (M, s) is itself a state of model $\mathcal{M}$, i.e., belief models (M) are nested in action models ($\mathcal{M}$). Strictly speaking, we should write $(\mathcal{M}, (\mathrm{M}, s)) \models \ldots$, but we drop the $\mathcal{M}$ for notational convenience.

Let $\mathcal{M} = \langle \mathcal{S}, \mathcal{R}, \mathrm{Aux}, \mathrm{Emo} \rangle$, $(\mathrm{M}, s) \in \mathcal{S}$, and $\mathrm{M} = \langle S, R, V \rangle$; formulas are then interpreted as follows.

Basic connectives:

$$\begin{array}{lll}
\mathrm{M}, s \models p & \text{iff} & p \in V(s) \qquad \text{for } p \in \text{ATM} \\
\mathrm{M}, s \models \neg\varphi & \text{iff} & \text{not } \mathrm{M}, s \models \varphi \\
\mathrm{M}, s \models \varphi \wedge \psi & \text{iff} & \mathrm{M}, s \models \varphi \quad \text{and} \quad \mathrm{M}, s \models \psi
\end{array}$$

Mental attitudes:

$$\begin{array}{lll}
\mathrm{M}, s \models \mathbf{B}_i\varphi & \text{iff} & \forall s' \in S : (s, s') \in R_i \quad \text{implies} \quad \mathrm{M}, s' \models \varphi \\
\mathrm{M}, s \models \mathbf{G}_i\varphi & \text{iff} & \varphi \in Goals(i)(\mathrm{M}, s) \\
\mathrm{M}, s \models \mathbf{A}_i\pi & \text{iff} & \pi \in Caps(i)(\mathrm{M}, s) \\
\mathrm{M}, s \models \mathbf{Com}_i(\pi) & \text{iff} & \pi \in Agd(i)(\mathrm{M}, s)
\end{array}$$

Dynamic and temporal operators:

$$\begin{array}{lll}
\mathrm{M}, s \models \langle i{:}\pi \rangle\varphi & \text{iff} & \exists(\mathrm{M}', s') \in \mathcal{S} : \mathrm{M}', s' \models \varphi \\
 & & \quad \text{and} \quad ((\mathrm{M}, s), (\mathrm{M}', s')) \in \mathcal{R}_{i:\pi} \\
\mathrm{M}, s \models \mathbf{Fut}\,\varphi & \text{iff} & \exists(\mathrm{M}', s') \in \mathcal{S} : \mathrm{M}', s' \models \varphi
\end{array}$$

$$\begin{array}{lll}
 & & \text{and} \quad ((\mathrm{M}, s), (\mathrm{M}', s')) \in (\bigcup \mathcal{R})^* \\
\mathrm{M}, s \models \mathbf{Past}\, \varphi & \text{iff} & \exists (\mathrm{M}', s') \in \mathcal{S} : \mathrm{M}', s' \models \varphi \\
 & & \text{and} \quad ((\mathrm{M}', s'), (\mathrm{M}, s)) \in (\bigcup \mathcal{R})^* \\
\mathrm{M}, s \models \mathbf{Prev}\, \varphi & \text{iff} & \exists (\mathrm{M}', s') \in \mathcal{S} : \mathrm{M}', s' \models \varphi \\
 & & \text{and} \quad ((\mathrm{M}', s'), (\mathrm{M}, s)) \in (\bigcup \mathcal{R})
\end{array}$$

Appraisal and judgment operators:

$$\begin{array}{lll}
\mathrm{M}, s \models \mathbf{Des}_i(\varphi) & \text{iff} & \varphi \in \mathit{Des}(i)(\mathrm{M}, s) \\
\mathrm{M}, s \models \mathbf{Undes}_i(\varphi) & \text{iff} & \varphi \in \mathit{Undes}(i)(\mathrm{M}, s) \\
\mathrm{M}, s \models \mathbf{Praisew}_i(j{:}\alpha) & \text{iff} & (j, \alpha) \in \mathit{Praisew}(i)(\mathrm{M}, s) \\
\mathrm{M}, s \models \mathbf{Blamew}_i(j{:}\alpha) & \text{iff} & (j, \alpha) \in \mathit{Blamew}(i)(\mathrm{M}, s) \\
\mathrm{M}, s \models \mathbf{Appeal}_i(x) & \text{iff} & x \in \mathit{Appeal}(i)(\mathrm{M}, s) \\
\mathrm{M}, s \models \mathbf{Unappeal}_i(x) & \text{iff} & x \in \mathit{Unappeal}(i)(\mathrm{M}, s) \\
\mathrm{M}, s \models \mathbf{CogUnit}_i(j) & \text{iff} & j \in \mathit{CogUnit}(i)(\mathrm{M}, s)
\end{array}$$

For clarity of presentation the interpretation of $\langle i{:}\pi \rangle \varphi$ is given. The future, past, and previous operators are interpreted over all action-accessibility relations in $\mathcal{R}$, which is done by taking the union $\bigcup \mathcal{R}$. $(\bigcup \mathcal{R})^*$ is then a relation connecting model–state pairs reachable in zero or more actions of agents. Notice that (M, s) and (M', s') are reversed for the future and past operators. As usual, $\models \varphi$ is used to denote that φ is valid, i.e., φ is satisfied in all possible model–state pairs.

5.4. APPRAISAL OPERATORS

Until now we have deferred the problem of specifying appraisal to the functions Des, Undes, $\mathit{Praisew}$, Blamew, Appeal, and $\mathit{Unappeal}$. In this section we have introduced (achievement) goals, which will allow us to give meaning to these functions, but only to a limited degree, because there are many other kinds of concerns, such as norms, interests, preservation, etc. Therefore, we will not *define* these appraisal functions; instead, we will *constrain* them such that they capture appraisal for agents with achievement goals only. The idea is that one can simply add more constraints to these appraisal function if the framework is enriched with more kinds of concerns.

Before introducing these constraints on the appraisal functions, we define two helper sets for matching subparts of goals and inverting goals, respectively:

$$\textsc{sub} = \{\, (\bigwedge \Phi_1, \bigwedge \Phi_2) \mid \Phi_1, \Phi_2 \in \textsc{csl},\ \Phi_1 \subseteq \Phi_2 \,\} \tag{5.10}$$

$$\textsc{inv} = \{\, (\bigwedge \Phi_1, \bigwedge \Phi_2) \mid \Phi_1 \in \textsc{csl},\ \Phi_2 = \{\, \mathit{neg}(\varphi) \mid \varphi \in \Phi_1 \,\} \,\} \tag{5.11}$$

where $neg(p) = \neg p$ and $neg(\neg p) = p$. Recall that we require achievement goals to be consistent conjunctions of literals. The set SUB will then be convenient for making subgoals, and the set INV will be convenient for inverting entire (sub)goals. For example, if $p_1 \wedge \neg p_2$ is a goal, then p_1 and $\neg p_2$ (and $p_1 \wedge \neg p_2$) are subgoals, and $\neg p_1 \wedge p_2$ is the inverted goal.

We will start now with the simplest case of desirability. Let *Des* be constrained such that every subgoal is desirable, and let *Undes* be constrained such that every inverted subgoal is undesirable:

$$\mathit{Des}(i)(\mathrm{M}, s) \supseteq \textsc{sub} \circ \mathit{Goals}(i)(\mathrm{M}, s) \tag{5.12}$$

$$\mathit{Undes}(i)(\mathrm{M}, s) \supseteq \textsc{inv} \circ \textsc{sub} \circ \mathit{Goals}(i)(\mathrm{M}, s) \tag{5.13}$$

where $\circ$ is used to denote relation composition.[12] These two constraints read just as they are written: *Des* contains all subgoals, and *Undes* contains all inverted subgoals. Let us illustrate the above constraint on *Undes*: if $\psi = p \wedge \neg q \wedge r$ is a goal, then the inverted subgoal $\varphi = \neg p \wedge q$ is undesirable, because $(\neg p \wedge q \wedge \neg r, p \wedge \neg q \wedge r) \in \textsc{inv}$ and $(\neg p \wedge q, \neg p \wedge q \wedge \neg r) \in \textsc{sub}$, so $(\neg p \wedge q, p \wedge \neg q \wedge r) \in \textsc{sub} \circ \textsc{inv}$.

The OCC model considers praiseworthiness (and its negative counterpart blameworthiness) to be determined with respect to one's standards. However, OCC note that the praiseworthiness of an action may be evaluated with respect to the desirability of the events caused by that action. Since we do not explicitly consider standards here, we will constrain *Praisew* and *Blamew* using *Des* and *Undes*, respectively. Of course, different or additional constraints may be studied if an explicit representation of standards were added to the logical framework.

We now constrain *Praisew* and *Blamew* as follows. An action of an agent is appraised as praiseworthy or blameworthy when the appraising agent believes that the action is related to a desirable or undesirable consequence, respectively.

$$\begin{aligned}&\mathit{Praisew}(i)(\mathrm{M}, s) \supseteq \\ &\quad \{\, (j, \alpha) \mid \exists \varphi \in \mathit{Des}(i)(\mathrm{M}, s) : \mathrm{M}, s \models \mathbf{B}_i\mathbf{Related}(j{:}\alpha, \varphi) \,\}\end{aligned} \tag{5.14}$$

$$\begin{aligned}&\mathit{Blamew}(i)(\mathrm{M}, s) \supseteq \\ &\quad \{\, (j, \alpha) \mid \exists \varphi \in \mathit{Undes}(i)(\mathrm{M}, s) : \mathrm{M}, s \models \mathbf{B}_i\mathbf{Related}(j{:}\alpha, \varphi) \,\}\end{aligned} \tag{5.15}$$

where $\mathbf{Related}$ is as defined in (4.13). We did not spell out the condition $\mathrm{M}, s \models \mathbf{B}_i\mathbf{Related}(j{:}\alpha, \varphi)$ using the semantics because that would have made these constraints considerably more difficult to read without becoming more enlightening.

[12] In the case where R is binary and Y in unary, the composition $R \circ Y$ is defined as: $R \circ Y = \{\, x \mid \exists y : (x, y) \in R,\ y \in Y \,\}$.

According to the OCC model, appealingness and unappealingness are determined with respect to one's attitudes. Here we will not constrain the functions *Appeal* and *Unappeal* with respect to objects that are not agents. Instead, we will only consider the appealingness of agents, as follows. An agent is appealing to the appraising agent if it has ever performed a praiseworthy action, and unappealing if it has ever performed a blameworthy action:

$$\mathit{Appeal}(i)(\mathrm{M},s) \supseteq \{\, j \mid \exists\alpha : \mathrm{M},s \models \mathbf{Past}\,\mathbf{Approving}_i^{\mathbf{T}}(j,\alpha)\,\} \quad (5.16)$$

$$\mathit{Unappeal}(i)(\mathrm{M},s) \supseteq \{\, j \mid \exists\alpha : \mathrm{M},s \models \mathbf{Past}\,\mathbf{Disapproving}_i^{\mathbf{T}}(j,\alpha)\,\} \quad (5.17)$$

Note that the definitions of 'approving' (3.8) and 'disapproving' (3.9) include the perception of the praiseworthy or blameworthy action. Again, we did not spell out these conditions using the semantics in order to keep the constraints concise and easy to read.

It should be emphasized that we have started out with just achievement goals as the only concerns of agents. We have then defined (or rather, constrained) desirability and undesirability in terms of achievement goals, then defined praiseworthiness and blameworthiness in terms of desirability and undesirability, and then defined appealingness and unappealingness in terms of praiseworthiness and blameworthiness.

5.5. Cognitive Unit

We do not constrain an agent's judgment of when it considers itself to be in a cognitive unit with another agent, except that we require each agent to at least be in a cognitive unit with itself.[13] This is expressed by the following constraint:

$$\mathit{CogUnit}(i)(\mathrm{M},s) \supseteq \{i\} \quad (5.18)$$

This constraint ensures that $\mathbf{Approving}_i^{\mathbf{T}}(i{:}\alpha)$ is equivalent to $\mathbf{Pride}_i^{\mathbf{T}}(i{:}\alpha)$ and that $\mathbf{Disapproving}_i^{\mathbf{T}}(i{:}\alpha)$ is equivalent to $\mathbf{Shame}_i^{\mathbf{T}}(i{:}\alpha)$, as anticipated in section 3 (*cf.* formulas (3.48) and (3.49)).

5.6. Confirmation and Disconfirmation

The only constructs yet undefined are **Confirms** and **Disconfirms**. Recall that $\mathbf{Confirms}_i(\varphi,\psi)$ expresses that agent i considers consequence φ to confirm consequence ψ, and likewise for disconfirmation. Now that we have restricted concerns of agents to achievement goals only and defined achievement goals as conjunctions of literals, representing (dis)confirmation has become quite straightforward.

[13] Even this constraint may be too strong in general, because it may preclude a kind of 'insanity' where one does not consider the self as the (cognitive) author of one's own actions.

For convenience, we first introduce some additional syntax. We will be using the operator $\sqsubseteq$ as the syntactic counterpart of the set SUB. The interpretation of $\sqsubseteq$ is thus as follows:

$$\mathrm{M}, s \models \varphi \sqsubseteq \psi \quad \text{iff} \quad (\varphi, \psi) \in \text{SUB}$$

$\varphi \sqsubseteq \psi$ is then read as "φ is a (logical) part of ψ." Furthermore, we add a syntactical variant of INV. For a conjunction or literals φ, writing $\overline{\varphi}$ in the object language means the 'inverse' of φ in the sense of INV. In other words, for all $(\varphi, \psi) \in$ INV, $\psi = \overline{\varphi}$. Note that INV is symmetric, so that $\overline{\overline{\varphi}} = \varphi$, as expected.

Using these new constructs, we define **Confirms** and **Disconfirms** as follows:

$$\mathbf{Confirms}_i(\varphi, \psi) \overset{\text{def}}{=} \mathbf{B}_i(\varphi \sqsubseteq \psi) \tag{5.19}$$

$$\mathbf{Disconfirms}_i(\varphi, \psi) \overset{\text{def}}{=} \mathbf{B}_i(\overline{\varphi} \sqsubseteq \psi) \tag{5.20}$$

These definitions express that a consequence φ confirms another consequence ψ when φ is a (logical) part of ψ. It is incorrect to require that φ be (logically) stronger than ψ, because the idea of 'confirms' is that it must also account for partial confirmations. For example, suppose Alice learns that a plane carrying four of her relatives has crashed; she will then fear they have all perished but hope for survivors. When later she learns that two of her relatives have survived the crash, this will both partially confirm her fear and partially confirm her hope. To account for partial confirmations, then, we use the construct $\varphi \sqsubseteq \psi$.

It should noted that it is not impossible to define confirmation more generally. For example, "φ (partially) confirms ψ" can also be expressed as "$\psi \models_{\mathrm{CL}} \varphi$," i.e., ψ logically entails φ. But then we would need a construct for representing entailment in the object language. In effect, $\sqsubseteq$ is a kind of entailment relation ($\varphi \sqsubseteq \psi$ implies $\psi \models_{\mathrm{CL}} \varphi$), but restricted to goal formulas (conjunctions of literals).

5.7. Properties

To finish this section, we will discuss some properties of the formalization in KARO of the eliciting conditions of emotions of the OCC model. As usual, $\models \varphi$ expresses that the formula φ is a validity, i.e., every state of every model satisfies φ. All definitions presented in sections 3, 4, and 5 are assumed to be in effect. Proofs of the properties below can be found in Appendix A.

The following properties show how the appraisal operators stem from just (achievement) goals and beliefs. Of course, concerns other than achievement goals can influence desirability, praiseworthiness, and appealingness, but we have restricted our study of appraisal in this paper.

$$\models \mathbf{G}_i\varphi \land \psi \sqsubseteq \varphi \rightarrow \mathbf{Des}_i(\psi) \land \mathbf{Undes}_i(\overline{\psi}) \tag{5.21}$$

$$\models \mathbf{Des}_i(\varphi) \land \mathbf{B}_i\mathbf{Related}(j{:}\alpha, \varphi) \rightarrow \mathbf{Praisew}_i(j{:}\alpha) \tag{5.22}$$

$$\models \mathbf{Undes}_i(\varphi) \land \mathbf{B}_i\mathbf{Related}(j{:}\alpha, \varphi) \rightarrow \mathbf{Blamew}_i(j{:}\alpha) \tag{5.23}$$

$$\models \mathbf{Past\,Approving}^{\mathbf{T}}_i(j{:}\alpha) \rightarrow \mathbf{Appeal}_i(j) \tag{5.24}$$

$$\models \mathbf{Past\,Disapproving}^{\mathbf{T}}_i(j{:}\alpha) \rightarrow \mathbf{Unappeal}_i(j) \tag{5.25}$$

Note that these properties correspond directly to the constraints specified in section 5.4. The notation $\overline{\psi}$ was explained in section 5.6. We emphasize again that desirability and undesirability, praiseworthiness and blameworthiness, and appealingness and unappealingness are not mutually exclusive; see properties (3.50), (3.51), and (3.52) on page 21. Furthermore, desirability and undesirability are not assumed to be individually consistent either. For example, $\mathbf{Des}_i(\varphi) \land \mathbf{Des}_i(\neg\varphi)$ and $\mathbf{Undes}_i(\varphi) \land \mathbf{Undes}_i(\neg\varphi)$ are not contradictions.

The following properties are restatements of (3.42), (3.43), (3.48), and (3.49), respectively.

$$\models \neg(\mathbf{Hope}^{\mathbf{T}}_i(\varphi) \land \mathbf{Joy}^{\mathbf{T}}_i(\varphi)) \tag{5.26}$$

$$\models \neg(\mathbf{Fear}^{\mathbf{T}}_i(\varphi) \land \mathbf{Distress}^{\mathbf{T}}_i(\varphi)) \tag{5.27}$$

$$\models \mathbf{Pride}^{\mathbf{T}}_i(i{:}a) \leftrightarrow \mathbf{Approving}^{\mathbf{T}}_i(i{:}a) \tag{5.28}$$

$$\models \mathbf{Shame}^{\mathbf{T}}_i(i{:}a) \leftrightarrow \mathbf{Disapproving}^{\mathbf{T}}_i(i{:}a) \tag{5.29}$$

Previously, additional assumptions were required to make these propositions derivable. In our formalization in KARO, we have turned these assumptions into constraints, making them truly properties.

In the following, we drop the agent subscripts (e.g., i and j) to ease notation; all terms requiring one are assumed to have the same agent subscript. The four properties below express the triggering conditions for the event-based (and self-based) emotion types in BDI-like terms, i.e., in terms of beliefs and goals.

$$\models \mathbf{G}\varphi \land \psi \sqsubseteq \varphi \land \mathbf{BelUpd}(\psi) \rightarrow \mathbf{Joy}^{\mathbf{T}}(\psi) \tag{5.30}$$

$$\models \mathbf{G}\varphi \land \psi \sqsubseteq \varphi \land \mathbf{BelUpd}(\overline{\psi}) \rightarrow \mathbf{Distress}^{\mathbf{T}}(\overline{\psi}) \tag{5.31}$$

$$\models \mathbf{G}\varphi \land \psi \sqsubseteq \varphi \land \mathbf{BelUpd}(\neg\psi \land \mathbf{Fut}\,\psi) \rightarrow \mathbf{Hope}^{\mathbf{T}}(\psi) \tag{5.32}$$

$$\models \mathbf{G}\varphi \land \psi \sqsubseteq \varphi \land \mathbf{BelUpd}(\neg\overline{\psi} \land \mathbf{Fut}\,\overline{\psi}) \rightarrow \mathbf{Fear}^{\mathbf{T}}(\overline{\psi}) \tag{5.33}$$

The first property states that ‘joy’ is triggered with respect to ψ if ψ is a subgoal of the agent and it has just updated its beliefs with ψ (i.e., subgoal ψ has just been achieved). Analogously, the second property states that ‘distress’ is triggered with respect to an inverted subgoal $\overline{\psi}$ if the subgoal ψ has just been undermined (i.e., part of ψ had previously been achieved but the agent now believes the inverse $\overline{\psi}$ to be true). The third and fourth properties have similar readings, save for being future-directed.

Of course it must be recognized that the current formalization also has its limitations. The OCC model describes emotions with respect to events, actions, and objects, and therefore the elicitation of emotions is naturally described in terms of the perception of events, actions, and objects. However, this emphasis on perception fails to incorporate changes in appraisal of goals, standards, and attitudes as triggers for emotions. For example, the perception of a desired consequence can trigger joy, but a new desire for a known consequence does not trigger joy in the current formalization. Formally, we now have:

$$\models \mathbf{New}\,\mathbf{B}\varphi \wedge \mathbf{Des}(\varphi) \rightarrow \mathbf{Joy}^{\mathbf{T}}(\varphi)$$
$$\not\models \mathbf{B}\varphi \wedge \mathbf{New}\,\mathbf{Des}(\varphi) \rightarrow \mathbf{Joy}^{\mathbf{T}}(\varphi)$$

but the latter should intuitively be valid as well. Of course it is possible to define the emotion triggers such that changes in appraisal are taken into account, but this will be a topic of future work.

The final properties that we will discuss relate intention with tracking of goal achievements and undermining. Given several (reasonable) assumptions, the notion of intention as used in KARO is related with a simultaneous elicitation of 'pride', 'joy', and 'gratification'. Specifically, if an agent (possibly) intends to perform action α to achieve goal φ, then after actually performing α, 'pride' about having done so will be triggered, as well as 'joy' about the achievement, and 'gratification' about the action leading to the achievement.

$$\Gamma \models \mathbf{PossIntend}(\alpha, \varphi) \rightarrow \\ [\alpha](\mathbf{Pride}^{\mathbf{T}}(\alpha) \wedge \mathbf{Joy}^{\mathbf{T}}(\varphi) \wedge \mathbf{Gratification}^{\mathbf{T}}(\alpha, \varphi)) \quad (5.34)$$

where Γ is the following set of assumptions:

- $\mathbf{B}[\alpha]\psi \rightarrow [\alpha]\mathbf{B}\psi$: α is accordant, i.e., the agent does not forget the results of α.
- $\langle\alpha\rangle\psi \rightarrow [\alpha]\psi$: action α is deterministic.
- $\neg(\mathbf{Done}(\alpha) \wedge \mathbf{Done}(\beta))$ for $\alpha \neq \beta$: the agent cannot perform actions in parallel to α.
- $\mathbf{BG}\psi \rightarrow \mathbf{G}\psi$: believed goals are true goals.
- $\mathbf{Prev}\,\mathbf{G}\psi \wedge \neg\mathbf{G}\psi \rightarrow \mathbf{Done}(\mathtt{drop}(\psi))$: only 'drop' actions can remove goals.
- $\neg\mathbf{PossIntend}(\mathtt{drop}(\psi), \psi)$: the agent never intends to achieve a goal by dropping it.

Interestingly, in contrast to 'pure' KARO, the current framework allows us to reason about *sub*goals. This makes it possible to define a less strict version of $\mathbf{PossIntend}$; namely, one expressing that an agent can achieve a subgoal with some action or plan (in contrast to a *complete* goal as required by $\mathbf{PossIntend}$). Analogously, we can define a construct expressing that an agent can undermine a subgoal with some action or plan. These two constructs are defined below as $\mathbf{PossAch}$ and $\mathbf{PossUnd}$, respectively (*cf.* formula (5.5) on page 30).

$$\mathbf{PossAch}_i(\pi, \psi, \varphi) \overset{\text{def}}{=} \mathbf{Can}_i(\pi, \psi) \wedge \mathbf{B}_i(\mathbf{G}_i\varphi \wedge \psi \sqsubseteq \varphi \wedge \overline{\psi}) \tag{5.35}$$

$$\mathbf{PossUnd}_i(\pi, \psi, \varphi) \overset{\text{def}}{=} \mathbf{Can}_i(\pi, \psi) \wedge \mathbf{B}_i(\mathbf{G}_i\varphi \wedge \overline{\psi} \sqsubseteq \varphi \wedge \overline{\psi}) \tag{5.36}$$

$\mathbf{PossAch}_i(\pi, \psi, \varphi)$ is read as "agent i can possibly achieve subgoal ψ of goal φ with plan π," and $\mathbf{PossUnd}_i(\pi, \psi, \varphi)$ is read as "agent i can possibly undermine subgoal $\overline{\psi}$ of goal φ with plan π" (here ψ is thus an inverted subgoal of φ). It may be interesting to note that $\mathbf{PossAch}_i(\pi, \varphi, \varphi)$ implies $\mathbf{PossIntend}_i(\pi, \varphi)$. Using $\mathbf{PossAch}$ we can strengthen property (5.34), and using $\mathbf{PossUnd}$ we can add an analogous case for the simultaneous elicitation of 'shame', 'distress', and 'remorse', as follows.

$$\Gamma \models \mathbf{PossAch}(\alpha, \psi, \varphi) \to \\ [\alpha](\mathbf{Pride}^{\mathbf{T}}(\alpha) \wedge \mathbf{Joy}^{\mathbf{T}}(\psi) \wedge \mathbf{Gratification}^{\mathbf{T}}(\alpha, \psi)) \tag{5.37}$$

$$\Gamma \models \mathbf{PossUnd}(\alpha, \psi, \varphi) \to \\ [\alpha](\mathbf{Shame}^{\mathbf{T}}(\alpha) \wedge \mathbf{Distress}^{\mathbf{T}}(\psi) \wedge \mathbf{Remorse}^{\mathbf{T}}(\alpha, \psi)) \tag{5.38}$$

where Γ is as in property (5.34).

6. Related Work

In this section we discuss several related attempts at adopting psychological models of emotions for modeling artificial agents. We will discuss similarities and differences with the presented approach.

6.1. Previous Work

In previous work, Meyer [22] and Dastani [6] proposed a functional approach to describe the role of four basic emotions in practical reasoning. According to this functional approach, an agent is assumed to execute domain actions in order to reach its goals. The effects of these domain actions cause and/or influence the elicitation of emotions according to a human-inspired model. These emotions in turn influence the deliberation operations of the agent, functioning as heuristics for determining which domain actions have to be chosen next, which completes the circle.

The specification and implementation of emotions carried out by Meyer [22] and Dastani [6] follows Oatley & Jenkins' model of emotions [24]. In contrast to our approach of capturing a broad and complete[14] range of emotion types, they consider only four emotions: *happy*, *sad*, *angry*, and *fearful*. Each emotion functions as a *label* of an aspect of an agent's cognitive state. The deliberation of an agent then behaves in accordance with heuristics associated with these four emotions. Later we have extended this approach by showing how interaction between *hope* and *fear* can influence an agent's deliberation [33].

In other previous work, we have shown how emotional experience can be modeled on top of a formalization of emotion triggers [34, 32]. We have then also shown how emotion regulation can be modeled on top of a formalization of emotional experience [35, 32]. This was done by introducing a notion of *action tendency* into the formalization, which indicates which action(s) an agent tends to perform to mitigate negative emotional experience. The present paper, however, is our first complete presentation of our formalization of the eliciting conditions of the emotion types of the OCC model.

6.2. Another Formalization of the OCC Model

The construction of a complete formalization of the OCC model in agent logic has previously been attempted by Adam, Herzig & Longin [1].[15] Our approach is similar to Adam's formalization in the sense that both use BDI-like logics (belief, desire, intention) to formalize the emotions of the OCC model and that both approaches are based on modal logic. Below we will briefly discuss major differences between the presented formalization of the OCC model and the one by Adam.

Just like us, Adam aims to be "as faithful as possible" to the OCC model. However, Adam's formalization of OCC's emotion types has been tailored to their BDI-based logical framework. In contrast, our formalization proceeds in three stages, where the first stage captures the logical structure of the OCC model and only the last stage commits to BDI. Furthermore, Adam's logical framework incorporates several very strong assumptions. For example, desires and ideals are assumed never to change and to be free of contradictions (thus excluding many forms of 'mixed feelings'); agents are assumed to have complete introspection with respect to their desires; and all actions are assumed to be deterministic, public, and accordant (i.e. no forgetting of effects). By refraining from making such assumptions, we believe our formalization

[14] That is, complete with respect to one psychological model of emotions, namely the OCC model.

[15] In the following, we simply use "Adam" to refer to Adam *et al.* [1].

is able to account for more situations in which emotions can arise (according to psychology).

Some of Adam's definitions of emotions do not capture all aspects of what is supposed to be formalized. For example, Adam's formalization of hope and fear does not account for future-directed prospects (only current uncertainty); 'easy' actions preclude pride and shame; and partial (dis)confirmations cannot trigger satisfaction, fears-confirmed, relief, or disappointment. Admittedly, the OCC model may be implicit or ambiguous with respect to these and other aspects, but ideally, the process of formalization should explicate such issues and offer clarifications.

There is some confusion in Adam's formalization between emotion elicitation and experience. Adam claims to formalize the eliciting conditions of emotions (as do we), and the action-based emotions indeed appear to incorporate a trigger, namely in the form of the perception of an action. However, Adam's formalizations of the event-based emotions do not incorporate any triggers. For example, joy is defined as $Joy_i\varphi \overset{\text{def}}{=} Bel_i\varphi \wedge Des_i\varphi$, but this expresses a 'state of joy' more than a trigger for joy. (Also note that this definition does not force φ to represent a consequence of an event.) Indeed, in the text Adam often identifies the satisfaction of an emotion formula with feeling the emotion in question. When Adam defines the compound emotions simply as conjunctions (e.g., $\mathit{Gratification}_i(i{:}\alpha, \varphi) \overset{\text{def}}{=} Pride_i(i{:}\alpha, \varphi) \wedge Joy_i\varphi$), it is then unclear what $\mathit{Gratification}_i(i{:}\alpha, \varphi)$ actually represents because it mixes triggering ($Pride_i(i{:}\alpha, \varphi)$) and experience ($Joy_i\varphi$). In our approach, we have made a clear distinction between emotion elicitation (treated in this paper) and experience (treated in [32]) in order to avoid such confusion.

Finally, Adam's formalization renders a number of properties of emotions that we find too strong. For example, Adam proves that $\vdash \neg(Joy_i\varphi \wedge Distress_i\varphi)$ and similarly for all pairs of opposing emotions applied to the same argument(s). Such formulas are not valid in our formalization because we allow goals, standards, and attitudes to be inconsistent. However, if their consistency would be adopted as a constraint, it would indeed be provable in our framework that opposing emotion triggers contradict. Furthermore, Adam derives complete introspection of emotions (i.e. $\vdash Emotion_i\varphi \leftrightarrow Bel_iEmotion_i\varphi$ and $\vdash \neg Emotion_i\varphi \leftrightarrow Bel_i\neg Emotion_i\varphi$). However, if $Emotion_i\varphi$ is supposed to be a formalization of the eliciting conditions of $Emotion$, as Adam intends, then we find this unintuitive; one does not have to be aware of what triggered an emotion. On the other hand, it is intuitive to suppose that one is aware of what one does and does not feel, i.e., if $Emotion_i\varphi$ were to represent the subjective experience of $Emotion$. Again, because of confusion between elicitation and experience in Adam's formalization, it is difficult to judge the status of these introspection properties.

6.3. A Computational Model of Emotions

Gratch and Marsella [13] have been working on a computational framework for modeling emotions. The framework is claimed to be domain-independent and they have implemented a process model, called EMA after the title of [19], for social training applications. The appraisal process used in EMA is inspired by the OCC model. As with our approach, the cognitive reasoning aspects of EMA are represented using BDI concepts and the emphasis of appraisal is on goal attainment.

In contrast to our approach, Gratch and Marsella take a computational, quantitative approach towards modeling appraisal. Specifically, the eliciting conditions of emotions modeled in EMA are based on quantitative measures of, e.g., desirability and likelihood. The calculation of these quantitative measures is facilitated by the usage of subjective probabilities for beliefs and assignment of utilities to states. However, precise triggering conditions for all emotions are not provided, so it is hard to judge how strictly Gratch and Marsella follow psychological models of emotions and how they deviate from or extend these. For example, in line with the OCC model, likelihood of a desirable event is given as a precondition for hope. However, in EMA likelihood of an event is equated with the believed probability of the event, such that likelihood can also be used as a precondition for joy (in particular, if the likelihood of the event equals one). But like Adam's notion of expectation, such a definition of likelihood models uncertainty about the current state but not prospects about *future* events.

Concerning the three main topics of modeling emotions discussed in the introduction (i.e. appraisal, experience, and regulation), Gratch and Marsella distinguish only between appraisal and regulation. Probably owing to their computational, quantitative approach, emotional experience appears to be merged with appraisal in EMA. Their main focus is on modeling how emotions influence behavior, emphasizing modeling of coping strategies for artificial agents. Behavioral effects of emotions have not been treated in this paper, but a detailed analysis using the presented framework can be found in [32].

7. Conclusions

In this paper we have studied the OCC model [27] of emotions and proceeded with the formalization of the eliciting conditions of emotions according to this psychological model. We have carried out this formalization in three stages. First, we have captured OCC's specifications of eliciting conditions in a semiformal manner, thus without committing to a particular formalism and semantics. Second, we have shown how OCC's main notions of events, consequences, actions, agents, and beliefs can be captured in dynamic doxastic logic. Third, we have represented OCC's main appraisal notions in the KARO framework, which is a BDI-based extension of dynamic doxastic logic, thereby firmly grounding the preceding two stages. It should be noted that the collection of emotion triggers that are satisfied in a certain state should *not* be regarded as a representation of an agent's full emotional state. Rather, it only represents which *new* emotions are triggered in that state, without specifying whether these newly triggered emotions are (or will ever be) experienced.

The idea is that the emotion triggers function as input for an additional quantitative model of emotions. Such a quantitative model should specify how triggered emotions are *experienced*. Elsewhere [34, 32], we have explained how emotional experience can be modeled on top of qualitative models of emotion triggers similar to the one presented in this paper. This was done by introducing functions representing different parameters of emotional experience. By setting thresholds for these parameters, different emotion words of the same emotion type can be modeled. For example, 'annoyed', 'livid', and 'outraged' can each be represented in the logic as different emotions of the type 'anger'.

Finally, it should be specified *what to do* with experienced emotions. Previous work on formalizing the behavioral effects of emotions either used a subset of the emotions as described in psychological model of Oatley & Jenkins [22, 6], or a subset of the OCC model [33, 35]. We are currently continuing in the line of [35] to formally specify emotion-based action tendency for all emotion types presented in this paper. Unfortunately, psychological literature on emotion regulation (i.e. the effect of elicited emotions on behavior) does not provide classifications and schemes as clear as those on appraisal, such as the OCC model. (For an overview of psychological research on the subject, we refer the reader to Gross [14].) Thus to proceed with formalizing in the direction of specifying the effects of emotions on the behavior and decision making of agents will require more creativity on the part of logicians and computer scientists. It should be noted that we have not included details on how to integrate experience and regulation in the presented formalization of emotions in order to limit the length of this paper. However, research in this direction can be found in other work [32, 35, 36].

Our future work on emotions is twofold. On the one hand, we are continuing our work on refining the formalization of quantitative aspects of emotions and specifying their effects on behavior and decision making. On the other hand, we are working on an implementation of this formal model of emotions on top of the interpreter of an agent programming language. This way we expect to validate the added value of emotions on the decision making and believability of artificial agents.

References

1. Carole Adam, Andreas Herzig, and Dominique Longin. A logical formalization of the OCC theory of emotions. *Synthese*, 168(2):201–248, 2009.
2. Cynthia L. Breazeal. *Designing Sociable Robots*. MIT Press, 2002.
3. Philip R. Cohen and Hector J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42:213–261, 1990.
4. Gilles Coppin. Emotion, personality and decision-making: Relying on the observables. In *Proceedings of the Third International Conference in Human Centered Processes (HCP-2008)*, 2008.
5. Antonio R. Damasio. *Descartes' Error: Emotion, Reason and the Human Brain*. Grosset/Putnam, New York, 1994.
6. Mehdi Dastani and John-Jules Ch. Meyer. Programming agents with emotions. In *Proceedings of the 17th European Conference on Artificial Intelligence (ECAI'06)*, pages 215–219, 2006.
7. Paul Ekman and Richard J. Davidson, editors. *The Nature of Emotion: Fundamental Questions*. Series in Affective Science. Oxford University Press, 1994.
8. Jon Elster. Rationality and the emotions. *Economic Journal*, 106(438):1386–1397, 1996.
9. E.A. Emerson. Temporal and modal logic. In J. van Leeuwen, editor, *Handbook of Theoretical Computer Science*, chapter 16, pages 995–1072. Elsevier/MIT Press, Amsterdam/Cambridge, MA, 1990.
10. Nico H. Frijda. *The Emotions*. Studies in Emotion and Social Interaction. Cambridge University Press, 1987.
11. Nico H. Frijda, Andrew Ortony, Joep Sonnemans, and Gerald L. Clore. The complexity of intensity: Issues concerning the structure of emotion intensity. In Margaret S. Clark, editor, *Emotion, Review of Personality and Social Psychology*, volume 13, pages 60–89. SAGE Publications, London, 1992.
12. Robert M. Gordon. *The Structure of Emotions: Investigations in Cognitive Philosophy*. Cambridge University Press, 1987.
13. Jonathan Gratch and Stacy Marsella. A domain-independent framework for modeling emotions. *Journal of Cognitive Systems Research*, 5(4):269–306, 2004.
14. James J. Gross and Ross A. Thompson. Emotion regulation: Conceptual foundations. In James J. Gross, editor, *Handbook of Emotion Regulation*. Guilford Press, 2007.
15. David Harel, Dexter Kozen, and Jerzy Tiuryn. *Dynamic Logic*. MIT Press, Cambridge, MA, 2000.
16. W. van der Hoek, Bernd van Linder, and John-Jules Ch. Meyer. An integrated modal approach to rational agents. In Michael J. Wooldridge and Anand S. Rao, editors, *Foundations of Rational Agency*, volume 14 of *Applied Logic Series*, pages 133–168. Kluwer, Dordrecht, 1999.
17. Robert R. Hoffman, John E. Waggoner, and David S. Palermo. Metaphor and context in the language of emotion. In Robert R. Hoffman and David S. Palermo, editors, *Cognition and the Symbolic Processes: Applied and Ecological Perspectives*, pages 163–185. Lawrence Erlbaum Associates, Hillsdale, NJ, 1991.
18. Michael Johns and Barry G. Silverman. How emotion and personality effect the utility of alternative decisions: A terrorist target selection case study. In *10th Conference On Computer Generated Forces and Behavioral Representation, SISO*, 2001.
19. Richard S. Lazarus. *Emotion and Adaptation*. Oxford University Press, 1994.
20. Joseph E. LeDoux. *The Emotional Brain: Mysterious Underpinnings of Emotional Life*. Simon & Schuster, 1996.

21. Robert P. Marinier and John E. Laird. Toward a comprehensive computational model of emotions and feelings. In *Proceedings of the International Conference on Cognitive Modeling (ICCM'04)*, pages 172–177, Pittsburgh, PA, 2004.
22. John-Jules Ch. Meyer. Reasoning about emotional agents. *International Journal of Intelligent Systems*, 21(6):601–619, 2006.
23. John-Jules Ch. Meyer, Wiebe van der Hoek, and Bernd van Linder. A logical approach to the dynamics of commitments. *Artificial Intelligence*, 113:1–40, 1999.
24. Keith Oatley and Jennifer M. Jenkins. *Understanding Emotions*. Blackwell Publishing, Oxford, UK, 1996.
25. Andrew Ortony, October 2009. Personal communication.
26. Andrew Ortony and Gerald L. Clore, April–June 2009. Personal communication.
27. Andrew Ortony, Gerald L. Clore, and Allan Collins. *The Cognitive Structure of Emotions*. Cambridge University Press, Cambridge, UK, 1988.
28. Rosalind W. Picard. *Affective Computing*. MIT Press, 1997.
29. Anand S. Rao and Michael P. Georgeff. Decision procedures for BDI logics. *Journal of Logic and Computation*, 8(3):293–344, 1998.
30. Klaus R. Scherer, Angela Schorr, and Tom Johnstone, editors. *Appraisal Processes in Emotion: Theory, Methods, Research*. Series in Affective Science. Oxford University Press, 2001.
31. Aaron Sloman. Beyond shallow models of emotion. *Cognitive Processing*, 2(1):177–198, 2001.
32. Bas R. Steunebrink. *The Logical Structure of Emotions*. PhD thesis, Utrecht University, Utrecht, The Netherlands, 2010.
33. Bas R. Steunebrink, Mehdi Dastani, and John-Jules Ch. Meyer. A logic of emotions for intelligent agents. In *Proceedings of the Twenty-Second Conference on Artificial Intelligence (AAAI'07)*. AAAI Press, 2007.
34. Bas R. Steunebrink, Mehdi Dastani, and John-Jules Ch. Meyer. A formal model of emotions: Integrating qualitative and quantitative aspects. In Ghallab Mali, Constantine D. Spyropoulos, Nikos Fakotakis, and Nikos Avouris, editors, *Proceedings of the 18th European Conference on Artificial Intelligence (ECAI'08)*, pages 256–260. IOS Press, 2008.
35. Bas R. Steunebrink, Mehdi Dastani, and John-Jules Ch. Meyer. A formal model of emotion-based action tendency for intelligent agents. In *Proceedings of the 14th Portuguese Conference on Artificial Intelligence (EPIA'09)*. Springer, 2009.
36. Bas R. Steunebrink, Mehdi Dastani, and John-Jules Ch. Meyer. Emotions to control agent deliberation. In van der Hoek, Kaminka, Lesprance, Luck, and Sen, editors, *Proceedings of the Ninth International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2010)*, 2010.

Appendix

A. Proofs of Propositions

All proofs use the semantics given on page 32.

Proposition (4.23). Take an arbitrary agent i and action α. Now for all $((\mathrm{M}, s), (\mathrm{M}', s')) \in \mathcal{R}_{i:\alpha}$ we obviously have that $((\mathrm{M}', s'), (\mathrm{M}, s)) \in \mathcal{R}_{i:\alpha^-}$ and $\mathrm{M}, s \models \top$, i.e., $\mathrm{M}', s' \models \langle i{:}\alpha^- \rangle \top$, which is the same as $\mathrm{M}', s' \models \mathbf{Done}(i{:}\alpha)$. Because (M', s') was an arbitrary $\mathcal{R}_{i:\alpha}$-successor of (M, s), we have that $\mathrm{M}, s \models [i{:}\alpha]\mathbf{Done}(i{:}\alpha)$. Because (M, s) was arbitrary, $[i{:}\alpha]\mathbf{Done}(i{:}\alpha)$ is valid. □

Proposition (4.24)**.** Assume $\mathrm{M}, s \models \mathbf{Done}(i{:}\alpha)$ for arbitrary (M, s), i, α. Then there exists a model–state pair (M', s') such that $((\mathrm{M}', s'), (\mathrm{M}, s)) \in \mathcal{R}_{i:\alpha}$. Because $\mathcal{R}_{i:\alpha} \subseteq \bigcup \mathcal{R}$, $((\mathrm{M}', s'), (\mathrm{M}, s)) \in (\bigcup \mathcal{R})$. Because $\mathrm{M}', s' \models \top$, we have that $\mathrm{M}, s \models \mathbf{Prev}\,\top$. Because (M, s) was arbitrary, $\mathbf{Done}(i{:}\alpha) \rightarrow \mathbf{Prev}\,\top$ is valid. □

Proposition (4.25)**.** Assume $\mathrm{M}, s \models \mathbf{Prev}\,\top$ for arbitrary (M, s). This implies that $\exists(\mathrm{M}', s') : ((\mathrm{M}', s'), (\mathrm{M}, s)) \in (\bigcup \mathcal{R})$. But $\bigcup \mathcal{R}$ is constrained to be injective, so $\exists(\mathrm{M}', s') : ((\mathrm{M}', s'), (\mathrm{M}, s)) \in (\bigcup \mathcal{R})$ and $\mathrm{M}', s' \models \varphi$ is true if and only if $\forall(\mathrm{M}', s') : ((\mathrm{M}', s'), (\mathrm{M}, s)) \in (\bigcup \mathcal{R})$ implies $\mathrm{M}', s' \models \varphi$ is true, i.e., $\mathrm{M}, s \models \mathbf{Prev}\,\varphi \leftrightarrow \neg\mathbf{Prev}\,\neg\varphi$. Because (M, s) was arbitrary, $\mathbf{Prev}\,\top \rightarrow (\mathbf{Prev}\,\varphi \leftrightarrow \neg\mathbf{Prev}\,\neg\varphi)$ is valid. And because $\mathbf{Prev}\,\varphi$ implies $\mathbf{Prev}\,\top$, this proposition can be rewritten as $\mathbf{Prev}\,\varphi \leftrightarrow \neg\mathbf{Prev}\,\neg\varphi \wedge \mathbf{Prev}\,\top$. □

Proposition (4.26)**.** If $\neg\mathbf{Prev}\,\top$ holds, then for any φ, $\neg\mathbf{Prev}\,\varphi$ holds. $\mathbf{New}\,\varphi$ is defined as $\varphi \wedge \neg\mathbf{Prev}\,\varphi$, so $\varphi \wedge \neg\mathbf{Prev}\,\top \rightarrow \mathbf{New}\,\varphi$ is valid. □

Proposition (4.27)**.** Assume $\mathrm{M}, s \models \mathbf{Done}(i{:}\alpha)$ and $\mathrm{M}, s \models \varphi$ for arbitrary (M, s), i, α, φ. Let (M', s') be the model–state pair such that $((\mathrm{M}', s'), (\mathrm{M}, s)) \in \mathcal{R}_{i:\alpha}$ (there can only be one such (M', s') because $\bigcup \mathcal{R}$ is injective). Now $\mathrm{M}', s' \models \langle i{:}\alpha\rangle\varphi$ and therefore $\mathrm{M}, s \models \mathbf{Prev}\,\langle i{:}\alpha\rangle\varphi$. Because (M, s) was arbitrary, $\mathbf{Done}(i{:}\alpha) \wedge \varphi \rightarrow \mathbf{Prev}\,\langle i{:}\alpha\rangle\varphi$ is valid. □

Proposition (4.28)**.** Assume $\mathrm{M}, s \models \mathbf{Prev}\,[i{:}\alpha]\varphi$ and $\mathrm{M}, s \models \mathbf{Done}(i{:}\alpha)$ for arbitrary (M, s), i, α, φ. Let (M', s') be the model–state pair such that $((\mathrm{M}', s'), (\mathrm{M}, s)) \in \mathcal{R}_{i:\alpha}$ (there can only be one such (M', s') because $\bigcup \mathcal{R}$ is injective). Now $\mathrm{M}', s' \models [i{:}\alpha]\varphi$ and therefore $\mathrm{M}, s \models \varphi$. Because (M, s) was arbitrary, we have that $\mathbf{Prev}\,[i{:}\alpha]\varphi \wedge \mathbf{Done}(i{:}\alpha) \rightarrow \varphi$ is valid. □

Proposition (4.29)**.** The expression $\mathrm{M}, s \models \mathbf{Fut}\,\varphi$ is interpreted as $\exists(\mathrm{M}', s') \in \mathcal{S} : ((\mathrm{M}, s), (\mathrm{M}', s')) \in (\bigcup \mathcal{R})^*$ and $\mathrm{M}', s' \models \varphi$. This is the same as $\mathrm{M}, s \models \varphi$ or $\exists n \in \mathbb{N} : \exists i_0, \ldots, i_n \in \text{AGT} : \exists \alpha_0, \ldots, \alpha_n \in \text{ACT} : \exists((\mathrm{M}, s), (\mathrm{M}', s')) \in \mathcal{R}_{i_0:\alpha_0} \circ \ldots \circ \mathcal{R}_{i_n:\alpha_n} : \mathrm{M}', s' \models \varphi$, i.e., $\mathrm{M}, s \models \varphi \vee \langle i_1{:}\alpha_1\rangle \cdots \langle i_n{:}\alpha_n\rangle\varphi$. So $\mathbf{Fut}\,\varphi \leftrightarrow \varphi \vee \exists i_0, \ldots, i_n \exists \alpha_0, \ldots, \alpha_n (\langle i_1{:}\alpha_1\rangle \cdots \langle i_n{:}\alpha_n\rangle\varphi)$ is valid (although strictly speaking we do not have quantification in our object language). □

Proposition (5.21)–(5.25)**.** These propositions follow immediately from constraints (5.12)–(5.17). □

Proposition (5.26) **and** (5.27)**.** These propositions follow immediately from propositions (4.19) and (4.20) and the fact that $\neg(\mathbf{B}_i\varphi \wedge \mathbf{B}_i\neg\varphi)$ follows from the seriality of R_i. □

Proposition (5.28) **and** (5.29)**.** These propositions follow immediately from propositions (3.48) and (3.49) and constraint (5.18). □

Proposition (5.30)–(5.33)**.** $\mathbf{Joy}_i^{\mathbf{T}}(\psi)$ is equivalent to $\mathbf{Des}_i(\psi) \wedge \mathbf{Actual}_i(\psi)$ and $\mathbf{Distress}_i^{\mathbf{T}}(\overline{\psi})$ is equivalent to $\mathbf{Undes}_i(\overline{\psi}) \wedge \mathbf{Actual}_i(\overline{\psi})$. By proposition (5.21), $\mathbf{G}_i\varphi \wedge \psi \sqsubseteq \varphi$ implies $\mathbf{Des}_i(\psi)$ and $\mathbf{Undes}_i(\overline{\psi})$. Because $\mathbf{Actual}_i(\psi)$ is equivalent to $\mathbf{BelUpd}_i(\psi)$, $\mathbf{G}_i\varphi \wedge \psi \sqsubseteq \varphi \wedge \mathbf{BelUpd}_i(\psi)$ implies $\mathbf{Joy}_i^{\mathbf{T}}(\psi)$ and $\mathbf{G}_i\varphi \wedge \psi \sqsubseteq \varphi \wedge \mathbf{BelUpd}_i(\overline{\psi})$ implies $\mathbf{Distress}_i^{\mathbf{T}}(\overline{\psi})$. Propositions (5.32) and (5.33) follow analogously, by noting that $\mathbf{BelUpd}_i(\mathbf{Fut}^+\psi)$ is equivalent to $\mathbf{FutUpd}_i(\psi)$, which implies $\mathbf{Prospective}_i(\psi)$. □

Proposition (5.34)**.** Assume $\mathrm{M}, s \models \mathbf{PossIntend}_i(\alpha, \varphi)$ for arbitrary (M, s), i, α, φ. Take an arbitrary model–state pair (M', s') such that $((\mathrm{M}, s), (\mathrm{M}', s')) \in \mathcal{R}_{i:\alpha}$. To prove: $\mathrm{M}', s' \models \mathbf{Pride}_i^{\mathbf{T}}(i{:}\alpha) \wedge \mathbf{Joy}_i^{\mathbf{T}}(\varphi) \wedge \mathbf{Gratification}_i^{\mathbf{T}}(i{:}\alpha, \varphi)$, i.e., $\mathrm{M}', s' \models \mathbf{PerceiveAction}_i(i{:}\alpha) \wedge \mathbf{Praisew}_i(i{:}\alpha) \wedge \mathbf{CogUnit}_i(i) \wedge \mathbf{PerceiveConseq}_i(\varphi) \wedge \mathbf{Actual}_i(\varphi) \wedge \mathbf{Des}_i(\varphi) \wedge \mathbf{Past}\,\mathbf{Pride}_i^{\mathbf{T}}(i{:}\alpha) \wedge \mathbf{Past}\,\mathbf{Joy}_i^{\mathbf{T}}(\varphi) \wedge \mathbf{PerceiveRelated}_i(i{:}\alpha, \varphi)$. We can immediately cross out several of the conjuncts: $\mathbf{CogUnit}_i(i)$ follows directly from constraint (5.18); by definition (4.10), $\mathbf{PerceiveConseq}_i(\varphi)$ follows from $\mathbf{Actual}_i(\varphi)$; and $\mathbf{Past}\,\mathbf{Pride}_i^{\mathbf{T}}(i{:}\alpha)$ and $\mathbf{Past}\,\mathbf{Joy}_i^{\mathbf{T}}(\varphi)$ will follow automatically because $\psi \to \mathbf{Past}\,\psi$ is a validity. Writing out definitions, we have to prove $\mathrm{M}', s' \models \mathbf{New}\,\mathbf{B}_i\mathbf{Past}\,\mathbf{Done}(i{:}\alpha) \wedge \mathbf{New}\,\mathbf{B}_i\varphi \wedge \mathbf{New}\,\mathbf{B}_i\mathbf{Related}(i{:}\alpha, \varphi) \wedge \mathbf{Des}_i(\varphi) \wedge \mathbf{Praisew}_i(i{:}\alpha)$. If $\mathrm{M}, s \models \neg\psi$ and $\mathrm{M}', s' \models \psi$, then $\mathrm{M}', s' \models \mathbf{New}\,\psi$. So for state (M, s) we have to show:

(i) $\mathrm{M}, s \models \neg\mathbf{B}_i\mathbf{Past}\,\mathbf{Done}(i{:}\alpha)$
(ii) $\mathrm{M}, s \models \neg\mathbf{B}_i\varphi$
(iii) $\mathrm{M}, s \models \neg\mathbf{B}_i\mathbf{Related}(i{:}\alpha, \varphi)$

and for state (M', s') we have to show:

(a) $\mathrm{M}', s' \models \mathbf{B}_i\mathbf{Past}\,\mathbf{Done}(i{:}\alpha)$
(b) $\mathrm{M}', s' \models \mathbf{B}_i\varphi$
(c) $\mathrm{M}', s' \models \mathbf{B}_i\mathbf{Related}(i{:}\alpha, \varphi)$
(d) $\mathrm{M}', s' \models \mathbf{Des}_i(\varphi)$
(e) $\mathrm{M}', s' \models \mathbf{Praisew}_i(i{:}\alpha)$

By proposition (5.22), (e) is implied by (d) and (c). Because $\psi \to \mathbf{Past}\,\psi$ is valid, so is $\mathbf{B}_i\psi \to \mathbf{B}_i\mathbf{Past}\,\psi$, which means that it suffices to show for (M', s') that:

(A) $\mathrm{M}', s' \models \mathbf{B}_i\mathbf{Done}(i{:}\alpha)$
(B) $\mathrm{M}', s' \models \mathbf{B}_i\varphi$
(C) $\mathrm{M}', s' \models \mathbf{B}_i\neg\mathbf{Prev}\,\varphi$
(D) $\mathrm{M}', s' \models \mathbf{Des}_i(\varphi)$

From proposition (4.23), we have that $[i{:}\alpha]\mathbf{Done}(i{:}\alpha)$ is valid. Then by necessitation $\mathbf{B}_i[i{:}\alpha]\mathbf{Done}(i{:}\alpha)$ is also valid, and by the assumption that $\mathbf{B}_i[i{:}\alpha]\psi \rightarrow [i{:}\alpha]\mathbf{B}_i\psi$, $[i{:}\alpha]\mathbf{B}_i\mathbf{Done}(i{:}\alpha)$ is also valid. But if $\mathrm{M}, s \models [i{:}\alpha]\mathbf{B}_i\mathbf{Done}(i{:}\alpha)$, then $\mathrm{M}', s' \models \mathbf{B}_i\mathbf{Done}(i{:}\alpha)$, which proves (A). $\mathrm{M}, s \models \mathbf{PossIntend}_i(\alpha, \varphi)$ implies $\mathrm{M}, s \models \mathbf{B}_i\langle i{:}\alpha\rangle\varphi$, which implies $\mathrm{M}, s \models \mathbf{B}_i[i{:}\alpha]\varphi$ (because α was assumed to be deterministic), which implies $\mathrm{M}, s \models [i{:}\alpha]\mathbf{B}_i\varphi$ (because α was assumed to be accordant). But then $\mathrm{M}', s' \models \mathbf{B}_i\varphi$, which proves (B). It is easy to verify that $\neg\varphi \rightarrow [i{:}\alpha]\neg\mathbf{Prev}\,\varphi$ is a validity; then $\mathbf{B}_i\neg\varphi \rightarrow \mathbf{B}_i[i{:}\alpha]\neg\mathbf{Prev}\,\varphi$ is also valid. Because $\mathbf{PossIntend}_i(\alpha, \varphi)$ implies $\mathbf{B}_i\neg\varphi$, $\mathrm{M}, s \models \mathbf{B}_i[i{:}\alpha]\neg\mathbf{Prev}\,\varphi$. But then $\mathrm{M}, s \models [i{:}\alpha]\mathbf{B}_i\neg\mathbf{Prev}\,\varphi$ and $\mathrm{M}', s' \models \mathbf{B}_i\neg\mathbf{Prev}\,\varphi$, which proves (C). $\mathbf{PossIntend}_i(\alpha, \varphi)$ implies $\mathbf{B}_i\mathbf{G}_i\varphi$, which implies $\mathbf{G}_i\varphi$ (because it was assumed that $\mathbf{B}_i\mathbf{G}_i\varphi \rightarrow \mathbf{G}_i\varphi$). So because $\mathrm{M}, s \models \mathbf{G}_i\varphi$, $\mathrm{M}', s' \models \mathbf{Prev}\,\mathbf{G}_i\varphi$. Furthermore, because is was assumed that $\neg\mathbf{PossIntend}_i(\mathtt{drop}(\varphi), \varphi)$ is valid, $\alpha \neq \mathtt{drop}(\varphi)$. It was assumed that $\neg(\mathbf{Done}(i{:}\alpha) \wedge \mathbf{Done}(i{:}\mathtt{drop}(\varphi)))$, so the fact that $\mathrm{M}', s' \models \mathbf{Done}(i{:}\alpha)$ implies $\mathrm{M}', s' \models \neg\mathbf{Done}(i{:}\mathtt{drop}(\varphi))$. But it was also assumed that $\mathbf{Prev}\,\mathbf{G}_i\varphi \wedge \neg\mathbf{G}_i\varphi \rightarrow \mathbf{Done}(i{:}\mathtt{drop}(\varphi))$, so it must be that $\mathrm{M}', s' \models \mathbf{G}_i\varphi$, which implies $\mathrm{M}', s' \models \mathbf{Des}_i(\varphi)$, which proves (D). It is easy to verify that the requirement of unique actions (see constraint (5.6)) validates $\mathbf{Past}\,\mathbf{Done}(i{:}\alpha) \rightarrow [i{:}\alpha]\bot$, i.e., $\langle i{:}\alpha\rangle\top \rightarrow \neg\mathbf{Past}\,\mathbf{Done}(i{:}\alpha)$. Then $\mathbf{B}_i\langle i{:}\alpha\rangle\top \rightarrow \mathbf{B}_i\neg\mathbf{Past}\,\mathbf{Done}(i{:}\alpha)$ is also valid. The antecedent is implied by $\mathbf{PossIntend}_i(\alpha, \varphi)$, so $\mathrm{M}, s \models \mathbf{B}_i\neg\mathbf{Past}\,\mathbf{Done}(i{:}\alpha)$. But by seriality of R_i, $\mathrm{M}, s \models \neg\mathbf{B}_i\mathbf{Past}\,\mathbf{Done}(i{:}\alpha)$, which proves (i). $\mathbf{PossIntend}_i(\alpha, \varphi)$ implies $\mathbf{B}_i\neg\varphi$, so by seriality of R_i, $\mathrm{M}, s \models \neg\mathbf{B}_i\varphi$, which proves (ii). It is easy to verify that $\mathbf{Past}\,(\mathbf{Done}(i{:}\alpha) \wedge \mathbf{New}\,\varphi) \rightarrow \mathbf{Past}\,\mathbf{Done}(i{:}\alpha)$ is a validity. But then $\neg\mathbf{B}_i\mathbf{Past}\,\mathbf{Done}(i{:}\alpha) \rightarrow \neg\mathbf{B}_i\mathbf{Past}\,(\mathbf{Done}(i{:}\alpha) \wedge \mathbf{New}\,\varphi)$ is also a validity. So (i) implies (iii), which proves (iii). We can now conclude that indeed $\mathrm{M}', s' \models \mathbf{Pride}_i^{\mathbf{T}}(i{:}\alpha) \wedge \mathbf{Joy}_i^{\mathbf{T}}(\varphi) \wedge \mathbf{Gratification}_i^{\mathbf{T}}(i{:}\alpha, \varphi)$. Because (M', s') and (M, s) were arbitrary, $\mathbf{PossIntend}_i(\alpha, \varphi) \rightarrow [i{:}\alpha](\mathbf{Pride}_i^{\mathbf{T}}(i{:}\alpha) \wedge \mathbf{Joy}_i^{\mathbf{T}}(\varphi) \wedge \mathbf{Gratification}_i^{\mathbf{T}}(i{:}\alpha, \varphi))$ is valid. □

Proposition (5.37) **and** (5.38)**.** The proofs of these propositions are largely the same as the proof of proposition (5.34) above. For example, assuming $\mathbf{PossAch}_i(\alpha, \psi, \varphi)$ or $\mathbf{PossUnd}_i(\alpha, \psi, \varphi)$ still implies point (ii), because we then have that $\mathrm{M}, s \models \mathbf{B}_i\overline{\psi}$, that $\overline{\psi} \rightarrow \neg\psi$ is valid, that R_i is serial, and thus that $\mathrm{M}, s \models \neg\mathbf{B}_i\psi$. To prove proposition (5.38), points (d) and (e) become $\mathrm{M}', s' \models \mathbf{Undes}_i(\psi)$ and $\mathrm{M}', s' \models \mathbf{Blamew}_i(i{:}\alpha)$, respectively. $\mathbf{PossUnd}_i(\alpha, \psi, \varphi)$ implies $\mathbf{G}_i\varphi \wedge \overline{\psi} \sqsubseteq \varphi$, which by proposition (3.50) implies $\mathbf{Undes}_i(\overline{\overline{\psi}}) = \mathbf{Undes}_i(\psi)$ and then by proposition (5.23) implies $\mathbf{Blamew}_i(i{:}\alpha)$. □