Comparing Formal Cognitive Emotion Theories

Koen V. Hindriks[1] and Joost Broekens[1]

Delft University of Technology, The Netherlands,
{k.v.hindriks,d.j.broekens}@tudelft.nl,
WWW home page: http://mmi.tudelft.nl/~koen

Abstract. We discuss an approach for evaluating cognitive emotion theories that explicitly relies on the use of formal logics of agency. This approach offers various benefits, which range from the fact that a formal logical approach is able to provide an explicit account of the inferential machinery used by an agent to elicit emotions to the fact that logical specifications provide a precise model of what is and what is not relevant in a computational model of emotions. Several *formal* emotion models have been proposed that *derive* emotions from the mental state of an agent using basic *logic*. The benefit of formal models of emotions is that they allow for a precise comparison. One of the key challenges, however, in comparing these models is that the conceptual basis that provides the starting point for deriving emotions is different. We present two simplified formal models and provide a preliminary comparison for illustrative purposes and identify some problems in comparing them. We also propose an alternative, more empirical approach for comparing and validating these models. To do so, we argue that there is a need for standard and detailed accounts of the evolution of emotions in particular scenarios that can be used as "benchmarks".

1 Introduction

Computational models of emotions are derived from psychological theories about emotions, and have been most influenced by the appraisal theory of [1] (the OCC model). The fact that appraisal theories or cognitive emotion theories have in particular drawn the attention of computational modellers of emotion perhaps may be explained by the fact that this model fits quite well with a more general notion of agency that incorporates a notion of mental state including beliefs and desires or goals and basic notions of action and perception. This model of agency, usually referred to as the BDI model of agency [2], is a computational model which provides a suitable starting point for constructing computational models of emotion derived from appraisal theory. One important question, however, is *how to obtain a computational model from an informal, psychological theory of emotions?* In other words, there is always the question whether the modellers did a faithful job of transforming the theory they used into a computational instantiation [3, 4].

An appealing approach for constructing a computational model of emotions is to first *formalize* a psychological theory of emotion and thereafter implement this formal model in a machine. This seems particularly appealing for emotion models based on appraisal theories as they rely on a notion of mental state that is present in agents and extensive work has been done on formalizing mental states and the associated common sense concepts.[1] Here the formalization provides a precise *specification* of the theory. This specification then is implemented to obtain a *computational model* of emotions. This approach aims at *translating* the psychological theory into a formal, *logical* specification of that theory. Such a translation requires a precise reconstruction of the theory in a formal logic which may uncover hidden assumptions, etc. This effort is useful if only because of the fact that it provides an *unambiguous statement* of the theory [5]. Apart from the usual benefits of formalization by means of logic, i.e. *conceptual clarification* by drawing "attention to some finer points in the logical structure of a theory" [6], the support it offers for evaluating the theory by e.g. checking the *consistency* of the theory, we are particularly interested in this approach because

[1] Note that other frameworks for decision making, e.g. decision theory, do not provide a similarly rich and mathematically precise model of the notion of mental state.

a formal specification can be *used as a starting point for implementation.*[2] One straightforward option could be to try and execute the logical theory that results itself. For agent logics, this has turned out to be more complicated than expected in practice, however.

One of the main benefits of taking the effort to first obtain a formal specification is that such a specification not only provides a precise model of what needs to be implemented but also clarifies *what is and what is not an important part of the computational model.* That is, only those features of the computational model that correspond to the abstract specification are relevant with respect to the theory that is being implemented whereas all other aspects including even important ones such as the software architecture may be considered irrelevant. This simply follows from the fact that *any* computational model that implements the specification is adequate. A formal specification also provides a clue as to what should be empirically validated since it highlights the foundational axioms of the theory and such a specification can be used to generate predictions that the theory makes.

Even though several approaches using formal logic to specify theories of emotions have been proposed, these formal specifications have not been used to empirically validate the theories that are formalized. We believe it may be very useful to investigate whether approaches using formal logic offer *a new point of view on evaluating emotion theories.* We expect that by starting and comparing different formal specifications of emotion theories a host of new questions may be generated that should be empirically validated. It is the purpose of this short paper to study some of the issues that are important to consider and to draw some initial conclusions on the usefulness of formal specifications of emotion theories for evaluating these theories. In essence, we ask ourselves whether we can reconnect pure, formal theory and empirical validation.

The approach we propose and briefly discuss in this short paper is to compare formal theories of emotions. To this end, in Section 2 we introduce and provide an overview of two formal theories of emotions proposed in the literature [9, 10]. In Section 3 we discuss various issues that need to be addressed when comparing, and, ultimately, empirically evaluating different formal theories of emotions. Section 4 concludes the paper. For now, we will not be particularly concerned with evaluating the benefits of using a formal approach per se although ultimately we would like to be able to draw some conclusions on the relative benefits of this approach compared to others.

2 Two Emotion Models based on the BDI Agent Model

One important observation concerning cognitive emotion theories is their implicit reliance on a more complicated cognitive architecture. Any cognitive emotion theory posits more or less implicitly certain other mental constructs, mechanisms, and attitudes that are basic for eliciting emotions and for describing changes in emotion. In fact, a basic theory of agency seems presupposed by cognitive emotion theories. One advantage of using a formal logic of agency to formalize such theories is that the *ontological commitments*, i.e. the things that are presupposed by the theory, can be clearly identified as well; in a similar spirit, though with somewhat different emphasis, [11] writes that "it is indispensable to consider the cognitive architecture that underlies emotions." Apart from clarifying *which mental constructs*, e.g. beliefs and goals, *are needed* in a cognitive emotion theory, it is of course also important to clarify their *role.* To do so, in a formal approach using logic, we are also forced to make these choices explicit by selecting the specific logical tool that provides the basic means for a formalization. As we will illustrate below, this introduces some issues of its own with respect to the predictions that the theory will generate.

We discuss two models of emotion that are built on top of the familiar BDI model of agency [2]. The first model is a formal, logical model based on the so-called KARO framework and aims to formalize the OCC model [10]. The second model is based on so-called Intention Logic [12, 13,

[2] Nevertheless it is worth emphasizing the benefits of conceptual clarification and conceptual coherence that are gained by using a logical formalism. What, for example, is the relationship between the concepts of "expectedness" and "likelihood", or, between "expectedness" and "causal attribution", used in [7]? The use of logic may provide a useful tool to address such questions. See also [8] which nicely illustrates our point.

9]. KARO is a rich, logical framework that provides a formal, logical account of agency proposed van Van der Linder, Van der Hoek, and Meyer [14]. Similarly, Intention Logic is a logic of agency proposed by Cohen and Levesque [15]. [11] also proposes a model that is partly formalized but not in a similarly rigorous way as is done by e.g. [10, 13].

2.1 Remarks on Notation

As we do not have the space to provide a comprehensive overview of the logical frameworks used nor of their application to obtain formalizations of cognitive emotion theories, we have simplified both presentation and notation quite a bit for the purpose of readability. First, we have used a more uniform presentation and notation to denote formulas from both the KARO framework as well as Intention Logic. We use $\mathsf{bel}(\varphi)$ to denote that the agent believes that φ; we do not distinguish between knowledge and beliefs here. We use $\mathsf{goal}(\varphi)$ to denote that the agent wants to be in a state where φ is true. The bel and goal operators are so-called *modal operators*. The usual operators for conjunction $\wedge$ and negation $\neg$ are used. We use $\Diamond\varphi$ to denote that φ is true at some time in the future.

KARO introduces a number of additional operators including $\mathsf{A}\pi$ which denotes that the agent is able (has the skills) to perform plan π, $\langle\pi\rangle\varphi$ which represents that the agent has the opportunity to realize φ by doing π, and $\mathsf{Comm}\pi$ which denotes that the agent is committed to performing plan π. Using these operators it is possible to define a range of additional operators. Most importantly, it is possible to define when an agent actually can realize a particular state of affairs.

2.2 The OCC Model Formalized in KARO

We only provide a simple illustration of how emotions are formalized in the KARO framework by means of the emotion of fear, and simplify the much more sophisticated model presented in [16]. Broadly speaking, a fear for φ in the OCC model is elicited by the new prospect that φ and the fact that the agent is displeased about φ. To represent the fact that a belief is new, a special operator new is introduced to model this aspect. Next we present a simplification of the actual formalization in [16] below as an illustration. We illustrate the basic scheme for formalizing the elicitation conditions for particular emotions. For example, using the operators discussed, for the emotion of fear we obtain:

$$fear(\varphi) ::= \mathsf{new}(\mathsf{bel}(\Diamond\varphi)) \wedge \mathsf{goal}(\neg\varphi)$$

Here, the formula defining the elicitation condition is a relative straightforward translation of the informal definition in natural language which says that the prospect that φ will happen - formalized by a belief that φ at some point in the future holds - and being displeased about this prospect - formalized by a goal that $\neg\varphi$ - triggers the emotion of fear.

2.3 Emotions Modelled in Intention Logic

Castelfranchi and colleagues put more emphasis on the anticipatory nature of various emotions. As a result, they use different agent logics to formalize the elicitation conditions of emotions. Here we base our discussion on [13], which is based in part of work of [15]. [15] introduced a logic also called Intention Logic.

We again only provide an illustration of a simplified version of a formalization of fear and include only the basic ingredients mentioned in [13]. It should be recognized that this does not do justice to the full analysis provided in [13] but we only use this simplified model to illustrate some basic points about the use of various agent logics. In [13] the basic ingredients are a notion of possibility that something may happen, to represent a basic anticipatory concept. We slightly abuse notation and use the $\Diamond$ operator again to represent this notion of possibility which is fitting here because the notion plays a very similar role representing a basic notion of *future* anticipation. To model the elicitation condition of fear, like [16] [13] argues that a basic belief must be in place

about a state that is considered possible (is "anticipated" or "expected") and a goal to avoid this state. Formally, we obtain a condition that is slightly more simple than that provided above:

$$fear(\varphi) ::= \mathsf{bel}(\Diamond\varphi)) \wedge \mathsf{goal}(\neg\varphi)$$

Basically, the condition is the same as that provided above based on [16]. However, the informal meaning associated with $\Diamond$ is richer and from this perspective it is not completely clear yet how to compare both formalizations. A deeper analysis of the logical axioms would be required to gain more insight into these differences.

More generally, we briefly discuss one interesting aspect related to the choice of agent logic used for formalizing elicitation conditions. Agent logics may introduce basic assumptions of their own. For example, Intention Logic assumes an axiom which is called *realism* in the agent literature which is not accepted by everyone. Realism imposes a relation on the beliefs and goals of an agent. The idea is that a rational agent should only have goals it considers feasible. Intuitively, this seems a reasonable demand. The discussion is about the particular formalization of this principle which makes the formula $\mathsf{bel}(\varphi) \rightarrow \mathsf{goal}(\varphi)$ valid. At first sight, this seems wrong. For example, if you believe that it is raining, intuitively it does not follow that you want it to rain. Given this simple instance it is clear that the formal operator goal itself does not correspond to our common sense notion of a goal. However, one might argue that this is a more primitive operator that is needed in order to represent our common sense concept of a goal correctly in a formal theory. This is not the place to discuss the details of this proposal, however; see e.g. [17]. This particular example has been discussed because it highlights that even the underlying logic itself may introduce controversial assumptions which may not be easily avoided. More generally, logical approaches often aim for a model of idealized rationality and this raises the issue whether it is reasonable to use models that make such assumptions in empirical research. Even so, this should not be taken as an obstacle per se, as empirical research in game theory which makes similar idealized rationality assumptions has been quite successful [18]. More importantly, there must be at least some meaning of these notions that is common and shared by all of us because such a common understanding is required for making sense of each other.

3 Comparing and Evaluating BDI-Based Emotion Models

Formalising cognitive emotion theories provides the benefit of a precise specification for implementation as well as precise predications that can be derived using a formal, logical model. Basically this offers two roads for evaluating a theory: formal and empirical.

It is possible, for example, using formal techniques to verify using basic model checking whether a particular evolution of a scenario involving emotions is consistent with a formal cognitive emotion theory. More interestingly, it becomes possible to verify whether the theory would actually predict a particular evolution of emotions in a scenario using semi-automated theorem-proving techniques. We have already discussed various issues that need to be taken into account with this method of evaluation above, but most importantly the key issue in comparing the emotion models we have introduced above is that the agent language used as a starting point to derive emotions is different. This does not invalidate the usefulness of the techniques mentioned for theory evaluation but does add a complicating dimension since the conceptual basis of the theories also needs to be compared and assessed on its merits.

In the approach we advocate here empirical validation should play a key part in the evaluation of emotion theories. In principle, we believe this requires making the step from specification to implementation and validating the theory by means of comparing simulation runs with empirical data. Two main goals would be to evaluate the *predictive power* and the *face value* of different models. That is, first, evaluate whether the model predicts the emotional state of someone correctly based on information about the particular mental state that person is in? And, secondly, by running computational simulations derived from these formal models, how do subjects evaluate the emotions predicted by the formal model given that they understand the basic mental attitudes present in agents? We believe that one appealing approach is to develop paradigm examples that can be used to "benchmark" emotion theories.

3.1 Empirical Evaluation: Example Paradigms

We thus argue that a formal comparison should be complemented with empirical validation. Apart from a formal comparison, which already shows potential for evaluating theories to some extent and would already provide insights on some of the main differences between approaches, there is a need to empirically evaluate emotion theories. As argued in the Introduction, logical theories provide a good starting point for evaluation because such theories clarify both the foundations as well as the predictions that a theory makes.

To this end, it is important to come up with good examples that can be used in experimental research. [19] argues that surprise might be a paradigm example for performing experimental research. One particularly interesting alternative is to use the bird example discussed in [7]. This example consists of a scenario where humans take part in an experiment which is interrupted by a pigeon that flies into the room through an open window. As the experiment was video-taped, so was the pigeon event and data was obtained by analysing the video about the emotions of the human participants. [7] provides a detailed account of the series of events that elicited and caused a change in emotions during an interruption of an experiment due to a pigeon that flew into the room through an open window.

As a first start, each of the theories discussed above might be tested on the different parts of the bird scenario of [7]. As an illustration, we briefly discuss the elicitation of the fear emotion. [7] introduce various propositional representations and introduce, for example, the proposition *birdApproaching* to represent that the agent believes the bird is getting closer. To model the example in the agent logics used here, we can already observe the importance of the choice of representation. In order faithfully represent the approaching of the bird, we need to make a slight modification in the representation and explicitly represent the temporal dimension by using a temporal operator, i.e. $\mathsf{bel}(\Diamond(birdHere))$. In other words, the formal logics used here require a more fine-grained analysis to be able to make the right inferences. This highlights yet another advantage of using a formal logic: it provides a more detailed account of the *inferential machinery* needed to model cognitive emotion theories. Continuing the example, as a consequence of the bird approaching, the agent infers that it might get injured, which clearly also is something projected into the future; we represent this conclusion by: $\mathsf{bel}(\Diamond(injured))$. The agent does not want to be injured by the bird, i.e. $\mathsf{goal}(\neg injured)$.[3]

The above would be sufficient to conclude that the agent fears $birdHere$ using the KARO formalization except for a minor detail that is missing. That is, we need $\mathsf{new}(\mathsf{bel}(\Diamond(birdHere))$ to be able to conclude that the belief is new. This is a very reasonable addition given the explanation in [7], however.

Given that the basic formalization based on [13] is even simpler but still consists of the minimal ingredients used to model the fear example related to the bird scenario, it is not surprising that we find we can derive the agent has fear according to the theory of [13] as well. We do not want to make strong claims here about the actual theories of [13] or [16], which are more sophisticated than we can illustrate here.

The main point that we do want to make is that even an example as simple as the conditions for fear that we have presented already raises questions related to the modelling of these conditions. Should we explicitly represent the fact that an agent comes to learn something new as in [16] or can we leave this more implicit by using a possibility operator as in [13]? Clearly, in the former case, different conditions for validating the theory should be checked than in the latter case.

4 Discussion and Conclusion

We have only illustrated some of the main ideas by means of simplified examples and much more work is needed to present a more thorough analysis and comparison of the frameworks. We believe,

[3] Note that saying that an agent does not want something requires shifting the negation inside the modal goal operator to obtain a correct formalization in this case.

however, that there is great potential in doing so and it is worthwhile to explore and investigate formal cognitive theories of emotion in greater depth.

Do logical theories provide computational models of emotions? As we have argued, they do not provide computational models per se but do provide specifications of such models. They highlight the aspects of these models that are in need of empirical investigation and clarify which aspects of computational models are not relevant from that perspective. Of course, before we can engage in empirical evaluation, there is a need to actually engineer a computational model from the logical specification and we have not discussed this. [11] which proposes a belief-desire theory of emotions which is similar in many respects to the more formal approach based on logics of agency discussed above provides a sketch of a computational architecture and argues that various mechanisms are needed. In particular, [11] argues for the need of belief-belief and belief-desire comparators.

We like to propose the use of existing BDI architectures for empirically evaluating cognitive emotion architectures. There are many benefits associated with this proposal. It avoids an ad hoc construction of an architecture for a computational model of emotions but instead builds on already well-established computational models of agency. Another benefit is that these architectures have been connected to e.g. games, virtual worlds, and robots, for example. This may also broaden the evaluative scope of emotion theories, as we should not only look for validation in virtual worlds but also where possible use e.g. robots. In the near future, we would like to test some of the experimental scenarios we introduced by means of the Goal agent platform [20] which incorporates the notions of beliefs and goals and facilitates reasoning with these mental attitudes. The Goal platform is firmly connected to agent logics and implements part of Intention Logic [21]. Interestingly, all of the so-called *cognitive operators* listed in [7] are readily available in the agent platform Goal, except maybe for speech recognition and generation. As a first step, we intend to add the belief-belief and belief-desire comparators of [11] and add a basic model of emotions incorporating type and intensity.

One challenge for the approach we propose is how to evaluate subtle distinctions made in formal logics. That is, how can we translate subtle differences made in different formalizations of appraisal theories into operational constructs that can be observed in experiments.

References

1. Orthony, A., Clore, G., Collins, A.: The cognitive structure of emotions. Cambridge University Press (1988)
2. Rao, A., Georgeff, M.P.: Bdi-agents: From theory to practice. In: Proceedings of the First International Conference on Multiagent Systems (ICMAS'95). (1995) 312–319
3. Broekens, J., DeGroot, D.: Formalizing cognitive appraisal: From theory to computation. In: Proceedings of the 18th European Meeting on Cybernetics and Systems Research (EMCSR'06). (2006) 595–600
4. Joost Broekens, D.D., Kosters, W.A.: Formal models of appraisal: Theory, specification, and computational model. Cognitive Systems Research **9**(3) (2008) 173–197
5. Kamps, J.: On criteria for formal theory building: Applying logic and automated reasoning tools to the social sciences. In: Proceedings of the AAAI'99, AAAI Press/The MIT Press (1999) 285–290
6. Masuch, M., Huang, Z.: A case study in logical deconstruction: Formalizing j.d. thompson's organizations in action in a multi-agent action logic. Computational & Mathematical Organization Theory **2**(2) (1996) 71–113
7. Marsella, S.C., Gratch, J.: EMA: A process model of appraisal dynamics. Cognitive Systems Research **10** (2009) 70–90
8. Steunebrink, B.R., Dastani, M., Meyer, J.J.C.: The OCC Model Revisited. In Reichardt, D., ed.: Proceedings of the 4th Workshop on Emotion and Computing - Current Research and Future Impact. (2009)
9. Castelfranchi, C., Lorini, E.: Cognitive anatomy and functions of expectations. In: Proceedings of IJCAI'03 Workshop on Cognitive Modeling of Agents and Multi-Agent Interactions, Acapulco, Mexico, August 9-11. (2003)
10. Steunebrink, B.R., Dastani, M., Meyer, J.J.C.: A Formal Model of Emotions: Integrating Qualitative and Quantative Aspects. In: Proceeding of the 18th European Conference on Artificial Intelligence (ECAI'08). (2008) 256–260

11. Reisenzein, R.: Emotions as metarepresentational states of mind: Naturalizing the beliefdesire theory of emotion. Cognitive Systems Research (2009) 6–20
12. Castelfranchi, C.: Affective appraisal *versus* cognitive evaluation in social emotions and interactions. In Paiva, A., ed.: Affective Interactions. Volume 1814 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg (2000) 76–106
13. Miceli, M., Castelfranchi, C.: The mind and the future: The (negative) power of expectations. Theory & Psychology **12** (2002) 335–366
14. van der Hoek, W., van der Linder, B., Meyer, J.J.C.: An Integrated Modal Approach to Rational Agents. In: Foundations of Rational Agency. Volume 14 of Applied Logic Series. Springer (1999) 133–168
15. Cohen, P.R., Levesque, H.J.: Intention is choice with commitment. Artificial Intelligence **42** (1990) 213–261
16. Steunebrink, B.R.: The logical structure of emotions. PhD thesis, Utrecht University (April 2010)
17. Hindriks, K.V., van der Hoek, W., van Riemsdijk, M.B.: Agent programming with temporally extended goals. In: Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'09). (2009) 137–144
18. Binmore, K.: Does Game Theory Work? The Bargaining Challenge. The MIT Press (2007)
19. Reisenzein, R., Meyer, W.U., Schützwohl, A.: Reactions to surprising events: A paradigm for emotion research. In Frijda, N., ed.: Proceedings of the 9th conference of the International Society for Research on Emotions (ISRE'96). (1996) 292–296
20. Hindriks, K.V.: The **Goal** agent programming language. `http://mmi.tudelft.nl/trac/goal` (2011)
21. Hindriks, K., van der Hoek, W.: Goal agents instantiate intention logic. In Hölldobler, S., Lutz, C., Wansing, H., eds.: Logics in Artificial Intelligence. Volume 5293 of Lecture Notes in Computer Science., Springer (2008) 232–244