

Agents with emotions: a logical perspective

Emiliano Lorini

Institut de Recherche en Informatique de Toulouse (IRIT), France
lorini@irit.fr

Abstract

The aim of this short paper is to discuss the role played by formal logic in the research on synthetic emotions. Indeed, logical methods have been recently exploited in order to provide a rigorous specification of how emotions should be implemented in an artificial agent and how agents should reason about and display some kind of emotions. I will emphasize that, although the application of logical methods to the formal specification of emotions has been quite successful, there is still much work to be done. Indeed, there is still no formal model in the literature which is able to characterize in an adequate way *complex emotions* such as regret, jealousy, envy, shame, guilt, reproach, admiration, remorse, pride, embarrassment. These emotions are complex in the sense that they involve very sophisticated forms of reasoning such as self-attribution of responsibility, counterfactual reasoning, reasoning about norms and ideals. In the last part of the paper I will show how current logical approaches can be extended in order to provide a comprehensive formal framework for the specification of complex emotions.

1 Introduction

In the recent years Internet has become a pervasive technology. Online services increase more and more every year. Among the different online services there are: ticket systems (for trains, movies, etc.); practical information systems (restaurants, hotels, tourism, etc.); cultural events (art, literature, history, etc.); tutoring systems. Interactive systems are at the heart of a new generation of Internet services. The general idea is that people must have direct access to a large set of services without any other help. Since the complexity of an interactive system increases with the complexity of the services it offers, we need, more than ever before, understandable and intuitive interfaces and systems. Therefore, a major objective in the area of information technologies is to develop interactive systems that are more attractive and closer to the users and that can be considered as believable interlocutors. These systems must be endowed with refined communicative capabilities. In this perspective, a technological challenge is to build machines which are capable: to reason about emotions, to predict and understand human emotions, and to process emotions in reasoning and during the interaction with a human user. This challenge concerns a large variety of interaction systems such as virtual agents, tutoring agents, personal scheduler agents, etc.

With the aim of creating a new generation of emotional interaction systems, the study of affective phenomena has become a “hot” topic in computer science and artificial intelligence (AI) where the domain of Affective Computing [33] has emerged in the last few years. Emotion has a long standing tradition in several disciplines such as psychology [12, 37, 14, 19, 29], economics [20, 13], cognitive neuroscience [18]. In the last years, research in artificial intelligence and computer science has addressed modelling and communication of expressive, emotional content. Consider for example research on Affective Computing at MIT [33] and on Kansei Information Processing in Japan [40]. Such researches led to the development of several systems prototypes for many different uses, such as expressive personal assistants and Embodied Conversational Agents (ECAs) [8], virtual environments conveying emotional information for enhanced user experience, robots displaying emotional behavior, virtual agents for entertainment (e.g. games, interactive storytelling).

In research on expression and recognition of emotion in computer science, the main interest is in the design of anthropomorphic systems that can interact with human users in a multi-modal way. Such systems are justified by the various forms of ‘anthropomorphic behavior’ that users show toward technologies. Intelligent ECAs use a model of emotions to simulate the user’s emotion and show their affective state and personality. This is in order to appear more believable [4], but also to adapt their behavior to the user’s emotions, preferences, attitudes and needs. Indeed, there are many evidences suggesting that virtual agents and robots (interacting with humans) that are capable to display emotions, to recognize the human users’ emotions, and to respond to their emotions in an appropriate way, allow to induce positive feelings in the humans during the interaction and to improve their performance. For instance, it has been shown that machines which express emotions and provide emotional feedback to the human user, enable to enhance the user’s enjoyability [35], his performance in task achievement [31], and his perception of the machine [6, 34].

Some authors in the field of Affective Computing have also analyzed empathy, sympathy as fundamental concepts for the design of socially intelligent agents [30, 28]. Empathy has especially been considered as a human-like quality which can be used to enhance the human user’s satisfaction and perception of the agent [6, 35].

2 Logical modelling of emotions

Recently, some researchers working in the field of multi-agent systems (MAS) have been interested in developing logical frameworks for the formal specification of emotions (see [25, 39, 11, 1, 42] for instance). Their main concern is to exploit logical methods in order to provide a rigorous specification of how emotions should be implemented in an artificial agent. The design of agent-based systems where agents are capable to reason about and display some kind of emotions can indeed benefit from the accuracy of logical methods. My contribution in this area has been a logical framework for the specification of graded beliefs and goals in which anticipatory emotions such as surprise, hope, fear, disappointment, relief can be formalized [21, 9, 24, 22].

All proposed logical frameworks for the specification of emotions are based on the so-called BDI (belief-desire-intention) logics which have been developed in the last fif-

teen years in the MAS domain (see [10, 23, 26, 36, 38, 44] for instance). BDI logics are multimodal logics in which agents' attitudes such as beliefs, goals, desires, intentions, etc. are formalized. Such concepts are generally expressed by corresponding modal operators and their interaction properties are specified. By way of example, consider a simple BDI logic with operators for belief, desire and time where time is supposed to be linear as in [10]. The operator Bel_i is used to express what a certain agent i believes. The operator $Goal_i$ is used to express what a certain agent i wants. The operator X is used to express those facts that will be true in the next state of system. Given an arbitrary formula ϕ of the logic:

- $Bel_i\phi$ is meant to stand for: agent i believes that ϕ ;
- $Goal_i\phi$ is meant to stand for: agent i wants ϕ to be true;
- $X\phi$ is meant for: ϕ will be true in the next state of the system.

An operator $Poss_i$ can be defined as the dual of the belief operator Bel_i , that is, $Poss_i\phi \stackrel{\text{def}}{=} \neg Bel_i\neg\phi$. $Poss_i\phi$ is meant to stand for: agent i thinks that ϕ is possible.

BDI logical approaches have been used to disambiguate the different dimensions of emotions which have been identified in the existing psychological models of emotions. Let me just present a couple of examples in order to show how the previous modal operators for belief, goal and time can be used for formalizing some basic emotions.

The fact a certain agent i feels joy about a certain fact ϕ (or rejoices about ϕ) can be expressed as follows:

$$Joy_i\phi \stackrel{\text{def}}{=} Bel_i\phi \wedge Goal_i\phi$$

According to this definition, agent i feels joy about ϕ if and only if i believes that ϕ is true and wants ϕ to be true. In this sense, i is pleased by the fact that he believes he has achieved a desirable result. The fact a certain agent i hopes that ϕ will be true can be expressed as follows:

$$Hope_i\phi \stackrel{\text{def}}{=} Poss_iX\phi \wedge Goal_iX\phi$$

According to this definition, agent i hopes that ϕ will be true in the next state if and only if i thinks that ϕ will be possibly true in the next state and i wants ϕ to be true in the next state.¹

Although the application of logical methods to the formal specification of emotions has been quite successful, there is still much work to be done. In fact, there is no formal model in the literature which is able to characterize in an adequate way *complex emotions* such as regret, jealousy, envy, shame, guilt, reproach, admiration, remorse, pride, embarrassment. With 'complex emotions' I mean those emotions that are based

¹This simple BDI logic of belief, goal and time could be extended to a concept of *graded belief* in order to account for *intensity* of emotions. Such an extension would enable reasoning about beliefs with different degrees of certainty in such a way that a concept of *graded hope* could be formalized. In the logic, the intensity of hope would depend on the degree of certainty of the corresponding belief.

on complex forms reasoning and on a larger set of appraisal variables than the set involved in ‘basic emotions’. As it has been stressed in the psychological literature such emotions involve very sophisticated forms of reasoning such as: self-attribution of responsibility [43], counterfactual reasoning [45, 17, 15], reasoning about norms and ideals [27].

For instance, several psychologists have emphasized the peculiar counterfactual dimension of regret: an agent regrets something he did only if he is aware to have taken a certain decision which turned out to be worse than expected, and the agent believes that if he had decided to do something different, he would have achieved better. Other psychologists working on guilt and shame [41, 27] have stressed the fact that these emotions involve not only counterfactual reasoning and the concepts of causality and responsibility, but also normative and moral aspects. In particular, such emotions are often based on an agent’s *internalization* of some social norm. For example, an agent feels guilty for something only if he thinks to be responsible for having violated a certain norm that he has internalized and he thinks that if he did something else he would not have incurred a violation.

In the next section I will provide a more detailed discussion of the main components of complex emotions. I will identify some necessary desiderata for the development of a comprehensive logical model of complex emotions. Indeed, I consider this as one of the main challenge in the field of Affective Computing.

2.1 Towards complex emotions

As emphasized above, the following are some of the most important dimensions that a logic of complex emotions should take into account.

- **Counterfactual dimension.** As stressed in the psychological literature some complex emotions such as regret and guilt are based on an agent’s counterfactual reasoning about its own choices and actions. For instance, an agent feels regrets if and only if he believes having failed to achieve one of its goal and believes that if it decided to do something different, it would have achieved the goal. In this sense, there are complex emotions which are based on the capacity to *imagine* alternative scenarios to the actual one that could have realized if something different was done. An aspect of counterfactual emotions which has been emphasized in the psychological literature is the distinction between *action* and *inaction*. For instance, an agent feels regret not only when it thinks that its decision *to do* a certain action led it to a bad outcome but also, when it thinks that its decision *to refrain from doing* a certain action led it to a bad outcome.
- **Moral dimension.** Many complex emotions such as guilt, remorse, pride, reproach, shame, admiration, embarrassment, involve normative and moral aspects. In particular, such emotions are often based on an agent’s *internalization* of some social norm (e.g. an obligation). A social norm internalized by an agent becomes an *ideal* of the agent that the agent is committed to pursue and to defend by punishing possible violators. For instance, to feel guilty for something it is not sufficient that an agent believes to be responsible for violating a certain prescription. I do not feel guilty if I do not respect Ramadan prescription “Do

not eat before the dawn during Ramadan”. Since I am not a Muslim and I am not identifying myself with that religion, Muslim norms are not parts of my ideals, and I simply do not care if I violate such norms. To feel guilty, an agent should think that it has done something which does not conform to the norms that it has internalized.

- **Responsibility dimension.** Some complex emotions such as regret, guilt, shame, reproach, remorse are based on an agent’s self-attribution of responsibility (e.g. guilt, shame, remorse) and an agent’s attribution of responsibility to other agents (e.g. reproach, admiration). For instance, an agent feels guilty only if he thinks to be responsible for having violated a certain norm that he has internalized (i.e. an ideal), and an agent feels admiration toward someone only if he thinks that the other was responsible for something that has happened which conforms to its ideals. The concept of responsibility must be carefully analyzed in order to understand the cognitive structure of complex emotions. In particular, *full responsibility* must be distinguished from *partial* or *co-responsibility*. In the former case, an agent i is said to be fully responsible for the occurrence of some state of affairs p if and only if the agent brought about p independently from what the other agents did. In latter case, the agent i is said to be *co-responsible* for the occurrence of p if and only if i brought about p together with some other agents, that is, p was the effect of a joint action of a group of agents which included i . This distinction between full and co-responsibility allows a better understanding of how the intensity of complex emotions should be calculated. For instance, the feeling of guilt is more intense when the agent believes to be fully responsible for having violated a norm that it has internalized rather than when he believes to be co-responsible for the the violation of the norm.

In order to build a logic which enables reasoning about the previous dimensions and allows to specify complex emotions (e.g. regret, jealousy, envy, shame, guilt, reproach, admiration, remorse, pride, embarrassment), it is necessary to go beyond the standard BDI approach discussed in the previous section 2. Indeed, a logic of complex emotions should be sufficiently expressive not only to characterize different types of agents’ mental attitudes (beliefs, desires, goals, intentions), but also to characterize the previous concepts of *responsibility*, *counterfactual*, *norms* and *ideals*.

I think that a promising solution for developing a logic of complex emotions is to combine a BDI logic of agents’ mental attitudes, with a logic of norms and ideals (usually referred to as deontic logic [3]), and with a logic of agency and multi-agent interaction [5, 2, 32]. The latter kind of logics are indeed well-suited to support both counterfactual reasoning and reasoning about responsibility of single agents and groups of agents.

There are several available logics of agency and multi-agent interaction in the literature in formal philosophy [5] and theoretical computer science [2, 32] such as Coalition Logic (CL) [32], Alternating-time temporal logic (ATL) [2] and the logic of *Seeing to It That* (STIT) [5, 16]. Here I only consider the modal logic STIT which has been proposed in the domain of formal philosophy in the 90ies. More recently, it has been imported into the field of theoretical computer science where its formal relationships with other logics for multi-agent systems have been studied [7].

STIT is a logic which supports reasoning about agents' responsibilities and powers, and counterfactual reasoning about agents' actions and choices. The semantics of STIT is based on a branching time structure with ordered *moments*. In STIT logic operators of the form $Stit_{\{i\}}$ for any individual agent i and $Stit_C$ for any group of agents C are introduced. Moreover, operators of historical necessity and historical possibility respectively noted \square and \diamond are used. The modal formulas $Stit_{\{i\}}\phi$ and $Stit_C\phi$ respectively express that: a certain agent i brings it about that the state of affairs ϕ is true no matter what the other agents do, and a group of agents C brings it about that the state of affairs ϕ is true no matter what the agents outside C do. The modal formula $\diamond\phi$ expresses that ϕ is true in at least one history passing through the moment in which the formula is evaluated (formula $\square\phi$ expresses that for all possible histories passing through the moment at which the formula is evaluated ϕ is true). STIT is endowed with temporal operators for the past and the future. For instance, X is used to express those facts which are true in the next moment and X^- is used to express those facts which were true in the previous moment. A reasonable assumption to be made in STIT is that, once all agents have decided to do something, the effect of the joint action of all agents is deterministic. This assumption is expressed by the formula $Stit_{AGT}X\phi \leftrightarrow X\phi$, where AGT denotes the group including all agents. This formula is meant to stand for "all agents bring it about that the state of affairs ϕ is true in the next moment if and only if ϕ will be true in the next moment".

An interesting aspect of STIT logic is that it is sufficiently expressive to reason about counterfactual situations involving actions and choices of agents. For example, formula $\phi \wedge X^- \neg Stit_{AGT \setminus \{i\}} X\phi$ captures an essential aspect of counterfactual reasoning about actions by denoting that ϕ is true at present, and the fact that ϕ is true at present depends on the action chosen by agent i at the previous moment, that is, agent i could have done otherwise and ensured ϕ to be false now.²

I think that an integration of STIT logic with a logic of agents' mental attitudes (in the style of BDI logics) will enable to characterize complex emotions such as regret, disappointment, jealousy, envy, shame, guilt, reproach, admiration, remorse, pride, embarrassment which are based on counterfactual reasoning and involve the concepts of causality and responsibility. For example, by adding modal operators for beliefs and goals to the standard constructions of STIT logic, we can come up with the following formal characterization of the concept of *regret*:

$$Regret_i\phi \stackrel{\text{def}}{=} Bel_i\phi \wedge Bel_iX^- \neg Stit_{AGT \setminus \{i\}} X\phi \wedge Goal_i \neg\phi$$

According to this definition, agent i regrets for ϕ (noted $Regret_i\phi$) if and only if i wants ϕ to be false (noted $Goal_i \neg\phi$), believes that ϕ is true (noted $Bel_i\phi$), and believes that he could have done otherwise and ensured ϕ to be false now (noted $Bel_iX^- \neg Stit_{AGT \setminus \{i\}} X\phi$).

Imagine a situation in which there are only two agents i and j , that is, $AGT = \{i, j\}$. Agent i decides to park his car in a no parking area. Agent j (the policeman) fines agent i 100 euros. Agent i regrets for having been fined 100 euros (noted $Regret_i fine$). This means that, i wants $fine$ to be false (noted $Goal_i \neg fine$), believes

²Indeed, the formula $X^- \neg Stit_{AGT \setminus \{i\}} X\phi$ expresses that "the group of agents $AGT \setminus \{i\}$ did not bring about ϕ no matter what agent i did".

that *fine* is true (noted $Bel_i fine$), and believes that he could have done otherwise (to park elsewhere) and ensured *fine* to be false now (noted $Bel_i X^{-} \neg Stit_{AGT \setminus \{i\}} X fine$).

References

- [1] C. Adam, B. Gaudou, A. Herzig, and D. Longin. OCC's emotions: a formalization in a BDI logic. In *Proceedings of the Twelfth International Conference on Artificial Intelligence: Methodology, Systems, and Applications (AIMSA'06), LNAI, vol. 4183*, pages 24–32. Springer Verlag, 2006.
- [2] R. Alur and T. Henzinger. Alternating-time temporal logic. *Journal of the ACM*, 49:672–713, 2002.
- [3] L. Åqvist. Deontic logic. In D. M. Gabbay and F. Geunther, editors, *Handbook of Philosophical Logic*. Kluwer Academic Publishers, Dordrecht, 2002.
- [4] J. Bates. The role of emotion in believable agents. *Communications of the ACM*, 37(7):122–125, 1994.
- [5] N. Belnap, M. Perloff, and M. Xu. *Facing the future: agents and choices in our indeterminist world*. Oxford University Press, New York, 2001.
- [6] S. Brave, C. Nass, and K. Hutchinson. Computers that care: Investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International Journal of Human-Computer Studies*, 62:161–178, 2005.
- [7] J. Broersen, A. Herzig, and N. Troquard. Embedding alternating-time temporal logic in strategic STIT logic of agency. *Journal of Logic and Computation*, 16(5):559–578, 2006.
- [8] J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors. *Embodied Conversational Agents*. MIT Press, Cambridge, MA, 2000.
- [9] C. Castelfranchi and E. Lorini. Cognitive anatomy and functions of expectations. In R. Sun, editor, *Proceedings IJCAI'03 Workshop on Cognitive modeling of agents and multi-agent interaction*, pages 29–36, 2003.
- [10] P. R. Cohen and H. J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42:213–61, 1990.
- [11] M. Dastani and J.-J. Ch. Meyer. Programming agents with emotions. In *Proceedings of 17th European Conference on Artificial Intelligence (ECAI 2006)*, pages 215–219. IOS Press, 2006.
- [12] P. Ekman. *Emotions in the human face*. Pergamon Press, New York, 1972.
- [13] J. Elster. Emotions and economic theory. *Journal of Economic Literature*, 36(1):47–74, 1998.
- [14] N. Frijda. *The Emotions*. Cambridge University Press, Cambridge, 1987.
- [15] T. Gilovich and V. H. Medvec. The experience of regret: What, when and why. *Psychological Review*, 102:379–395, 1995.
- [16] J. F. Horty. *Agency and Deontic Logic*. Oxford University Press, Oxford, 2001.
- [17] D. Kahneman. Varieties of counterfactual thinking. In N. J. Roese and J. M. Olson, editors, *What might have been: the social psychology of counterfactual thinking*, pages 375–396. Erlbaum, Mahwah, NJ, 1995.
- [18] R. Lane and L. Nadel, editors. *The cognitive neuroscience of emotions*. Oxford University Press, New York, 2000.
- [19] R. S. Lazarus. *Emotion and adaptation*. Oxford University Press, New York, 1991.
- [20] G. Loewenstein. Emotions in economic theory and economic behavior. *American Economic Review*, 90(2):426–432, 2000.
- [21] E. Lorini and C. Castelfranchi. The cognitive structure of surprise: looking for basic principles. *Topoi: an International Review of Philosophy*, 26(1):133–149, 2007.
- [22] E. Lorini and R. Falcone. Modeling expectations in cognitive agents. In C. Castelfranchi, C. Balke-nius, M. Butz, and A. Ortony, editors, *Proceedings of AAAI 2005 Fall Symposium-From Reactive to Anticipatory Cognitive Embodied Systems*, pages 114–121, Menlo Park, 2005. AAAI Press.

- [23] E. Lorini, A. Herzig, and C. Castelfranchi. Introducing *attempt* in a modal logic of intentional action. In M. Fisher, W. Van der Hoek, and B. Konev, editors, *Proceedings Tenth European Conference on Logics in Artificial Intelligence (JELIA 2006)*, volume 4160 of *Lecture Notes in Artificial Intelligence*. Springer, 2007.
- [24] E. Lorini and M. Piunti. The benefits of surprise in dynamic environments: from theory to practice. In R. Paiva, A. Prada and R. W. Picard, editors, *Affective Computing and Intelligent Interaction (ACII 2007)*, volume 4738 of *LNCS*, pages 314–325. Springer Verlag, 2007.
- [25] J.-J. Ch. Meyer. Reasoning about emotional agents. *International Journal of Intelligent Systems*, 21(6):601–619, 2006.
- [26] J. J. Ch. Meyer, W. van der Hoek, and B. van Linder. A logical approach to the dynamics of commitments. *Artificial Intelligence*, 113(1-2):1–40, 1999.
- [27] M. Miceli and C. Castelfranchi. How to silence one’s conscience: cognitive defenses against the feeling of guilt. *Journal for the Theory of Social Behaviour*, 28:287 – 318, 1998.
- [28] M. Ochs, C. Pelachaud, and D. Sadek. An empathic rational dialog agent. In A. Paiva, R. Prada, and R. W. Picard, editors, *Proceedings of Affective Computing and Intelligent Interaction, Second International Conference (ACII 2007)*, volume 4738 of *LNCS*, pages 338–349. Springer Verlag, 2007.
- [29] A. Ortony, G. Clore, and A. Collins. *The cognitive structure of emotions*. Cambridge University Press, Cambridge, 1988.
- [30] A. Paiva, J. Dias, D. Sobral, R. Aylett, P. Sobreperes, S. Woods, C. Zoll, and L. E. Hall. Caring for agents and agents that care: Building empathic relations with synthetic agents. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2004)*, pages 194–201. IEEE Computer Society, 2004.
- [31] T. Partala and V. Surakka. The effects of affective interventions in human-computer interaction. *Interacting with computers*, 16:295–309, 2004.
- [32] M. Pauly. A modal logic for coalitional power in games. *Journal of Logic and Computation*, 12(1):149–166, 2002.
- [33] R. W. Picard. *Affective Computing*. MIT Press, 1997.
- [34] R.W. Picard and K.K. Liu. Relative Subjective Count and Assessment of Interruptive Technologies Applied to Mobile Monitoring of Stress. *International Journal of Human-Computer Studies*, 65:396–375, 2007.
- [35] H. Prendinger and M. Ishizuka. The empathic companion: A character-based interface that addresses users’ affective states. *International Journal of Applied Artificial Intelligence*, 19:297–285, 2005.
- [36] A. S. Rao and M. P. Georgeff. Modelling rational agents within a BDI-architecture. In *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning (KR’91)*, 473–484, 1991. Morgan Kaufmann.
- [37] K. R. Scherer and P. Ekman, editors. *Approaches to Emotion*. Erlbaum, Hillsdale, N. J., 1984.
- [38] Y. Shoham. Agent-oriented programming. *Artificial Intelligence*, 60:51–92, 1993.
- [39] B. R. Steunebrink, M. Dastani, and J.-J. Ch. Meyer. A logic of emotions for intelligent agents. In *Proceedings of the 22th AAI conference on Artificial Intelligence (AAAI’07)*, pages 142–147. AAAI Press, 2007.
- [40] K. Suzuki, A. Camurri, P. Ferrentino, and S. Hashimoto. Intelligent agent system for human-robot interaction through artificial emotion. In *Proceedings of the 1998 IEEE International Conference on Systems, Man, and Cybernetics (SMC’98)*, pages 1055–1060, New York, 1998. IEEE Computer Society Press.
- [41] J. P. Tangney, J. Stuewig, and D. J. Mashek. Moral emotions and moral behavior. *Annual Review of Psychology*, 58:345–372, 2007.
- [42] P. Turrini, J.-J. Ch. Meyer, and C. Castelfranchi. Rational agents that blush. In R. Paiva, A. Prada and R. W. Picard, editors, *Affective Computing and Intelligent Interaction (ACII 2007)*, volume 4738 of *LNCS*, pages 314–325. Springer Verlag, 2007.

- [43] B. Weiner, editor. *An Attributional Theory of Motivation and Emotion*. Springer, New York, 1986.
- [44] M. Wooldridge. *Reasoning about rational agents*. MIT Press, Cambridge, 2000.
- [45] M. Zeelenberg and E. van Dijk. On the psychology of “if only”: regret and the comparison between factual and counterfactual outcomes. *Organizational Behavior and Human Decision Processes*, 97(2):152–160, 2005.