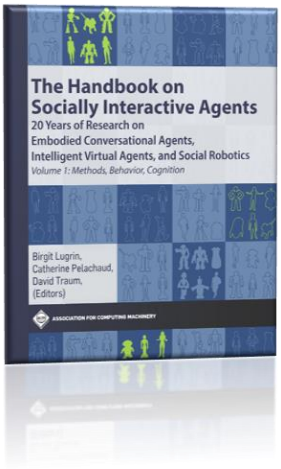
Emotion

Joost Broekens



Author note:

This is a preprint. The final article is published in "The Handbook on Socially Interactive Agents" by ACM books.

Citation information:

Broekens, J. (2021). Emotion. In B. Lugrin, C. Pelachaud, D. Traum (Eds.), *Handbook on Socially Interactive Agents – 20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics*, Volume 1: Methods, Behavior, Cognition (pp. 349-384). ACM.

DOI of the final chapter: 10.1145/3477322.3477333

DOI of volume 1 of the handbook: 10.1145/3477322

Correspondence concerning this chapter should be addressed to Joost Broekens (joost.broekens@gmail.com)

10 Emotion

Joost Broekens

In this chapter I cover computational modelling of emotion in Socially Interactive Agents. I focus on the computational representation of emotion and other affective concepts such as mood and attitude, and on computational modelling of appraisal, that is, the assessment of personal relevance of a situation. In Section 1, I define essential affective concepts in the study and modelling of emotion and discuss three different psychological perspectives towards studying emotion that have strongly influenced emotion modelling in SIAs. I will motivate why SIAs can make constructive use of emotions. In Section 2, I cover computational representation of affective concepts and four approaches towards computational modelling of appraisal. I also give four working examples. In Section 3-5, I cover the history, state of the art and outlook of emotion modelling in SIAs.

10.1 Motivation

10.1.1 What are emotions

Emotions are about feelings. Emotions tell us how a situation matters to us. Emotions also motivate us to do something about that situation, and we express them to let others know how we feel. Emotion is a multifaceted phenomenon involving a relation between action, motivation, expression, information processing, language, feelings and social interaction, as well as showing complex interactions with other affective and cognitive phenomena such as mood, attitudes, beliefs and decision making [Barrett et al. 2007, Damasio 1994, Fischer and Manstead 2008, Frijda et al. 2000]. An instance of an emotion is a specific combination of jointly active bodily and mental features, including expression, arousal, assessment of the situation in terms of personal relevance, often with an associated (learned) label [Broekens et al. 2013, Hoemann et al. 2019]. The core of an emotion is an assessment of the personal relevance of a situation thereby in some way providing feedback on the suitability of past, current, or future behaviour [Baumeister et al. 2007, Broekens et al. 2013, Frijda 2004, Lazarus 1991, Moors et al. 2013, Van Reekum and Scherer 1997]. Even in modern constructionist views this is an important underlying mechanism ('emotional events ... fundamentally occur within a brain that anticipates the body's energy needs in relation to the current context.') [Hoemann et al. 2019]. In this chapter, I will use the term *appraisal* for this process of assessment, independent of *how* this process is implemented in agent or human, or what the potential consequences on behaviour or further information processing are. This

is useful for our discussion of emotion simulation in Socially Interactive Agents (SIA) later on, as it distances us from debates around the nature of the appraisal process (in SIA's this is always grounded in binary information processing), it's causal role (in SIA's appraisal always causes the emotion), and the exact information processing involved in this assessment (in SIAs this is computationally implemented in many different ways).

However, what about mood, attitude, and other affective phenomena? To facilitate a clear discussion of emotion in SIAs, in this section I define essential concepts in the study and modelling of emotion. I cover affect, emotion, mood, attitude, relation, and personality (the latter two to a limited extent as chapter 11, 12, 18 and 19 cover these topics in more depth). Then I discuss three different psychological perspectives towards studying emotion that have strongly influenced emotion modelling in SIAs: the categorical view; the dimensional-constructionist view; and, the cognitive appraisal view. I will explain their main differences but also their commonalities. Please note that *cognitive appraisal* refers to this particular perspective on emotion elicitation, while *appraisal* is used as stated above. Finally, I briefly highlight the interplay between affect and cognition in humans.

10.1.1.1 Definitions

The study of emotion falls within the field of affective sciences. *Affect* in affective science is an umbrella term that refers to anything related to emotion, emotion processing and emotion in social interaction. Affective science thus deals with the study of emotion in the broadest sense. For the purpose of this chapter we introduce the most commonly used terminology when it comes to emotion modelling in Socially Interactive Agents. I will borrow some of the terminology from [Scherer 2005].

Affect also refers to the positiveness and negativeness associated with an emotion or other psychological construct (an attitude, a mood, a thought, a relation, etc...). For example, in mood induction studies [Dreisbach and Goschke 2004] (valence), in core affect [Russell 1980] (valence-arousal), and in affect associated with textual stimuli [Bradley and Lang 2007] (valence-arousal-dominance).

Emotion refers to an event-related affective reaction (it is about something) typically of short duration and relatively intense (one feels the emotion and is conscious of it). For example, joy is a strong and short-term reaction resulting from an event with an associated positive feeling. In psychology and neuroscience a distinction is made between emotions and feelings, where feeling is sometimes reserved for the subjective experience of the emotion ([Damasio 1994],p139), while othertimes emotion is reserved for the mental categorisation of the affective experience [Barrett 2005]. In this chapter I refer to feeling when I mean experience.

Mood refers to the longer term affective state an individual is in, is usually less intense, unrelated to a specific event, and less differentiated [Beedie et al. 2005]. For example, I can be in a cheerful mood, in which case I feel positive (i.e., positive associated affect) for no

particular reason (it is not directed at something specific) and although I feel good I am not necessarily laughing all the time (it is not intense). Mood influences emotion elicitation; pre-existing moods intensify congruent emotional responses [Neumann et al. 2001]. For example, being in a positive mood makes me more, and more easily joyful, just like a grumpy mood will make me more easily angry. Moods can be caused by psychological and biological events including repeated emotions (e.g. repeated exposure to negative events), thoughts (e.g., when ruminating or mind wandering), and changes in physiological state (e.g. hunger). Moods are also difficult to identify for people and can be unconscious.

Attitude refers to affect that has been associated with something or someone. For example, I like Chinese food (I have a positive association with Chinese food). Attitudes form due to repeated exposure to and appraisal of a stimulus. Attitude is also referred to as opinion or sentiment. For example in opinion mining and sentiment analysis [Liu and Zhang 2012], one tries to automatically identify the attitude the public has for a particular thing or person based on text data.

Related to attitudes are *relations* (interpersonal stance [Scherer 2005]), which are social attitudes attributed to other agents, typically other people but not exclusively. For example, I like my boss (i.e., I have a positive attitude towards my boss), and, I love my children (i.e., I have a positive attitude towards my children and I feel a bond). Relations are complex social constructs (see related chapters), but when it comes to emotion modelling in socially interactive agents this definition is sufficient.

Personality (affect dispositions [Scherer 2005]) refers to generic and stable characteristics of a person in terms of behaviour, emotion and thought. Usually a person's personality is expressed as values on several traits. These personality traits are the result of large factor analysis studies of personality adjectives with the aim of expressing as much variation as possible in as little number of factors. Then such factor's are transformed into questionnaires and validated for measuring personality traits. A well known personality model that is used often in SIA's is the big five factor model consisting of Openness, Conscientiousness, Extraversion, Agreeableness, and Neurotism (aka OCEAN) [Goldberg 1990, McCrae and Costa 1987]. A newer, related personality model that puts more emphasis on emotional and relational factors by introducing a new factor Honesty-Humility is HEXACO [Lee and Ashton 2004].

10.1.1.2 Emotion

Now that we have defined the most commonly used terms related to affect, we move on to the three most influential perspectives on emotion in psychology that have influenced emotion modelling in SIA's: the categorical perspective; the dimensional-constructionist perspective; and, the cognitive appraisal perspective (for an excellent comparison of the fundamentals behind categorical versus dimensional perspectives see [Zachar and Ellis 2012]). Emotion can be studied from these different, complementary perspectives. Each of these bring unique insights

into what emotion is, and how emotions are produced and represented as psychological constructs in human minds. Furthermore, these perspectives offer opportunities but also present limitations to computational modelling of emotion. It is therefore important to understand these perspectives before modelling emotion in a SIA, because these perspectives ultimately shape what you can expect from (interaction with) your emotional agent.

When emotions are studied from a *categorical* perspective, an emotion is a specific multimodal response resulting from an assessment of the situation in terms of survival potential for the individual. All humans have the same survival needs, and many related animals as well. As a result, many emotions are similar in different individuals of one species, and probably even between species [Bekoff 2008, De Waal 2019]. The modalities of this reflex typically consist of an affective assessment of the situation (is it good or bad), a specific feeling (how does this feel), a specific action tendency (what do I do), and if evolution had a need for it, a typical expression pattern (how do I show this internal state). For example, anger is a negative feeling due to someone doing you harm. Anger has an associated tendency to act aggressively, a particular facial grimace and an approach posture. The categorical view emphasizes the evolutionary roots of emotion and its role in shaping behaviour and communication. Key historical theories that represent this view include Darwin's emotions as serviceable habits (see [Barrett 2011] for a critical analysis), Ekman's basic emotions [Ekman and Friesen 1971], and Frijda's action tendencies [Frijda 1988]. Jack [Jack et al. 2014] presents more recent work in this line, refining the notion of basic emotion categories into biologically plausible hierarchies by studying perception of computationally generated dynamic facial expressions. Indeed there is developmental evidence that emotional categories and the labelling thereof develops over time and becomes more refined when children grow older [Widen and Russell 2008]. The categorical perspective is useful when one is interested in communication, labelling of emotions, emotion specificity and embodied approaches.

When emotions are studied from a *dimensional* perspective, an emotion is the person's interpretation of currently felt core affect, where core affect is described in terms of affective dimensions [Russell 1980]. This relates to the constructionist view [Barrett 2005] that emphasizes that many important emotions that we experience are not related to any expression or action tendency, even though we do have words and feelings that clearly identify these emotions as specific mental constructs with an affective feeling. We learn to classify core affect, together with the context of it's emergence, just like we learn to classify colours or car brands. These affective dimensions typically include *valence* (aka pleasure) and *arousal* (not the same as emotional intensity), and sometimes a third dimension called *dominance* (related to motivational stance and social verticality [Mast and Hall 2017]). Valence refers to the positive and negative aspect of the emotion, arousal to the associated physiological activation, and dominance to the amount of influence and control the individual feels over the situation. Emotions are the labels we learn to attach to specific values of core affect together with the context. For example, sadness is a label we identify with a feeling of low valence, low arousal, and

low dominance, when something happens that is irreversible; while elation (extreme joy) is a label we identify with a feeling of high valence, high arousal and high dominance [Mehrabian 1980]. Affective dimensions can also be associated with other psychological constructs including moods, thoughts, opinions, and even representations of objects. It is important to keep in mind that while it is possible to describe an emotion in terms of its affective dimensions, these dimensions have low specificity and without knowing what a PAD value triplet represents in context it is hard to deduce what it means. For example, interpreting high valence, high arousal, and high dominance as elated is not necessarily correct; the individual might feel extreme pride instead. More contextual information is needed to find the 'correct' emotion label because many emotion labels map to similar PAD values. The dimensional perspective is useful in SIA's when one is interested in a common representation for different affective phenomena (e.g. when modelling emotion, mood and attitudes) or when emotional continuity is important (e.g. when modelling emotions that dynamically change from one into the other).

When emotions are studied from a *cognitive appraisal* perspective, an emotion is the result of the evaluation of the situation on a set of cognitive dimensions in light of the individual's concerns in order to motivate the individual into appropriate action [Arnold 1960, Ortony et al. 1988, Roseman and Smith 2001, Scherer 2001, Smith and Lazarus 1990, Van Reekum and Scherer 1997]. In short, emotion results from concern-based reasoning. Some evaluations are simple assessments of stimulus properties, e.g., the suddenness of a stimulus or the intrinsic pleasantness of that stimulus. Others are complex assessments of the consequences and causes of the stimulus, e.g., goal congruence and attribution of responsibility. However, the core of this view is that emotion is largely the result of a cognitive evaluation of the situation. As mentioned, this cognitive appraisal process is organized into different processes often referred as *appraisal dimensions*. For example, if a car is nearing me at great speed, this is a sudden stimulus that is of personal relevance, not conducive to my concern of survival, and I have limited control. This combination of appraisal dimension 'activations' (sudden, high relevance, low goal congruence, and low control) is typically associated with the emotion we would label as 'fear' [Scherer 2001]. Cognitive appraisal theory thus links cognitive processing to the elicitation of emotion. Note, however, that modern appraisal theory does not claim that all assessments are due to conscious reasoning, for example, when the taste of candy is assessed as pleasant. Cognitive appraisal theories are less concerned with exact emotion labelling of the appraisal outcome, nor are they with the resulting specific behaviour. Although many SIA researchers have interpreted appraisal as a goal/belief-derived reasoning process aimed at action planning (see e.g. [Gratch and Marsella 2004, Rosis et al. 2003]), this view has been advocated explicitly only recently by a subset of cognitive emotion theorists [Moors et al. 2017, Reisenzein 2009a]. The cognitive appraisal view is useful when one is interested in modelling the emotion elicitation process, but lacks a precise mapping between the appraisal and the resulting emotion label. An exception to the latter is the Ortony-Clore-

Collins model [Ortony et al. 1988] (see [Bartneck 2002] for an analysis of the model), explaining its popularity in SIA's [Popescu et al. 2014, Rosis et al. 2003] as well as formal modelling of emotion [Adam et al. 2009, Meyer 2006, Steunebrink et al. 2007]. The cognitive appraisal perspective is helpful when emotions are needed for agents that have a cognitive basis for their AI.

These three perspectives are complementary. Affective dimension-constructionist views give us a generic representation of affect for a wide variety of affective phenomena while emphasising a common emotional core and statistical categorization principles explaining individual emotional development and variation. Categorical emotion research brings us structure in affective expression and communication while emphasising the biological roots of emotions as coordinated behaviours to address immediate concerns. Cognitive appraisal theory brings us information processing mechanisms for the elicitation of emotion while emphasizing that specific emotions are elicited by thought processes that are mental, individual, and contextual. There are also important similarities. First of all, all views emphasize that emotion is about experiencing the positiveness and negativeness of a situation related to the well-being of the organism. All views therefore acknowledge that emotions at the core are about assessing the utility of the current situation with respect to survival of the individual. The valence dimension represents positive/negativeness in the dimensional view, categorical emotions are hierarchically structured around positive and negative emotions, and cognitive appraisal resolves around an affective evaluation related to personal concerns. Second, the different views reserve an important role for power, including social power. The dominance dimension represents the extend to which one influences or is influenced by the external environment (including others), an important aspect of categorical emotions is whether the emotion is an approach versus avoidance emotion, and many cognitive appraisal theories propose coping-related appraisal processes related to the perception of power and influence [Scherer 2001]. Finally, all views emphasize the importance of bodily activation. The arousal dimension represents the extend to which a stimulus, thought, relation, etc.. has associated bodily activation, categorical emotions are strongly tied to action and bodily activation through action tendencies [Frijda 1988], and cognitive appraisal theories have processes related to the urgency and novelty of stimuli (e.g. [Scherer 2001] that predict alertness of the individual.

10.1.1.3 Affect and Cognition

Modern psychology, neuroscience and computational modelling research strongly suggests that affective processing and cognitive processing are interdependent [Broekens 2018, Damasio 1994, Hoemann et al. 2019, Marsella and Gratch 2009, Moors et al. 2017, Reisenzein 2009a, Rolls 2014]. In fact, many emotion theorists nowadays suspect that a hard distinction between cognition and emotion is arbitrary and not helping in advancing our understanding of affect and cognition. However, here I will not go into that debate but simply list several well-known interactions between affect and cognition. First, mood influences information process-

ing in that positive moods typically favour high-level processing of information and creative problem solving while negative moods favour attention to detail and critical reflection [Dreisbach and Goschke 2004, Vosburg 1998], and certain information processing styles are more prone to influences of mood than others [Forgas 2000]. Second, memory recall is mood congruent [Matt et al. 1992]. Third, emotional processing needs to be intact for decision making [Damasio 1994] and affect influences how decisions are made, e.g. in negotiation [Broekens et al. 2010, Kleef et al. 2004]. Forth, memories with strong associated emotions are easier to remember [Reisberg and Hertel 2003], affect-based attitudes are relatively stable compared to cognition-based attitudes [Edwards 1990], and affect plays an important role in how attitudes [Maio et al. 2018] and judgments [Greifeneder et al. 2010] are formed. Many mechanisms have been proposed that might be responsible for this including current affect as a source of information about something, current affect that triggers associations with congruent attitudes, arousal as intensity measure for the importance of beliefs and memories, emotion processing as a way to value alternative outcomes, etc... It would go to far to review this field here, but I hope to have convinced you that cognition and emotion are intertwined.

10.1.2 Why do Socially Interactive Agents need Emotions?

Now that we have covered some background in affective science, I explain why SIAs, particularly those that need to interact with humans, need some form of artificial emotional intelligence [Picard 1997, Schuller and Schuller 2018]. Emotional Intelligence is defined as the ability to carry out accurate reasoning about emotions and the ability to use emotions and emotional knowledge to enhance thought [Mayer et al. 2008]. This ability should - for now - be considered a 'holy grail' for SIA research as this is still a long way to go. For this to be possible, many things need to be in place including proper recognition of emotion in humans, plausible emotion elicitation simulation and incorporation of affective information in the agent's information processing, and finally reliable emotion expression synthesis. Further, agent reasoning, machine learning and pattern recognition is needed as well for a solid understanding of the context.

However, why would we want this in the first place? In general, there are two main reasons for using affect in SIAs: expression of affect and recognition of affect can be used as a means to enhance SIAs' communication abilities, and, the modelling of affective processes can enhance the SIA's decision making abilities. This closely follows the function of emotion in humans. On an interpersonal level, emotion has a communicative function: the expression of an emotion is used to communicate social feedback as well as empathy (or distance) (see, e.g., [Fischer and Manstead 2008]). On an intra-personal level emotion has a reflective function [Oatley 2010]: emotions shape behaviour by providing feedback on past, current and future situations [Baumeister et al. 2007] as well as help to make important decisions [Damasio 1994].

Communication of affect is essential for the development of children [Buss and Kiel 2004, Chong et al. 2003, Klinnert 1984, Saint-georges et al. 2013]. If SIAs are implemented either as tutor/coach or as teachable robot or agent, then it is likely that both roles are difficult to fulfil without the use of affective processing in the SIA (See e.g., [Castellano et al. 2013, Heylen et al. 2003]). Indeed, evidence indicates that emotion helps for both roles [Broekens 2007, Leyzberg et al. 2011]. Communication of affect is also essential for the development of relationships and the building of trust [Weber et al. 2004], and the communication of empathy [Dimberg et al. 2011]. These are important elements in the building of rapport with a conversational agent (see chapters 11 and 12 on empathy and rapport). Indeed there is evidence that SIA's that express emotions and mood as part of their behaviour are perceived to be more empathic resulting in higher trust [Cramer et al. 2010] than those that do not, and, that emotions expressed by agents influence how the users respond to the agent [de Melo et al. 2011, 2014]. Also, there is a long line of research in emotional and sociable robots showing that users generally attribute all kinds of human abilities and characteristics to the robots (see, e.g., [Turkle et al. 2006])

Emotions are essential in decision making [Damasio 1994]. Classically, artificial reasoning and decision making is approached from the perspective of optimality: the process should give the best possible outcome given the input data/knowledge base. However, in many cases, reaching a 'good enough' solution, i.e., satisficing solution, is fine as well. Further, sometimes the goal is to represent human decision making, not to reach optimality per se [Baarslag et al. 2017]. Indeed, work on embedding emotions in decision making architectures clearly showed diverse benefits for reaching better solutions or good enough solutions faster [Belavkin 2001, Salichs and Malfaz 2012]. A long history of research into cognitive-affective architectures shows that emotions can play a useful role in agent learning, exploration and reasoning [Franklin and Graesser 1999, Franklin et al. 2014, Hogewoning et al. 2007, Marinier III and Laird 2004, Velasquez 1998] (see historical section for more references).

In short, emotions are in various was crucial for social interactive agents. We will next look at the computational methods to implement emotions in such agents.

10.2 Computational Models and Approaches

This section focuses on computational modelling of affect, excluding both expression synthesis (see Section II in this book, and Section II in [Calvo et al. 2014]) and automated affect detection (see Section II in [Calvo et al. 2014] and Section III in [Burgoon et al. 2017]). I discuss the most commonly used approaches to modelling emotion and directly related affective phenomena including mood, attitude, relation and personality. First, I cover computational *representations* of emotion, mood, attitude, relation and personality. These representations are the computational constructs maintaining the values of the affective variables. Then, I cover four approaches to the computational modelling of appraisal, i.e., *emotion elicitation*,

the mechanism responsible for simulating the values of artificial emotions when the SIA evaluates its situation. Finally, I cover four working examples, one for each approach.

10.2.1 Computational representations of affect

In this section I cover the most commonly used computational representation of the affective phenomena introduced in the definition sections. Note that SIA architectures do not need to implement all of these phenomena.

The *emotion* of the agent (aka affective state) is usually represented as a vector with intensities $e = [i_{E_1}, ..., i_{E_n}]$ for each represented emotion $E = \{E_1, ..., E_n\}$ [Kaptein et al. 2016, Ochs et al. 2009]. Vector elements typically represent categorical emotions or dimensions. For example, if an agent models the six Ekman emotions, then $E = \{joy, surprise, anger, fear, disgust, sadness\}$ with for example an emotional state equal to $e = [1, 0.5, 0, 0, 0, 0]$ denoting maximum intensity for joy, and half intensity for surprise. If the emotion is represented as emotion dimensions, then, for example $E = \{valence, arousal, dominance\}$ and the emotional state could be $e = [1, 1, 0]$. Note that when the agent needs to express or reason upon this state, there are many different ways to do this. For example, one could take $Express(Max(e))$ to express only the emotion with the highest intensity, or one could express the interpolation of the expressions based on all intensities in $ExpressInterpolate(e)$, if the expression/rendering allows this. For reasoning upon the state, similar choices need to be made. The emotional state is changed due to appraisals (covered in detail below). Appraisals eventually result in emotion intensities, which are integrated in the emotional state. In the below dynamics we assume an appraisal *sets* the emotional state, however, one can also *add* the appraisal to it and use a bound for the emotion intensities. Further, in the absence of appraisals the emotional state typically decays over time. So, the complete abstract dynamics for the emotional state can be written as follows:

$$e_{t+1} = a_t(situation) \vee \{e_t * \gamma\} \tag{10.1}$$

with $\gamma = [0, 1]$ and $a_t(situation)$ the outcome of the appraisal process at time t represented as an emotion vector with dimensions E.

The *mood* of the agent is also commonly represented as a vector with intensities $m = [i_{M_1}, ..., i_{M_n}]$ for each represented emotion $M = \{M_1, ..., M_n\}$. Moods are typically represented as a vector of dimensions [Jones and Sabouret 2013, Peña et al. 2011], but one also sees approaches with a categorical mood state. The mood state typically is a function of the history of the last n emotional states, so:

$$m_t = f(g(e_{t-n}), ..., g(e_t)) \tag{10.2}$$

Usually, $f()$ is some form of averaging, i.e., $f() = avg()$, with $g(x) \rightarrow x$, i.e., function $g()$ does nothing. Sometimes the mood representation is different form the emotion representation, for example when the emotional state is represented as a vector of basic emotion intensities

while the mood is represented as a vector of affective dimensions. In those cases, function $g()$ maps the emotional state to a different representation first. We will use $g()$ to denote a mapping function between affective dimensions throughout this chapter. For example, when $M = \{valence, arousal, dominance\}$ and the emotional state is again based on Ekman, $E = \{joy, surprise, anger, fear, disgust, sadness\}$, then a possible element from this mapping could be $g([1, 0.5, 0, 0, 0, 0]) \rightarrow [1, 1, 1]$. Such mappings are usually continuous functions based on findings from the literature (see e.g. the word-affect lists from Bradley and Lang [Bradley and Lang 2007] or Mehrabian [Mehrabian 1980]). Moods can also influence the emotional state. One of the more common ways is to have the emotional state decay to the mood state:

$$e_{t+1} = a_t(situation) \vee \{g^{-1}(m) * (1-\gamma) + e_t * \gamma\} \tag{10.3}$$

with $g^{-1}(m)$ the inverse mapping from the agent's mood representation to emotional state representation. Note that such inverse mapping is not trivial if the emotional state has a higher dimensionality, and requires a representation of the emotions in the lower dimensional mood space, for example as points representing prototype emotions, and a distance function defining the intensities [Breazeal 1998].

An *attitude* is usually represented as an association between a piece of affective information and a piece of knowledge (a belief, a state, an entry in a knowledge base, a 'chunk', a thought, an image, a word, etc...). The affective information typically originates from the emotional state or the appraisal process. Again, the actual representation varies depending on the affective dimensions used, but attitudes are commonly represented with the valence dimension alone. Computationally this means that an attitude $at_k = i$ is a tuple of an intensity i and a piece of knowledge k. Attitudes form due to the attribution of an appraisal or emotion to a piece of knowledge k. In order to do this, the agent appraises the current situation, potentially identifies salient aspects of the situation, and stores the result of the appraisal, or the resulting emotional state, as an association with the situation or one or more aspects. For example, if a virtual math coach identifies that the currently proposed exercise is too hard for a child, and the child reacts with anger, it could appraise based on the OCC model [Ortony et al. 1988] that it's own action is blameworthy resulting in the emotion of guilt, storing a negative attitude for that exercise $at_{exercise} = g([joy = 0, .., guilt = 1])$ with $g([joy = 0, .., guilt = 1]) = -1$, assuming that attitudes get mapped from the OCC emotions to a one dimensional valence representation by a mapping function $g()$ taking either an emotional state e or appraisal a as input [Jones and Sabouret 2013]. Also attitudes have dynamics, with a simple abstract form equal to:

$$at_{k,t+1} = g(e \vee a) * (1-\gamma) + at_{k,t} * \gamma \tag{10.4}$$

with t the time dimension, usually moments of attributions so that attitudes will be averaged over attributions, and γ another discount factor. Attitudes can also be stored as fuzzy or probabilistic links between beliefs and emotions in e.g. a Belief Net.

When it comes to computational representations of *relations* we have to distinguish relations from social emotions. Relations are typically represented as $r_{agent} = i$ with *agent* being the agent and i being again the intensity of the relation on some affective dimensions (see e.g. [Ochs et al. 2009]). The rest is similar to attitudes: relations develop over time due to emotions or appraisals and can be represented on different dimensions than the emotional state. Social emotions are emotions that exist by the virtue of relations and other agents, are attributed *to* other agents, and influence relations with other agents. To highlight the difference: one can have a positive relation with someone and feel proud of, envious, angry at, disappointed with, or thankful for that person. This might influence the relation (following the attribution as explained above), but it is something different. To give a concrete example: if I have a positive relation with Peter $r_{Peter,t} = 1$, and Peter does something that makes me feel disappointed, then following Equation 10.4 my relation could become:

$$\begin{aligned} r_{Peter,t+1} &= g([..., disappointment = 1, ...]) * 0.2 + r_{Peter,t} * 0.8) \\ r_{Peter,t+1} &= -1 * 0.2 + 1 * 0.8) \\ r_{Peter,t+1} &= 0.6 \end{aligned} \tag{10.5}$$

with $g([..., disappointment = 1, ...]) = -1$ and $\gamma = 0.8$. If Peter repeatedly disappoints me, my relation towards him will gradually move to -1 with a speed depending on γ.

Personality is usually represented as a vector of personality trait values $p = [i_{T_1}, ..., i_{T_n}]$, with $T = T_1, ..., T_n$ being the traits, e.g. based on OCEAN [Goldberg 1990]. Contrary to the above affective constructs, the personality is assumed to be stable, so once set, p does not change for an individual agent. There are many ways in which personality can influence the emotional state, including changing the way information is processed, changing the sensitivity of particular emotions by having e.g. an emotion-specific personality-dependent γ, changing the weight for particular appraisal processes in the calculation of the appraisal of the current situation, and using the personality as a default agent mood [Gebhard 2005]. In most cases some static mapping $g(p)$ is introduced linking p to the factors that change the emotional outcome (appraisal weights, emotion sensitivity) [Jones and Sabouret 2013].

10.2.2 Emotion elicitation

The core of an emotion is the assessment of personal relevance of a situation, thereby in some way providing feedback on the suitability of past, current, or future behaviour [Baumeister et al. 2007, Broekens et al. 2013, Moors et al. 2013, Van Reekum and Scherer 1997]. As mentioned, in this chapter I will use the term *appraisal* for this process of assessment. An emotion occurs when something happens that is personally meaningful to the agent. In this

chapter, I cover four approaches towards computational modelling of appraisal: cognitive-agent based modelling, embodied modelling, reinforcement-learning modelling, and hard-wired appraisal. Each approach is based on different modelling principles, in particular with respect to the concept of *goal*, and the representation of the agent-environment relation. Cognitive-agent based appraisal models use some form of cognitive agent formalism - such as BDI logic or utility-based planning - to represent the agent-environment relation, with an explicit representation of a goal that can be used in planning and reasoning (for review see [Gratch and Marsella 2014]). Embodied modelling uses a goal representation derived from homeostasis, and places less emphasis on symbolic processing to assess the agent-environment relation (for a recent treatise see [Canamero 2019]). Reinforcement learning-based modelling proposes that goals are derived from some value or reward signal, and the agent-environment relation is based on (models of) state-action-reward-state-action sequences (see [Moerland et al. 2018] for a recent review). Hard-wired appraisal directly or indirectly encodes the appraisal outcome in the stimuli perceived by the agent. Although each of these approaches are different, they all share a utility-based view of appraisal. Therefore, from a theoretical perspective all of these approaches are related to the psychological concept of appraisal. There is always some agent's need and some form of discrepancy or distance between the current state and that need. Emotion is derived from this discrepancy and the intensity of the need.

10.2.2.1 Cognitive-agent based modelling

Cognitive-agent based appraisal models are based on cognitive theories of emotion. In *cognitive appraisal theory* emotion is often defined as a valenced reaction resulting from the assessment of personal relevance of an event [Moors et al. 2013, Ortony et al. 1988, Scherer 2001]. The assessment is based on what the agent believes to be true and what it aims to achieve as well as its perspective on what is desirable for others. The basis is that a collection of computational processes analyse the current situation in terms of desirability for and impact on the agent [Dias and Paiva 2005, Marsella and Gratch 2009, Popescu et al. 2014, Steunebrink et al. 2007]. Cognitive appraisal models in SIA's mostly come in two flavours: those developed from theories that propose specific appraisals to assess the affective outcome of stimuli, we refer to this as *stimulus appraisal*, and, those developed from theories that propose that emotions result from belief-desire structures, we refer to this as *belief-desire theory of emotion* or BDTE.

We first cover *stimulus appraisal*. Each appraisal process is responsible for a particular aspect of the analysis of the stimulus and together they result in an emotion. Most computational models are based on Ortony Clore and Collins' model [Ortony et al. 1988]. This OCC model proposes a *tree-structure of evaluations, resulting in one or more specific emotions for a given situation*. The appraisal process is described in abstract computational terms including

statements such as:

$$if\ desirable(event) \wedge approveAction(otherAgent) \rightarrow emotion(gratitude) \tag{10.6}$$

Although this does not tell us how to calculate $desirable(event)$ or $approveAction(otherAgent)$ or the intensity of $emotion(gratitude)$, it does give a precise structure to implement appraisal rules and the resulting emotions. Many computational models use this structure as a guideline for the emotion elicitation process of virtual agents and robots. Further, due to this clarity it facilitates selecting an expression when interacting with humans. Computational models need some appraisal logic to decide how to implement the appraisal processes and the emotional intensities. For example based on modal logic, desirable can be expressed as accomplishing a goal:

$$if\ \kappa \in K \wedge \kappa = true \rightarrow desirable(\kappa) \tag{10.7}$$

This still needs to be extended with proper intensities and a logic for consequences of events, see e.g., [Steunebrink et al. 2008] for more detail on such an approach. Other approaches are more agent-logic agnostic and focus on the appraisal framework ([Dias et al. 2014, Jiang et al. 2007, Ochs et al. 2009, Popescu et al. 2014]), or use a fuzzy-logic approach based on OCC to determine the emotional state [El-Nasr et al. 2000]. Several models are available as open source appraisal engines, including FATIMA [Dias et al. 2014] and GAMYGDALA [Popescu et al. 2014].

A second influential stimulus appraisal idea is that *appraisal processes evaluate stimuli in order to motivate appropriate behaviour, with a looser connection to the specific emotion that results from those processes*. This can be found in Smith and Lazarus appraisal theory as well as in Scherer's Stimulus Evaluation Checks (SEC). Both propose that situations are checked by specific appraisal processes, in line with OCC, but the appraisals are different and follow a different process. SEC, for example, proposes that simple appraisals assess the stimulus first, including relevance and pleasantness of that stimulus, after which more complex processes kick in, including goal congruence and even later coping. Computational models based on SEC follow similar lines as those based on OCC, namely, they have to select and implement specific appraisals, which together give an indication of the resulting emotion. For example, when a stimulus is sudden and unpleasant, fear is likely to be the emotional result. A computational model will now have to implement

$$if\ sudden(event) \wedge unpleasant(event) \rightarrow emotion(fear) \tag{10.8}$$

While Scherer's model goes into quite some depth on the appraisal processes, the link between appraisal activation and emotion is less clear. Therefore it is harder for the agent designer to decide based on this theory what specific emotion comes out of the reasoning process. This is not problematic when interested in simulating appraisal (such as in the work by Marinier and Laird [Marinier III and Laird 2004]), but may become a problem if

clear emotion signals need to be send to the user of the SIA. A well-known computational model that is inspired by this idea of sequential checking in light of an organisms adaptation and functioning is EMA [Marsella and Gratch 2009]. EMA implements appraisals as fast automatic and parallel evaluators of the current cognitive state. As such the appraisal has no causal influence (although the information may be used for coping later on) but provides a moment-to-moment *affective summary* of the situation, an idea that also resonates with the ideas of BDTE discussed next.

We now cover the cognitive-motivational view also known as *belief-desire theory of emotion* (BDTE). BDTE assumes that appraisal of beliefs and desires, rather than stimuli, is the core of what emotions are. The key difference is that BDTE proposes that *emotions result from an assessment of the current belief-desire structure of an agent*, while previous theories propose that perceived stimuli are assessed with a set of appraisal processes in the context of desires of the agent. There are subtle differences in the psychological and philosophical underpinnings and ramifications of different BDTE approaches (see [Reisenzein 2009b]), but from a SIA engineer's perspective all BDTE approaches place important emphasis on the concepts of beliefs and desires (goal states in particular). To explain this view we take Reisenzein's cognitive model as example. In this model there are two core appraisal processes: belief-belief and belief-desire congruence. Reisenzein proposes that these processes are sufficient to explain eight basic emotions [Reisenzein 2009a]. For example, joy results from the belief that a state *s* is true and desired (i.e., *s* is a goal state). Fear would result from the belief that a state *s* becomes more true but not entirely, and *s* is not desired. And so on...

In practice, though, the difference between BDTE models and stimulus appraisal models is not so big from a computational point of view. In both cases, the computational model needs to explicate the appraisal processes and this is usually done with a form of utility planning/goal-based agent formalism. For example, while Steunebrink uses OCC as basis [Steunebrink et al. 2007] and Kaptein uses Reisenzein's BDTE [Kaptein et al. 2016], both use similar agent logic to decide upon the desirability of events (whether or not a belief helps achieving a desired goal, that is). Further, in both approaches, typical agent implementations will not reason upon just a holistic state representation, but instead will reason over beliefs, plans and goals [Castelfranchi, Kaptein et al. 2016, Marsella and Gratch 2009, Meyer 2006, Reilly 1996, Steunebrink et al. 2007]. For example, if a particular belief brings closer (e.g. in terms of time, or effort of the agent) a particular goal, then hope could result. As the final actions of these agents are often also informed by this same process of reasoning, action and emotion are in line, which is consistent with current cognitive views on emotion [Moors et al. 2017].

10.2.2.2 Embodied Models

In theories that emphasize biology, behaviour, and evolutionary benefit [Frijda 2004, Panksepp 1982], or core-affect [Russell 1980] the emotion is more directly related to ac-

tion selection, the body, hormones, biological drives and particular behaviours but the core of the appraisal is similar: an assessment of harm versus benefit resulting in action aimed at adapting the behaviour of the agent. Computationally, *Embodied models* of emotion emphasize that agents have drives, limited decision making resources, an artificial body, and an action selection problem. Core in most of these models is that at some point the agent needs to select an action in real time and this action needs to be consistent with the need to keep a set of homeostatic variables in check [Arkin et al. 2003, Cañamero 2003, Cos et al. 2013]. This process of homeostasis is the basis for the *goal* representations. For example, the emotion from the homeostasis process may be used as additional - or modulation of the - reward signal (e.g., [Cos et al. 2013] motivated actor-critic approach). Such emotion models are thus implemented on top of *homeostatic machines* [Man and Damasio 2019]. Most implementations of such models are used in studying how robots can solve relatively simple resource gathering-like tasks [Avila-Garcia and Canamero 2005, Kiryazov et al. 2013], although these principles can be applied to human-robot and human agent interaction as well [Arkin et al. 2003, Breazeal 1998, Verdijk et al. 2015].

An important concept in embodied models of emotion is *grounding*: emotion is emerging from the organism's assessment of (and is functionally meaningful to) its bodily state and well-being. For example, fear is the anticipation of bodily harm, resulting in avoidance behaviour. If a robot has sensors to detect body integrity (which is a homeostatic variable it wants to keep up), and it has an association between a certain stimulus and a decrease in body integrity, then the perception that body integrity is anticipated to drop triggers a drive to do something about that, eventually resulting in an action to move away from the stimulus. Notice that in this process we could add that fear is triggered with an intensity equal to the predicted drop in body integrity, but in this embodied example this does not add much to the whole process. Indeed, in many embodied approaches the emotions are considered emergent phenomena, consisting of the collective activation of processes including affect grounded in the robots body and activity to meet robotic needs. As such, they fit well with core-affect and constructionist views as well. For example in the work by Kiryazov [Kiryazov et al. 2013], arousal is a representation of the robots electrical energy processes. In the work by Avila-Garcia and Canamero [Avila-Garcia and Canamero 2005] the 'emotion' of fear can be observed when the robot's subsystems trigger behaviour to avoid a competitor robot in a resource gathering task when in a high-risk health state.

10.2.2.3 Reinforcement Learning Models

Most *reinforcement learning models* of emotion are in essence cognitive appraisal theories implemented on top of the reinforcement learning paradigm. With RL an agent tries to solve a Markov Decision Problem by effective exploration, receiving after each action it takes as only feedback the reward and the next state it arrives in [Kaelbling et al. 1996, Sutton and Barto 2018]. The goal of the agent is to learn an action selection policy that will maximize

utility, expressed as the sum of future rewards. The reward is a scalar $R(s,a)$, the utility of a state is expressed as the value $V(s)$, the value of an action is $Q(s,a)$ and the MDP model is usually represented as conditional transition probabilities $T(s'|s,a)$. RL models of emotion have been extensively surveyed in [Moerland et al. 2018]. Here we summarize two of the four main approaches towards emotion elicitation (not how the simulated emotion is subsequently used). The other two are very similar to either embodied modelling or hard-wired appraisal.

In the first approach the agent learns and acts in the environment and the emotions are derived from the reward, the value function or the temporal difference signal (the update to the value of the state, see below in the Examples section). The default assumption is similar to that of cognitive appraisal theory, namely, that affect is related to an assessment of utility. First ideas emerged as early as the 1980's with the work of Bosinovski [Bozinovski 1982] interpreting the state-value as the emotion associated to that state. Also in [Broekens et al. 2015], the value of the state is used as a signal for fear and hope, while in [Moerland et al. 2016] the temporal difference signal is taken as basis for the simulation of joy and distress, hope and fear. Salichs and Malfaz [Salichs and Malfaz 2012] model fear for a particular state as the worst historical Q-value associated with that state (remembering a particular bad situation that it should fear). Other approaches compute a mood-like signal from normalized averages of rewards over time [Hogewoning et al. 2007, Schweighofer and Doya 2003].

In the second approach, the agent appraises situations based on its model and environment states. Typically, appraisal processes are implemented based on a cognitive appraisal theory that take state and model information as input, and output either emotion intensities or appraisal intensities (e.g., novelty, desirability, etc.). These emotions can then be used as meta-learning parameters or as additional reward signals. Key approaches include Marinier and Lairds approach [Marinier and Laird 2008] and Sequeira's [Sequeira et al. 2014], both based on Scherer [Scherer 2001]. Notice that when it comes to emotion elicitation, this approach is in fact a cognitive appraisal based approach, albeit using RL state and model as input.

10.2.2.4 Hard-wired appraisal

Finally we briefly cover an emotion elicitation approach that I would refer to as *hard-wired appraisal*. Here, events or environmental stimuli have a predetermined meaning in terms of the appraisal processes or even in terms of emotion. For example, for the stimulus *snake* there would be a predetermined emotional outcome $fear(snake) = 1$. These primary emotional responses are often related to the *low route* of LeDoux's emotion processing proposal, whereby primary emotions are evolutionary shaped complex responses [LeDoux 1996]. Secondary emotions are more difficult to simulate because each event has a predetermined emotional meaning, and for these secondary emotions cognitive processing is assumed to be needed. To add some flexibility, some approaches do not directly annotate events with emotional or appraisal consequences but indirectly using the input needed for the appraisal process. For example, in both [Ochs et al. 2009] and [Popescu et al. 2014] the events in a simulation can be

annotated with appraisal-relevant information including to which goal the event is contributing and the likelihood of the event being true, after which the 'black-box' appraisal engine will interpret the event and compute emotional consequence and effects on e.g. relations between agents.

10.2.3 Examples of cognitive-affective architectures

In this section, we will go through the design of four fictional cognitive-affective architectures for a socially interactive agent inspired by the four modelling approaches just described. We focus on the appraisal (not the expression). It is important to keep in mind that emotions are added to an agent for a particular purpose. This can be theoretical exploration, but also practical purposes such as enabling a robot to simulate emotions to children. These design goals change the way the model is developed and evaluated.

10.2.3.1 Cognitive-agent based appraisal

Assume we want a virtual agent that can help tutor children with math problems. Inspired by [Castellano et al. 2013, D'Mello et al. 2007], we assume the children want to learn and we assume the robot is empathic, i.e., it's own emotions will mimic those of the child. As such we assume the robot has the same goal as the child, namely, $goal(understand(X))$. The robot can give exercises to the child in the form of actions pushed to a tablet interface $action(exercise(X,nr))$, and perceive the answers $percept(answer(X,nr))$ pushing $answer(X,nr)$ in the belief base of the agent. It further has a knowledge base of correct answers $correct(X,nr)$ and some rules stating that:

$$if\ answer(X,nr) = correct(X,nr) \rightarrow likelyhood(X,l+1)\ else\ likelyhood(X,l-1) \quad (10.9)$$

$$if\ likelyhood(X) > 10 \rightarrow understand(X) \quad (10.10)$$

$$if\ answer(X,nr) \wedge \neg understand(X) \rightarrow action(exercise(X,nr+1)) \quad (10.11)$$

which keeps pushing actions. Granted, this is a simple agent, but it will start pushing actions as long as the child does not understand a particular goal X, which we add in the following way to the goal base $goal(understand(fractions))$.

Now we implement a simple emotion model based on Reisenzein's ideas that emotions are belief-belief and belief-desire comparators. In fact we cheat a bit because one of the comparators, the likelihood of the goal being true, is build in the logic in the form of the $likelyhood(X)$ predicate. We can now express the appraisal process for joy as follows:

$$if\ understand(X) \wedge goal(understand(X)) \rightarrow joy(X) \quad (10.12)$$

The agent is happy when it believes the child reaches the learning goal X. Now this is not a very interactive agent, and it would be helpful to also express some hope when the child is

doing a good job. For this the agent needs to know if the situation improved or not. We add:

$$if\ answer(X,nr) = correct(X,nr) \rightarrow improved(X) \tag{10.13}$$

Now we can simply appraise this as follows:

$$if\ improved(X) \wedge goal(understand(X)) \rightarrow hope(X) \tag{10.14}$$

The agent is hopeful (and of course expresses this to the child) when improvement is made. I leave the formalisation of distress and fear as well as a an actual working simulation of this system as an exercise to the reader.

10.2.3.2 Embodied appraisal

Assume we want to investigate the relation between resource gathering, survival and emotions. Inspired by [Avila-Garcia and Canamero 2005, Kiryazov et al. 2013], consider the following homeostatic machine (animat) with homeostatic variables $H = \{hunger, thirst\}$, behavioural urges $B = \{search, drink, eat\}$, potential stimuli $S = \{food, water\}$ and drives

$$\begin{aligned} D_{eat} &= (hunger * food), \\ D_{drink} &= (thirst * water), \\ D_{search} &= (hunger * thirst) \end{aligned} \tag{10.15}$$

where we assume that the intensity of the drive D_b equals the behavioural urge B and behaviour selection is based on $b_{active} = argmax(B)$. The problem this animat needs to solve is how to survive by keeping hunger and thirst low, while food and water are scattered around the world. It needs to solve an action selection problem and balance searching, eating and drinking. If hunger or thirst goes up, search will be triggered and will become the $argmax(B)$ resulting in searching behaviour. When food is found, this triggers eating. When hunger is lower again, the animat will start searching because thirst is triggering search behavior but the hunger is gone due to eating. When finding water it will start drinking, lowering thirst after which it will start searching again and so on. Now what could emotion be in this system?

Emotion can be simulated in different ways (the following are examples). First we can assume a *hedonic* approach, and interpret the homeostatic state as pleasure, i.e. $pleasure = 1 - (avg(H))$. This means that whenever the animat is doing well in terms of its homeostatic variables, it is also feeling good reflecting the idea of 'core affect' [Russell 1980]. Second, we can assume an *emotion as feedback* approach, and interpret changes to the homeostatic state as signals of joy and distress, i.e. $pleasure = \delta(avg(H))$. This means that whenever something happens that moves the homeostatic variables in the desired direction, the animat will feel good (and vice versa), reflecting the idea that emotions are abstract feedback signals about the appropriateness of actions for the individual's well-being [Baumeister et al. 2007]. With respect to arousal, there are also different choices to be made. For example, we can interpret

the overall behavioural urge as arousal, i.e. $arousal = avg(B)$, reflecting the idea that arousal is related to (preparation of) physiological activity [Frijda 2004, Russell 1980]. Second, we can interpret arousal in a more holistic way such that arousal is 'all activity including information processing', i.e. $arousal = avg(B,S)$, presence of stimuli also increases arousal.

Let's pick an emotion as feedback and physiological activity approach. This means that activity is linked to arousal, and changes in the homeostatic state are linked to pleasure. If the animat is hungry, it will search and have high drives for eating. In the absence of food, the animat will feel aroused and on top of that it feels displeasure every time $avg(H)$ decreases. When food is found, it will switch to eating behaviour. The animat will still feel aroused (nothing changed there yet), but will feel pleasure due to the first bite of food reducing the hunger drive. While eating, hunger goes down and eventually vanishes. At this point, the animat will either stay there or start wandering around a bit, feels low arousal and neutral pleasure (no changes). Emotionally we can thus observe the following: high arousal and displeasure when hungry and searching (fear?); high arousal and pleasure when food is found (excitement?); low arousal and neutral pleasure when finished eating (relaxed?). It is left as an exercise to the reader to implement this, perhaps use the emotion as an actual feedback signal for the agent, and define additional evaluation criteria.

10.2.3.3 Reinforcement Learning appraisal

Assume we want an adaptive household service robot able to communicate to us the extend to which the learning process is converging and whether or not consequences of events were anticipated. Inspired by [Broekens and Chetouani 2019, Moerland et al. 2016, Thrun et al. 1999], consider a Reinforcement Learning service robot. The robot receives rewards when the user praises the robot and for the amount of dust and dirt it collects. While it is learning, it experiences TD errors and updates $Q(s,a)$ accordingly. Temporal difference errors are interpreted as signals of joy and distress [Broekens 2018]. For Q-learning this would mean that Joy and Distress are defined as follows:

$$if\ (TD > 0) \rightarrow Joy = TD \tag{10.16}$$

$$if\ (TD < 0) \rightarrow Distress = TD \tag{10.17}$$

With the TD error defined in the standard way for Q-learning:

$$TD = r + \gamma \max_{a'} Q(s',a') - Q(s,a)_{old} \tag{10.18}$$

In terms of additional emotion dynamics, whenever an emotion is triggered, it is added to the current emotional state intensity for that emotion using a logarithmic function with decay [Reilly 2006] to not saturate the emotion but keep gain at low intensities and allow for decay over time. At every point in time the agent thus has a vector $E = [i_{joy}, i_{distress}]$. It

expresses this vector continuously. While learning the particular tasks in a household it will experience positive and negative TD's expressing joy and distress to the user. By the time the tasks are converted to known RL policies, the robot will become emotionally neutral (and thus show that the learning has converged) and only express emotions when TD's occur due to unexpected outcomes. It is left to the reader to implement a simulation of this, and again, define additional evaluation criteria.

10.2.3.4 Hard-wired appraisal

The last example is simple. Inspired by [Ochs et al. 2009, Popescu et al. 2014], assume we want a Non-Play-Character, a villager, in a video game to be able to simulate rudimentary emotions based on events in the game. If the villager can perceive the events $S = \{player\_near, monster\_near, gold\_stolen, thief\_near\}$, and has the following emotions $E = \{joy, fear, sadness, anger\}$, then we can annotate the events as appraisals as follows:

$$\begin{aligned} A = \{&a(player\_near) = [1,0,0,0], a(monster\_near) = [0,1,0,0], \\ &a(gold\_stolen) = [0,0,1,0], a(thief\_near) = [0,0,0,1]\} \end{aligned} \tag{10.19}$$

Upon perceiving an event, the emotional state can be updated according the equation (10.1). When more flexibility is needed, one can annotate the event with input needed for the appraisal process instead, for example:

$$\begin{aligned} likelihood(gold\_stolen) &= 0.5, \\ conduciveness_{get\_rich}(gold\_stolen) &= -1 \end{aligned} \tag{10.20}$$

This can now be fed into an appraisal model and leave the emotion calculation to the model, in the spirit of [Ochs et al. 2009, Popescu et al. 2014]. We leave playing around with this as an exercise to the reader.

10.3 History / Overview

In this section I give a short overview of the history of computational modelling of emotion, including important milestones that influenced the different approaches introduced in the previous section. It will be a brief history, enhanced with some recent work from the last 5 years. For more history on this field including an excellent taxonomy of models pre 2014, readers are referred to [Gratch and Marsella 2014], while readers should look at [Pfeifer 1988] for a review of the treatment of emotion and affect in computer models before the field of emotion modelling existed.

10.3.1 The early period

The computational study of emotion was initiated by the cognitive revolution in psychology. Computational study of emotion was for the first time explicitly suggested in the early 1980's by Rolf Pfeifer [Pfeifer 1982], and around the same time by Aaron Sloman and

Monica Croucher who wrote the influential paper *why robots will have emotions* [Sloman and Croucher 1981]. In the late 1980's psychologists such as Nico Frijda, with his student Jaap Swagerman [Frijda and Swagerman 1987] started formalising Frijda's action tendencies theory and around that same time the influential OCC model was developed [Ortony et al. 1988] by Andrew Ortony, Gerald Clore and Allan Collins. These developments spurred agent-oriented research into emotion simulation resulting in the famous work by Clark Elliot, the affective reasoner [Elliott 1992], which was the first full blown cognitive appraisal-based implementation of the OCC model using goal-based agent reasoning. Emotion simulation work soon began to be applied in intelligent virtual agents, for example in the work on believable agents in the Oz project by Scott Reilly and Joseph Bates [Reilly 1996, Reilly and Bates 1992]. Fuelled by Damasio's ideas on the importance of emotion on decision making, emotions also made their appearance in the first social robots such as Kismet [Breazeal 1998] of which the emotional system was based on Velasquez's work on modelling emotion-based decision making [Velasquez 1998].

10.3.2 The diversification period

During the 2000's a surge in interest was seen in trying to understand the role of emotion in interaction with agents [Conati et al. 2005, Hudlicka 2003, Paiva 2000]. This was not in the least due to the book *affective computing* by Rosalind Picard [Picard 1997] in 1997, who for the first time defined emotion modelling as part of a field. Simulated emotions were added to social robots (iCat) and virtual agents (Greta, Steve) and applied to different settings including pedagogical agents [Gratch and Marsella 2001], negotiation [Core et al. 2006], game characters [Ochs et al. 2009], and human-robot interaction [Leite et al. 2008]. We see the development of Virtual Humans (now we would call these Intelligent Virtual Agents, or SIA's) that included emotions in their reasoning, their decision making and expressive repertoire [Allbeck and Badler 2002, Gratch et al. 2002], Also we see that agent researchers started to look at how to structure the appraisal process based on different formalisms including planning [Gratch and Marsella 2004], BDI logic [Meyer 2006] and set theoretic approaches [Broekens 2007] as well as how to embed emotions in complete cognitive agent architectures [Dias and Paiva 2005, Hudlicka 2005, Marinier III and Laird 2004]. We also see modelling emotion in relation to complete agents with personality, mood and expression [Rosis et al. 2003]. This was especially seen in embodied conversational agents [Egges et al. 2004]. Other appraisal theories, including Scherer's, are also being modelled [Broekens 2007, Marinier and Laird 2008]. In parallel during the 2000's, different approaches to emotion modelling appeared that were not emphasizing the cognitive appraisal but the interplay between emotion and cognition in agents as well as more embodied (cybernetic) approaches [Belavkin 2001, Cañamero 2001] as well as first attempts at linking emotion to reinforcement learning and metalearning and optimization [Hogewoning et al. 2007, Schweighofer and Doya 2003].

However, most of the work remained focussed on computational investigation and application of cognitive appraisal and in particular the OCC model in interactive agents.

10.3.3 Current work

By the end of the 2000's it became clear that the field needed to focus more on the evaluation of models of emotion [Broekens et al. 2013, Gratch et al. 2009]. This shifts the focus to the question why emotions were added in the first place. We see that validity (is the emotion theoretically valid) and user experience (how does the model impact the human in the loop) become important evaluation criteria. With regards to user experience, we see, for example, a focus on applying emotions in robots and agents for specific reasons including robot empathy [Paiva et al. 2017] in human robot interaction (See also Chapter 11), the building of rapport between SIA and human (see Chapter 12), enhancing non-player character flexibility in games [Chowanda et al. 2016], for review see [Yannakakis and Paiva 2014] and Chapter 27, and enhancing cognitive-assistive technologies [Robillard and Hoey 2018]. With regards to validity, we see approaches that focus on studying particular theoretical aspects of artificial emotion elicitation, such as how can emotions result from temporal difference reinforcement learning [Broekens and Dai 2019, Moerland et al. 2016], how can appraisal be modelled on top of reinforcement learning state-value information [Sequeira et al. 2014], how can we ground emotion in robot physiology [Kiryazov et al. 2013, Lowe et al. 2016] and interaction with robots [Jung 2017], how can appraisal be conceptualized as an iterative affective summary process [Marsella and Gratch 2009], and how can Damasio's as-is loop be simulated using a dynamical systems approach [Bosse et al. 2008]. Finally we see a large body of research working towards integrating emotion and other affective phenomena such as personality, relations as well as user emotions (user modelling) into the decisions-making process of SIA's.

10.4 Similarities and Differences in IVAs and SRs

The main approach for emotion elicitation in Socially Interactive Agents is cognitive appraisal, although models of emotions in social robots tend to also investigate embodied approaches (such as grounding affective dimensions in robot's physiology). Further, in both fields emotions are often used in interaction with people. Other chapters go into more detail on this (such as Chapter 8, 11, and 12). Overall, the two fields are relatively well aware of what each other does when it comes to modelling emotion elicitation through cognitive appraisal. RL-based and embodied modelling are more seen in Robots and more theoretical agent simulation studies that do not involve interaction but mainly task-based agent learning or adaptation scenarios. These modelling approaches are at this point still more theoretical in nature.

10.5 Current Challenges and Future Directions

In this section we discuss seven (somewhat arbitrary) challenges in the modelling of emotion. We discuss challenges originating from the core of emotion simulation and challenges in applying emotions in SIA's. First, it is still not clear how to select the appropriate frame of reasoning for generating emotions based on cognitive appraisal theory. As most theories assume emotions are triggered by the appraisal of a stimulus or appraisal of the belief-desire structure, the question in AI agents remains, *what goals do I take into account, and do I focus on the hope or fear side of things*. Many emotions may result from a single percept, and it is not clear if all of these arise, only the strongest, or only the last, etc.... It remains to be seen if this issue can be solved at all. Second, the intensity of emotions stays a difficult issue. Up until now there is no widely accepted model for emotion intensity based on appraisal theoretic simulations. Third, how can we ground and user-test emotions in other architectures than classical cognitive agents, including adaptive agents and simple robots? Will we be stuck with human perception studies only, or is there more to be done, e.g. replicating animal studies of emotion? Fourth, how to incorporate recent evidence that emotions follow a hierarchy with little basic emotions [Jack et al. 2014]. Should this perspective change the way we develop computational appraisal models? Fifth, we need to test the plausibility and effects of SIA emotion, as generated by an elicitation process, in non-trivial and longer-term interaction domains. For example, the effect of an emotional agent that is always supportive and empathic might be counterproductive in the long run and raise frustration. Sixth, there is still a lack of standard benchmarks for testing (often quite complex) emotion models. For this, human-agent negotiation can be a good basis as this involve many aspects of emotion including reactive emotions, appraisal, utility, norms, values, and strategic use if emotions [Gratch et al. 2015]. Related to this is the fact that there are many models that implement emotion, mood, personality and relations, but there is in fact no way to test and compare these complex models. We need simple interaction effects between affective phenomena to be replicated (such as the impact of mood on emotion and vice versa), including what this might bring to the user in terms of experience. Seventh, emotions are used by humans to explain their point of view and perspective. In AI there is the potential to investigate the use of emotion modelling in explanation, transparency, ethics and simulated emotions. Emotional expression grounded in the decision making process of the agent could be a form of transparency of the SIA's functioning [Broekens and Chetouani 2019, Kaptein et al. 2017].

10.6 Summary

We have covered important affective concepts including emotion, mood, attitude, personality and relation. We have covered how these concepts are computationally represented. Then we covered four approaches to the simulation of appraisal in SIA's and given practical working

examples of these approaches. Finally, we surveyed the history of the field and pointed out current challenges in emotion modelling.

Bibliography

C. Adam, A. Herzig, and D. Longin. 2009. A logical formalization of the occ theory of emotions. *Synthese*, 168(2): 201–248. ISSN 1573-0964. https://doi.org/10.1007/s11229-009-9460-9. DOI: 10.1007/s11229-009-9460-9.

J. Allbeck and N. Badler. 2002. Toward representing agent behaviors modified by personality and emotion. *Embodied Conversational Agents at AAMAS*, 2: 15–19.

R. C. Arkin, M. Fujita, T. Takagi, and R. Hasegawa. 2003. An ethological and emotional basis for human–robot interaction. *Robotics and Autonomous Systems*, 42(3-4): 191–201. ISSN 0921-8890.

M. B. Arnold. 1960. *Emotion and personality. Vol. I: Psychological aspects*. New York: Columbia University Press.

O. Avila-Garcia and L. Canamero. 2005. *Hormonal modulation of perception in motivation-based action selection architectures*. SSAISB.

T. Baarslag, M. Kaisers, E. Gerding, C. M. Jonker, and J. Gratch. August 2017. When will negotiation agents be able to represent us? the challenges and opportunities for autonomous negotiators. In C. Sierra, ed., *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pp. 4684–4690. International Joint Conferences on Artificial Intelligence.

L. F. Barrett. 2005. *Feeling Is Perceiving: Core Affect and Conceptualization in the Experience of Emotion*, pp. 255–284. Guilford Press, New York, NY, US. ISBN 1-59385-188-X (Hardcover).

L. F. Barrett. 2011. Was darwin wrong about emotional expressions? *Current Directions in Psychological Science*, 20(6): 400–406.

L. F. Barrett, B. Mesquita, K. N. Ochsner, and J. J. Gross. 2007. The experience of emotion. *Annual Review of Psychology*, 58(1): 373–403. DOI: doi:10.1146/annurev.psych.58.110405.085709.

C. Bartneck. 2002. *Integrating the occ model of emotions in embodied characters*, p. 39–48.

R. F. Baumeister, K. D. Vohs, and C. Nathan DeWall. 2007. How emotion shapes behavior: Feedback, anticipation, and reflection, rather than direct causation. *Personality and Social Psychology Review*, 11(2): 167. ISSN 1088-8683.

C. Beedie, P. Terry, and A. Lane. 2005. Distinctions between emotion and mood. *Cognition and Emotion*, 19(6): 847–878. ISSN 0269-9931. https://doi.org/10.1080/02699930541000057.

M. Bekoff. 2008. *The emotional lives of animals: A leading scientist explores animal joy, sorrow, and empathy–and why they matter*. New World Library. ISBN 1577316290.

R. V. Belavkin. 2001. The role of emotion in problem solving. In *Proceedings of the AISB'01 Symposium on emotion, cognition and affective computing, Heslington, York, England*, pp. 49–57.

T. Bosse, C. M. Jonker, and J. Treur. 2008. Formalisation of damasio's theory of emotion, feeling and core consciousness. *Consciousness and Cognition*, 17(1): 94–113. ISSN 1053-8100. http://www.sciencedirect.com/science/article/pii/S1053810007000633. DOI: 10.1016/j.concog.2007.06.006.

S. Bozinovski. 1982. A self-learning system using secondary reinforcement. *Cybernetics and Systems Research*, pp. 397–402.

M. Bradley and P. Lang. 2007. Affective norms for english text (anet): Affective ratings of text and instruction manual. *Techical Report. D-1, University of Florida, Gainesville, FL.*

C. Breazeal. 1998. A motivational system for regulating human-robot interaction. In *AAAI*, pp. 54–61.

J. Broekens. 2007. *Emotion and Reinforcement: Affective Facial Expressions Facilitate Robot Learning*, pp. 113–132. http://dx.doi.org/10.1007/978-3-540-72348-6_6.

J. Broekens. 2018. A temporal difference reinforcement learning theory of emotion. *arXiv preprint arXiv:1807.08941*.

J. Broekens and M. Chetouani. 2019. Towards transparent robot learning through tdrl-based emotional expressions. *IEEE Transactions on Affective Computing*, in press.

J. Broekens and L. Dai. 2019. A tdrl model for the emotion of regret. In *8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 150–156. IEEE. ISBN 1728138884.

J. Broekens, C. M. Jonker, and J.-J. C. Meyer. 2010. Affective negotiation support systems. *J. Ambient Intell. Smart Environ.*, 2(2): 121–144. ISSN 1876-1364.

J. Broekens, T. Bosse, and S. C. Marsella. 2013. Challenges in computational modeling of affective processes. *Affective Computing, IEEE Transactions on*, 4(3): 242–245. ISSN 1949-3045.

J. Broekens, E. Jacobs, and C. M. Jonker. 2015. A reinforcement learning model of joy, distress, hope and fear. *Connection Science*, pp. 1–19. ISSN 0954-0091. http://dx.doi.org/10.1080/09540091.2015.1031081. DOI: 10.1080/09540091.2015.1031081.

J. K. Burgoon, N. Magnenat-Thalmann, M. Pantic, and A. Vinciarelli. 2017. *Social signal processing*. Cambridge University Press. ISBN 1108124585.

K. A. Buss and E. J. Kiel. 2004. Comparison of sadness, anger, and fear facial expressions when toddlers look at their mothers. *Child Development*, 75(6): 1761–1773. ISSN 1467-8624. http://dx.doi.org/10.1111/j.1467-8624.2004.00815.x. DOI: 10.1111/j.1467-8624.2004.00815.x.

L. Cañamero. 2001. Emotions and adaptation in autonomous agents: a design perspective. *Cybernetics and Systems*, 32(5): 507–529. ISSN 0196-9722.

L. Cañamero. 2003. *Designing emotions for activity selection in autonomous agents*, pp. 115–148.

R. A. Calvo, S. D'Mello, J. Gratch, and A. Kappas. 2014. *The Oxford Handbook of Affective Computing*. Oxford University Press. ISBN 0199942234.

L. Canamero. 2019. Embodied robot models for interdisciplinary emotion research. *IEEE Transactions on Affective Computing*, pp. 1–1. ISSN 2371-9850. DOI: 10.1109/TAFFC.2019.2908162.

C. Castelfranchi. Affective appraisal versus cognitive evaluation in social emotions and interactions. In *International Workshop on Affective Interactions*, pp. 76–106. Springer.

G. Castellano, A. Paiva, A. Kappas, R. Aylett, H. Hastie, W. Barendregt, F. Nabais, and S. Bull. 2013. *Towards Empathic Virtual and Robotic Tutors*, pp. 733–736. Springer Berlin Heidelberg, Berlin, Heidelberg. ISBN 978-3-642-39112-5.

S. Chong, J. F. Werker, J. A. Russell, and J. M. Carroll. 2003. Three facial expressions mothers direct to their infants. *Infant and Child Development*, 12(3): 211–232. ISSN 1522-7219.

A. Chowanda, M. Flintham, P. Blanchfield, and M. Valstar. 2016. Playing with social and emotional game companions. Intelligent Virtual Agents, pp. 85–95. Springer International Publishing. ISBN 978-3-319-47665-0.

C. Conati, S. Marsella, and A. Paiva. 2005. *Affective interactions: the computer in the affective loop*, pp. 7–7. ACM, San Diego, California, USA. ISBN 1581138946.

M. Core, D. Traum, H. C. Lane, W. Swartout, J. Gratch, M. van Lent, and S. Marsella. 2006. Teaching negotiation skills through practice and reflection with virtual humans. *SIMULATION*, 82(11): 685–701. http://sim.sagepub.com/cgi/content/abstract/82/11/685. DOI: 10.1177/0037549706075542.

I. Cos, L. Cañamero, G. M. Hayes, and A. Gillies. 2013. Hedonic value: enhancing adaptation for motivated agents. *Adaptive Behavior*, 21(6): 465–483. http://adb.sagepub.com/content/21/6/465.abstract. DOI: 10.1177/1059712313486817.

H. Cramer, J. Goddijn, B. Wielinga, and V. Evers. 2010. *Effects of (in)accurate empathy and situational valence on attitudes towards robots*, pp. 141–142. ISBN 2167-2148. DOI: 10.1109/HRI.2010.5453224.

A. R. Damasio. 1994. *Descartes' Error: emotion reason and the human brain*. Putnam, New York.

C. M. de Melo, P. Carnevale, and J. Gratch. 2011. *The effect of expression of anger and happiness in computer agents on negotiations with humans*, pp. 937–944. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 098265717X.

C. M. de Melo, P. J. Carnevale, S. J. Read, and J. Gratch. 2014. Reading people's minds from emotion expressions in interdependent decision making. *Journal of Personality and Social Psychology*, 106(1): 73. ISSN 1939-1315.

F. De Waal. 2019. *Mama's Last Hug: Animal Emotions and what They Tell Us about Ourselves*. WW Norton and Company. ISBN 0393635074.

J. Dias and A. Paiva. 2005. *Feeling and reasoning: A computational model for emotional characters*, pp. 127–140. Springer. ISBN 3540307370.

J. Dias, S. Mascarenhas, and A. Paiva. 2014. *Fatima modular: Towards an agent architecture with a generic appraisal framework*, pp. 44–56. Springer.

U. Dimberg, P. Andréasson, and M. Thunberg. 2011. Emotional empathy and facial reactions to facial expressions. *Journal of Psychophysiology*, 25(1): 26–31. https://econtent.hogrefe.com/doi/abs/10.1027/0269-8803/a000029. DOI: 10.1027/0269-8803/a000029.

S. D'Mello, R. W. Picard, and A. Graesser. 2007. Toward an affect-sensitive autotutor. 22: 53–61. ISSN 1541-1672. http://doi.ieeecomputersociety.org/10.1109/MIS.2007.79.

G. Dreisbach and K. Goschke. 2004. How positive affect modulates cognitive control: Reduced perseveration at the cost of increased distractibility. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(2): 343–353.

K. Edwards. 1990. The interplay of affect and cognition in attitude formation and change. *Journal of Personality and Social Psychology*, 59(2): 202–216. ISSN 1939-1315(Electronic),0022-3514(Print). DOI: 10.1037/0022-3514.59.2.202.

A. Egges, S. Kshirsagar, and N. Magnenat-Thalmann. 2004. Generic personality and emotion simulation for conversational agents. *Computer Animation and Virtual Worlds*, 15(1): 1–13. ISSN 1546-4261. https://onlinelibrary.wiley.com/doi/abs/10.1002/cav.3. DOI: 10.1002/cav.3.

P. Ekman and W. Friesen. 1971. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2): 124. ISSN 1939-1315.

M. S. El-Nasr, J. Yen, and T. R. Ioerger. 2000. Flame—fuzzy logic adaptive model of emotions. *Autonomous Agents and Multi-agent systems*, 3(3): 219–257. ISSN 1387-2532.

C. D. Elliott. 1992. *The Affective Reasoner: A process model of emotions in a multi-agent system*. Thesis.

A. H. Fischer and A. Manstead. 2008. *Social Functions of Emotion*, pp. 456–468. Guilford Press.

J. P. Forgas. 2000. *Feeling is believing? The role of processing strategies in mediating affective influences in beliefs*, pp. 108–143. Cambridge University Press.

S. Franklin and A. Graesser. 1999. A software agent model of consciousness. *Consciousness and Cognition*, 8(3): 285–301. ISSN 1053-8100. DOI: http://dx.doi.org/10.1006/ccog.1999.0391.

S. Franklin, T. Madl, S. D'Mello, and J. Snaider. 2014. Lida: A systems-level architecture for cognition, emotion, and learning. *IEEE Transactions on Autonomous Mental Development*, 6(1): 19–41. ISSN 1943-0604. DOI: 10.1109/TAMD.2013.2277589.

N. H. Frijda. 1988. The laws of emotion. *American Psychologist*, 43(5): 349. ISSN 1935-990X.

N. H. Frijda. 2004. *Emotions and action*, p. 158–173. Cambridge University Press.

N. H. Frijda and J. Swagerman. 1987. Can computers feel? theory and design of an emotional system. *Cognition and Emotion*, 1(3): 235–257. ISSN 0269-9931. https://doi.org/10.1080/02699938708408050. DOI: 10.1080/02699938708408050.

N. H. Frijda, A. S. R. Manstead, and S. Bem. 2000. *Emotions and Beliefs: How Feelings Influence Thoughts*. Cambridge University Press.

P. Gebhard. 2005. Alma: a layered model of affect. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pp. 29–36.

L. R. Goldberg. 1990. An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology*, 59(6): 1216–1229. ISSN 1939-1315(Electronic),0022-3514(Print).

J. Gratch and S. Marsella. 2001. *Tears and fears: modeling emotions and emotional behaviors in synthetic agents*, pp. 278–285. ACM, Montreal, Quebec, Canada. DOI: http://doi.acm.org/10.1145/375735.376309.

J. Gratch and S. Marsella. 2004. A domain-independent framework for modeling emotion. *Cognitive Systems Research*, 5(4): 269–306. ISSN 1389-0417. http://www.sciencedirect.com/science/article/B6W6C-4C56KYY-1/2/e21f759dcf674531f63aa07c171a0f31.

J. Gratch and S. Marsella. 2014. *Appraisal Models*, pp. 54–67. ISBN 0199942234.

J. Gratch, J. Rickel, E. Andr #233, J. Cassell, E. Petajan, and N. Badler. 2002. Creating interactive virtual humans: Some assembly required. *IEEE Intelligent Systems*, 17(4): 54–63. ISSN 1541-1672. DOI: 10.1109/mis.2002.1024753.

J. Gratch, S. Marsella, N. Wang, and B. Stankovic. 2009. *Assessing the validity of appraisal-based models of emotion*.

J. Gratch, D. DeVault, G. M. Lucas, and S. Marsella. 2015. Negotiation as a challenge problem for virtual humans. In W.-P. Brinkman, J. Broekens, and D. Heylen, eds., *Intelligent Virtual Agents*, pp. 201–215. Springer International Publishing, Cham.

R. Greifeneder, H. Bless, and M. T. Pham. 2010. When do people rely on affective and cognitive feelings in judgment? a review. *Personality and Social Psychology Review*, 15(2): 107–141. ISSN 1088-8683. https://doi.org/10.1177/1088868310367640. DOI: 10.1177/1088868310367640.

D. Heylen, A. Nijholt, R. o. d. Akker, and M. Vissers. 2003. *Socially intelligent tutor agents*, pp. 341–347. Lecture Notes in Artificial Intelligence.

K. Hoemann, F. Xu, and L. F. Barrett. 2019. Emotion words, emotion concepts, and emotional development in children: A constructionist hypothesis. *Developmental psychology*, 55(9): 1830.

E. Hogewoning, J. Broekens, J. Eggermont, and E. Bovenkamp. 2007. *Strategies for Affect-Controlled Action-Selection in Soar-RL*, pp. 501–510. Springer, Berlin.

E. Hudlicka. 2003. To feel or not to feel: The role of affect in human-computer interaction. *International Journal of Human-Computer Studies*, 59(1-2): 1–32. ISSN 1071-5819. http://www.sciencedirect.com/science/article/B6WGR-48SNKXG-2/2/c172889a95be64cb4419763a5cefa852.

E. Hudlicka. 2005. *Modeling interactions between metacognition and emotion in a cognitive architecture*, pp. 55–61.

R. E. Jack, O. G. Garrod, and P. G. Schyns. 2014. Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current Biology*, 24: 1–6.

H. Jiang, J. M. Vidal, and M. N. Huhns. 2007. Ebdi: an architecture for emotional agents. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, pp. 1–3.

H. Jones and N. Sabouret. 2013. Tardis-a simulation platform with an affective virtual recruiter for job interviews. In *IDGEI (Intelligent Digital Games for Empowerment and Inclusion)*.

M. F. Jung. 2017. Affective grounding in human-robot interaction. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, pp. 263–273.

L. P. Kaelbling, M. L. Littman, and A. W. Moore. 1996. Reinforcement learning: A survey. *Journal of Articial Intelligence Research*, 4: 237–285.

F. Kaptein, J. Broekens, K. V. Hindriks, and M. Neerincx. 2016. *CAAF: A Cognitive Affective Agent Programming Framework*, pp. 317–330. Springer International Publishing.

F. Kaptein, J. Broekens, K. Hindriks, and M. Neerincx. 2017. The role of emotion in self-explanations by cognitive agents. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pp. 88–93. IEEE.

K. Kiryazov, R. Lowe, C. Becker-Asano, and M. Randazzo. 2013. The role of arousal in two-resource problem tasks for humanoid service robots. In *2013 IEEE RO-MAN*, pp. 62–69. ISBN 1944-9437.

G. v. Kleef, C. De Dreu, and A. Manstead. 2004. The interpersonal effects of emotions in negotiations: A motivated information processing approach. *Journal of Personality and Social Psychology*, 87(4): 510–528.

M. D. Klinnert. 1984. The regulation of infant behavior by maternal facial expression. *Infant Behavior and Development*, 7(4): 447–465. ISSN 0163-6383. http://www.sciencedirect.com/science/article/pii/S0163638384800053. DOI: http://dx.doi.org/10.1016/S0163-6383(84)80005-3.

R. S. Lazarus. 1991. Cognition and motivation in emotion. *American psychologist*, 46(4): 352.

J. LeDoux. 1996. *The Emotional Brain*. Simon and Shuster, New York.

K. Lee and M. C. Ashton. 2004. Psychometric properties of the hexaco personality inventory. *Multivariate Behavioral Research*, 39(2): 329–358. ISSN 0027-3171.

I. Leite, C. Martinho, A. Pereira, and A. Paiva. 2008. icat: an affective game buddy based on anticipatory mechanisms. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 3*, pp. 1229–1232. International Foundation for Autonomous Agents and Multiagent Systems.

D. Leyzberg, E. Avrunin, J. Liu, and B. Scassellati. 2011. *Robots that express emotion elicit better human teaching*, pp. 347–354. ISBN 2167-2121. DOI: 10.1145/1957656.1957789.

B. Liu and L. Zhang. 2012. *A Survey of Opinion Mining and Sentiment Analysis*, pp. 415–463. Springer US, Boston, MA. ISBN 978-1-4614-3223-4.

R. Lowe, E. Barakova, E. Billing, and J. Broekens. 2016. Grounding emotions in robots–an introduction to the special issue. *Adaptive Behavior*, 24(5): 263–266. ISSN 1059-7123.

G. R. Maio, G. Haddock, and B. Verplanken. 2018. *The psychology of attitudes and attitude change*. Sage Publications Limited. ISBN 1526454122.

K. Man and A. Damasio. 2019. Homeostasis and soft robotics in the design of feeling machines. *Nature Machine Intelligence*, 1(10): 446–452. ISSN 2522-5839. https://doi.org/10.1038/s42256-019-0103-7. DOI: 10.1038/s42256-019-0103-7.

R. Marinier and J. E. Laird. 2008. *Emotion-driven reinforcement learning*, pp. 115–120.

R. P. Marinier III and J. E. Laird. 2004. *Toward a Comprehensive Computational Model of Emotions and Feelings*, pp. 172–177.

S. Marsella and J. Gratch. 2009. Ema: A process model of appraisal dynamics. *Cognitive Systems Research*, 10(1): 70–90. ISSN 1389-0417. http://www.sciencedirect.com/science/article/B6W6C-4SX9G35-1/2/484cdc830f9bfa8c5d12b87ddd5bace7.

M. S. Mast and J. A. Hall. 2017. *The Vertical Dimension of Social Signaling*. Cambridge University Press.

G. E. Matt, C. Vázquez, and W. K. Campbell. 1992. Mood-congruent recall of affectively toned stimuli: A meta-analytic review. *Clinical Psychology Review*, 12(2): 227–255. ISSN 0272-7358. http://www.sciencedirect.com/science/article/pii/027273589290116P. DOI: https://doi.org/10.1016/0272-7358(92)90116-P.

J. D. Mayer, R. D. Roberts, and S. G. Barsade. 2008. Human abilities: Emotional intelligence. *Annu. Rev. Psychol.*, 59: 507–536. ISSN 0066-4308.

R. R. McCrae and P. T. Costa. 1987. Validation of the five-factor model of personality across instruments and observers. *Journal of personality and social psychology*, 52(1): 81. ISSN 1939-1315.

A. Mehrabian. 1980. *Basic Dimensions for a General Psychological Theory*. OG and H Publishers.

J.-J. C. Meyer. 2006. Reasoning about emotional agents. *International Journal of Intelligent Systems*, 21(6): 601–619. ISSN 1098-111X. http://dx.doi.org/10.1002/int.20150.

T. Moerland, J. Broekens, and C. Jonker. 2016. *Fear and Hope Emerge from Anticipation in Model-Based Reinforcement Learning*, pp. 848–854. AAAI Press.

T. M. Moerland, J. Broekens, and C. M. Jonker. 2018. Emotion in reinforcement learning agents and robots: a survey. *Machine Learning*, 107(2): 443–480. ISSN 1573-0565. https://doi.org/10.1007/s10994-017-5666-0. DOI: 10.1007/s10994-017-5666-0.

A. Moors, P. C. Ellsworth, K. R. Scherer, and N. H. Frijda. 2013. Appraisal theories of emotion: State of the art and future development. *Emotion Review*, 5(2): 119–124.

A. Moors, Y. Boddez, and J. De Houwer. 2017. The power of goal-directed processes in the causation of emotional and other actions. *Emotion Review*, 9(4): 310–318. ISSN 1754-0739. https://doi.org/10.1177/1754073916669595. DOI: 10.1177/1754073916669595.

R. Neumann, B. Seibt, and F. Strack. 2001. The influence of mood on the intensity of emotional responses: Disentangling feeling and knowing. *Cognition and Emotion*, 15(6): 725–747. ISSN 0269-9931.

K. Oatley. 2010. Two movements in emotions: Communication and reflection. *Emotion Review*, 2(1): 29–35. http://emr.sagepub.com/content/2/1/29.abstract. DOI: 10.1177/1754073909345542.

M. Ochs, N. Sabouret, and V. Corruble. 2009. Simulation of the dynamics of nonplayer characters' emotions and social relations in games. *Computational Intelligence and AI in Games, IEEE Transactions on*, 1(4): 281–297. ISSN 1943-068X. DOI: 10.1109/tciaig.2009.2036247.

A. Ortony, G. L. Clore, and A. Collins. 1988. *The Cognitive Structure of Emotions*. Cambridge University Press.

A. Paiva. 2000. *Affective Interactions: Toward a New Generation of Computer Interfaces?*, pp. 1–8. http://dx.doi.org/10.1007/10720296_1.

A. Paiva, I. Leite, H. Boukricha, and I. Wachsmuth. 2017. Empathy in virtual agents and robots: A survey. *ACM Trans. Interact. Intell. Syst.*, 7(3): Article 11. ISSN 2160-6455. https://doi.org/10.1145/2912150. DOI: 10.1145/2912150.

J. Panksepp. 1982. Toward a general psychobiological theory of emotions. *Behavioral and Brain Sciences*, 5(03): 407–422. ISSN 1469-1825.

L. Peña, J.-M. Peña, and S. Ossowski. 2011. Representing emotion and mood states for virtual agents. In F. Klügl and S. Ossowski, eds., *Multiagent System Technologies*, pp. 181–188. Springer Berlin Heidelberg, Berlin, Heidelberg. ISBN 978-3-642-24603-6.

R. Pfeifer. 1982. Cognition and emotion: An information processing approach. *CIP working paper 436*.

R. Pfeifer. 1988. *Artificial Intelligence Models of Emotion*, pp. 287–320. Springer Netherlands, Dordrecht. ISBN 978-94-009-2792-6.

R. W. Picard. 1997. *Affective Computing*. MIT Press.

A. Popescu, J. Broekens, and M. v. Someren. 2014. Gamygdala: An emotion engine for games. *IEEE Transactions on Affective Computing*, 5(1): 32–44. ISSN 1949-3045. http://doi.ieeecomputersociety.org/10.1109/T-AFFC.2013.24. DOI: 10.1109/T-AFFC.2013.24.

S. N. Reilly. 2006. Modeling what happens between emotional antecedents and emotional consequents. *ACE*, 2006(19).

W. S. Reilly. 1996. Believable social and emotional agents. Report, Carnegie-Mellon Univ Pittsburgh, Dept of Computer Science.

W. S. Reilly and J. Bates. 1992. Building emotional agents. Report, School of Computer Science, Carnegie Mellon University.

D. Reisberg and P. Hertel. 2003. *Memory and emotion*. Oxford University Press. ISBN 019534796X.

R. Reisenzein. 2009a. Emotional experience in the computational belief-desire theory of emotion. *Emotion Review*, 1(3): 214–222. http://emr.sagepub.com/cgi/content/abstract/1/3/214. DOI: 10.1177/1754073909103589.

R. Reisenzein. 2009b. Emotions as metarepresentational states of mind: Naturalizing the belief-desire theory of emotion. *Cognitive Systems Research*, 10(1): 6–20. ISSN 1389-0417. http://www.sciencedirect.com/science/article/B6W6C-4SSNDC1-1/2/6c65804e7c6115a7e230a17e0285bfd1.

J. M. Robillard and J. Hoey. 2018. Emotion and motivation in cognitive assistive technologies for dementia. *Computer*, 51(3): 24–34. ISSN 1558-0814. DOI: 10.1109/MC.2018.1731059.

E. T. Rolls. 2014. Emotion and decision-making explained: A précis. *cortex*, 59: 185–193. ISSN 0010-9452.

I. J. Roseman and C. A. Smith. 2001. Appraisal theory. *Appraisal processes in emotion: Theory, methods, research*, pp. 3–19.

F. d. Rosis, C. Pelachaud, I. Poggi, V. Carofiglio, and B. D. Carolis. 2003. From greta's mind to her face: modelling the dynamics of affective states in a conversational embodied agent. *International Journal of Human-Computer Studies*, 59(1-2): 81–118. ISSN 1071-5819. http://www.sciencedirect.com/science/article/B6WGR-487DHWN-1/2/125e0d74aaa75243f67ca1711d527201.

J. A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6): 1161–1178.

C. Saint-georges, M. Chetouani, R. Cassel, F. Apicella, A. Mahdhaoui, F. Muratori, M. Laznik, and D. Cohen. 2013. Motherese in interaction: at the cross-road of emotion and cognition? (a systematic review). *PLoS ONE*, 8(10): e78103.

M. A. Salichs and M. Malfaz. 2012. A new approach to modeling emotions and their use on a decision-making system for artificial agents. *IEEE Transactions on Affective Computing*, 3(1): 56–68. ISSN 2371-9850. DOI: 10.1109/T-AFFC.2011.32.

K. Scherer. 2001. *Appraisal considered as a process of multilevel sequential checking*, pp. 92–120.

K. R. Scherer. 2005. What are emotions? and how can they be measured? *Social science information*, 44(4): 695–729.

D. Schuller and B. W. Schuller. 2018. The age of artificial emotional intelligence. *Computer*, 51(9): 38–46. ISSN 1558-0814. DOI: 10.1109/MC.2018.3620963.

N. Schweighofer and K. Doya. 2003. Meta-learning in reinforcement learning. *Neural networks*, 16(1): 5–9. ISSN 0893-6080. http://www.sciencedirect.com/science/article/pii/S0893608002002289. DOI: 10.1016/s0893-6080(02)00228-9.

P. Sequeira, F. S. Melo, and A. Paiva. 2014. Learning by appraising: an emotion-based approach to intrinsic reward design. *Adaptive Behavior*, 22(5): 330–349. https://journals.sagepub.com/doi/abs/10.1177/1059712314543837. DOI: 10.1177/1059712314543837.

A. Sloman and M. Croucher. 1981. Why robots will have emotions. Report, Sussex University.

C. A. Smith and R. S. Lazarus. 1990. *Emotion and adaptation*, pp. 609–637. The Guilford Press, New York, NY, US.

B. R. Steunebrink, M. Dastani, and J.-J. C. Meyer. 2007. *A Logic of Emotions for Intelligent Agents*, pp. 142–147. AAAI Press.

B. R. Steunebrink, M. Dastani, and J.-J. C. Meyer. 2008. *A Formal Model of Emotions: Integrating Qualitative and Quatitative Aspects*, pp. 256–260. IOS Press.

R. S. Sutton and A. G. Barto. 2018. *Reinforcement learning: An introduction*. MIT press. ISBN 0262352702.

S. Thrun, M. Bennewitz, W. Burgard, A. Cremers, F. Dellaert, D. Fox, D. Hähnel, C. Rosenberg, N. Roy, J. Schulte, and D. Schulz. 1999. *MINERVA: A Tour-Guide Robot that Learns*, pp. 696–696. http://dx.doi.org/10.1007/3-540-48238-5_2.

S. Turkle, C. Breazeal, O. Dasté, and B. Scassellati. 2006. Encounters with kismet and cog: Children respond to relational artifacts. *Digital media: Transformations in human communication*, 120.

C. M. Van Reekum and K. R. Scherer. 1997. Levels of processing in emotion-antecedent appraisal. In *Advances in Psychology*, volume 124, pp. 259–300. Elsevier.

J. Velasquez. 1998. *Modeling emotion-based decision making.* AAAI Press.

J. W. Verdijk, D. Oldenhof, D. Krijnen, and J. Broekens. 2015. Growing emotions: Using affect to help children understand a plant's needs. In *Affective Computing and Intelligent Interaction (ACII), International Conference on*, pp. 160–165. DOI: 10.1109/ACII.2015.7344566.

S. K. Vosburg. 1998. The effects of positive and negative mood on divergent-thinking performance. *Creativity Research Journal*, 11(2): 165–172. ISSN 1040-0419. https://doi.org/10.1207/s15326934crj1102_6.

K. Weber, A. Johnson, and M. Corrigan. 2004. Communcating emotional support and its relationship to feelings of being understood, trust, and self-disclosure. *Communication Research Reports*, 21(3): 316–323. ISSN 0882-4096. https://doi.org/10.1080/08824090409359994. DOI: 10.1080/08824090409359994.

S. C. Widen and J. A. Russell. 2008. Children acquire emotion categories gradually. *Cognitive Development*, 23(2): 291–312. ISSN 0885-2014. http://www.sciencedirect.com/science/article/pii/S0885201408000038. DOI: https://doi.org/10.1016/j.cogdev.2008.01.002.

G. N. Yannakakis and A. Paiva. 2014. *Emotion in games*, pp. 459–471. Oxford University Press.

P. Zachar and R. D. Ellis. 2012. *Categorical versus dimensional models of affect: a seminar on the theories of Panksepp and Russell*, volume 7. John Benjamins Publishing. ISBN 9027274754.