CHAPTER 2

BACKGROUND AND RELATED WORK

2.1 Introduction

In this chapter, we review the background of prominent collaboration theories including the the SharedPlans theory [103] as one of the two theoretical foundations of our work. We also provide the similarities and differences between these theories as well as both theoretical and practical related work and applications of them. We continue by discussing the concept of affective computing and the social and communicative aspects of emotions from a psychological point of view. Understanding the social aspects of emotions is important in our work, since it is focused on collaboration which is a social phenomenon in human environments. We also present the concept of artificial emotions and provide some examples of the existing computational models of emotions. Then, we provide the background of the cognitive appraisal theory of emotions as the second theoretical foundation in our work as well as other computational models of emotions and the related concepts such as some examples of cognitive architectures and the influence of affect in decision-making procedures. This chapter continues with the description of motives and the related theories in psychology and artificial intelligence. The role of motives as goal-driven affective components is crucial in our work, since the collaboration structure is built based on the concept of a shared goal between collaborators. Finally, a brief description and the related work in psychology and artificial intelligence is provided about theory of mind as another concept used in a small limited-scale in our work.

2.2 Computational Collaboration Theories

The construction of computer systems and robots that are intelligent, collaborative problem-solving partners is important in Artificial Intelligence (AI) and its applications. It has always been important for us to make computer systems better at helping us do whatever they are designed for. To build collaborative systems, we need to identify the capabilities that must be added to individual agents so that they can work with us or other agents. As Grosz says, collaboration must be designed into systems from the start; it cannot be patched on [97].

Collaboration is a special type of coordinated activity in which the participants work jointly, together performing a task or carrying out the activities needed to satisfy a shared goal [101]. Collaboration involves several key properties both in structural and functional levels. For instance, most collaborative situations involve participants who have different beliefs and capabilities; most of the time collaborators only have partial knowledge of the process of accomplishing the collaborative activities; collaborative plans are more than the sum of individual plans; collaborators are required to maintain mutual beliefs about their shared goal throughout the collaboration; they need to be able to communicate with others effectively; they need to commit to the group activities and to their role in it; collaborators need to commit to the success of others; they need to reconcile between commitments to the existing collaboration and their other activities; and they need to interpret others' actions and utterances in the collaboration context [98]. These collaboration properties are captured by the existing computational collaboration theories.

As we mentioned, to be collaborative, partners, e.g., a robot and a human, need to meet the specifications stipulated by collaboration theories. These theories argue for an essential distinction between a collaboration and a simple interaction or even a coordination in terms of commitments [96, 152]. This section briefly provides descriptions of major computational collaboration theories, their similarities and differences, and their application in AI and robotics. It primarily focuses on Joint

Intention, SharedPlans and hybrid theories of collaboration. In this section, we do not present the theories in formal language, but simply describe their features in general terms.

The prominent collaboration theories are mostly based on plans and joint intentions [54] [103] [150], and they were derived from the Belief-Design-Intention (BDI) paradigm developed by Bratman [26] which is fundamentally reliant on folk psychology [199]. The two theories, Joint Intentions [54] and SharedPlans [103], have been used extensively to examine and describe teamwork and collaboration.

The SharedPlans theory is based on the theories of Bratman and Pollack [29, 190, 191], who outline a mental-state view of plans in which having a plan is not just knowing how to do an action, but also having the intention to do the actions entailed. Bratman's views of intention goes back to the philosophical views of Anscombe [6] and *Castañeda* [44] about intention. Also, as Grosz and Sidner mention in [103] the natural segmentation of discourse reflects intentional behaviors in each segment. These intentions are designated as Discourse Segment Purposes (DSPs) which are the basic reasons for engaging in different segments of discourse. DSPs are a natural extension of Gricean intentions at the utterance level [177].

Cohen and Levesque also mention that in Joint Intentions theory their view of intention is primarily future-directed [55] which makes their view similar to Bratman's theory of intention [27], contra Searle [231].

Commitment – One of the most important concepts of teamwork and collaboration is the concept of commitment. Collaboration theories are required to meet the notion of commitment, otherwise the participants are just doing some coordinated works. Since the prominent computational collaboration theories, reviewed in this paper, are based on Bratman's view of intention, we briefly provide his view of commitment here before describing these theories. Bratman defines certain prerequisites for an activity to be considered shared and cooperative [28]. He stresses the importance of:

a) **Mutual commitment to joint activity** – which can be achieved by agreement on the joint activity, and prevention of abandoning the activity without involving teammates;

b) **Mutual support** – which can be achieved by team members if they actively try to help teammate activity;

c) **Mutual responsiveness** – which means team members should take over tasks from teammates if necessary.

In the following sections, we are also going to see how each collaboration theory addresses the notion of commitment.

2.2.1 Shared-Plans Theory

The SharedPlans model of collaborative action, presented by Grosz and Sidner [100, 101, 103], aims to provide the theoretical foundations needed for building collaborative robots/agents [97]. SharedPlans is a general theory of collaborative planning that requires no notion of joint intentions (see Section 2.2.2), accommodates multi-level action decomposition hierarchies and allows the process of expanding and elaborating partial plans into full plans (see Section 2.2.1). SharedPlans theory explains how a group of agents can incrementally form and execute a shared plan that then guides and coordinates their activity towards the accomplishment of a shared goal. SharedPlans is rooted in the observation that collaborative plans are not simply a collection of individual plans, but rather a tight interleaving of mutual beliefs and intentions of different team members. In [101] Grosz and Kraus use first-order logic to present the formalization of SharedPlans.

Grosz and Sidner in [103] present a model of plans to account for how agents with partial knowledge collaborate in the construction of a domain plan. They are interested in the type of plans that underlie discourse in which the agents are collaborating in order to achieve a shared goal. They propose that agents are building

a shared plan in which participants have a collection of beliefs and intentions about the actions in the plan. Agents have a library of how to do their actions, i.e. recipes (see Section 2.2.1). These recipes might be partially specified as to how an action is executed, or contributes to a goal (see Section 2.2.1). Then, each agent communicates their beliefs and intentions by making utterances about what actions they can contribute to the shared plan. This communication leads to the construction of a shared plan, and ultimately termination of the collaboration with each agent mutually believing that there exists one agent who is going to execute an action in the plan, and the fact that that agent has intention to perform the action, and that each action in the plan contributes to the goal [103] [153].

Later, we are going to see that to successfully complete a plan the collaborators must mutually believe that they have a common goal and have agreed on a sequence of actions for achieving that goal. They should believe that they are both capable of performing their own actions and intend to perform those actions while they are committed to the success of their plans.

Recipes

The SharedPlans theory differentiates between knowing how to accomplish a goal (a recipe) and having a plan, which includes intentions. The SharedPlans definition of mutual beliefs states that when agents have a shared plan for doing some action, they must hold mutual beliefs about the way in which they should perform that action [101, 103]. Following Pollack [191], the term recipe refers to what collaborators know when they know a way of doing an action. Recipes are specified at a particular level of detail. Although the agents need to have mutual beliefs about actions specified in the recipe, they do not need to have mutual beliefs about all levels of performing actions. Therefore, having mutual beliefs of the recipe means that the collaborators hold the same beliefs about the way in which an action should be accomplished. Consequently, the collaborators need to agree on how to execute an action. Recipes are aggregations of action-types and relations among them. Action-types, rather

than actions, are the main elements in recipes. Grosz and Sidner in their earlier work [103] have considered only simple recipes in which each recipe consisted of only a single action-type relation [153]. Recipes can be partial, meaning they can expand and be modified over time.

Grosz and Sidner propose that collaboration must have the following three elements, which also indicates the importance of the shared plan:

1. the participants must have commitment to the shared activity;

2. there must be a process for reaching an agreement on a recipe for the group action;

3. there must be commitment to the constituent actions.

Shared plan is an essential concept in the collaboration context. The definition of the shared plan is derived from the definition of plans Pollack introduced in [190, 191] since it rests on a detailed treatment of the relations among actions and it distinguishes the intentions and beliefs of an agent about those actions. However, since Pollack's plan model is just a simple plan of a single agent, Grosz and Sidner extended that to plans of two or more collaborative agents. The concept of the shared plan provides a framework in which to further evaluate and explore the roles that particular beliefs and intentions play in collaborative activity [153]. However, this formulation of shared plans (a) could only deal with activities that directly decomposed into single-agent actions, (b) did not address the requirement for the commitment of the agents to their joint activities, and (c) did not adequately deal with agents having partial recipes [101]. Grosz and Kraus in [101], reformulate Pollack's definition of the individual plans [191], and also revise and expand the SharedPlans to address these shortcomings.

Figure 2.1 shows what we need to add to individual plans in order to have plans for group actions. The top of the figure lists the main components for individual plans. First, an individual agent needs to know the recipe for an action, whereas

agents in a group need to have a mutual belief of a recipe for an action (bottom of the figure). In the case of a group plan, having a mutual belief of a recipe, leads the agents to agree on how they are going to execute the action. Then, similar to individual agents that need to have the ability to perform the constituent actions in an individual plan and must have intentions to perform them, the participants in a group activity need to have individual or group plans for each of the constituent actions in the mutually agreed recipe [97, 103].

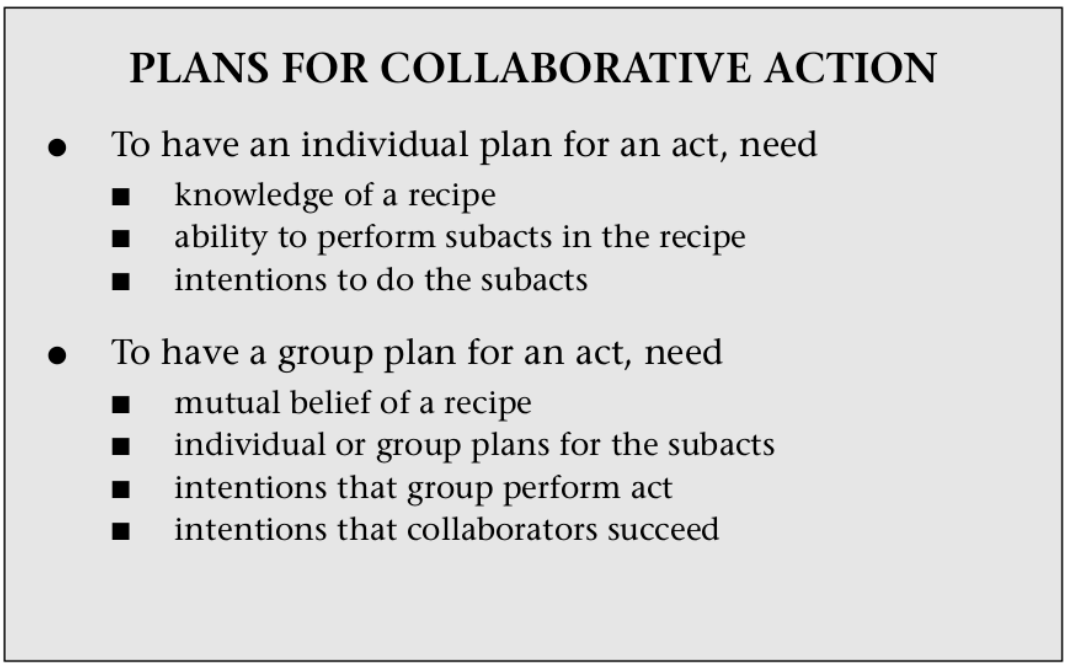

Figure 2.1: Plans for collaborative action [97].

As shown in Figure 2.1 (bottom), plans for group actions include two essential constituents that do not have correlates in the individual plan. First, the agents need to have a commitment to the group activity; All the agents need to intend that (see Section 2.2.1) the group will do the action. For instance, a robot and an astronaut need to have intentions that they install solar panels together. Among other things, these intentions will keep them both working on the panels until the panels are installed. Second, the participants need to have some commitment to the other agents to succeed in their own their actions. For instance, the robot must

have an intention that the astronaut be able to measure the quality of installation successfully. This intention will prevent the robot from interrupting the astronaut's measurement action or prevent the robot from using the astronaut's measurement tool [97, 103].

Full Vs. Partial Shared Plan

The SharedPlans formalization distinguishes complete plans and partial plans. A shared plan can be either a *Full Shared Plan (FSP)* or a *Partial Shared Plan (PSP)*. An *FSP* is a complete plan in which agents have fully determined how they will perform an action. A *PSP* definition provides a specification of the minimal mental state requirements for collaboration to exist and gives criteria governing the process of completing the plan.

An *FSP* to do α represents a situation where every aspect of a joint activity α is fully determined. This includes mutual belief and agreement in the complete recipe to do α. A recipe is a specication of a set of actions A_i, which constitutes the performance of α when executed under specified constraints. *FSP(*$\boldsymbol{P}$, $\Theta, \alpha, T_p, T_\alpha, \boldsymbol{R}_\alpha$*)* denotes a group Θ's plan $\boldsymbol{P}$ at time T_p to do action α at time T_α using recipe $\boldsymbol{R}_\alpha$. In short, *FSP* holds if and only if the following conditions are satisfied:

1. All members of group Θ mutually believe that they intend to do α.
2. All members of group Θ mutually believe that $\boldsymbol{R}_\alpha$ is the recipe for α.
3. For each step A_i in recipe $\boldsymbol{R}_\alpha$:
   - A subgroup Θ_j has an *FSP* for A_i, using recipe $\boldsymbol{R}_{Ai}$.
   - Other members of group Θ believe that there exists a recipe such that subgroup Θ_j can bring about A_i and have an FSP for A_i.
   - Other members of group Θ intend that subgroup Θ_j can bring about A_i using some recipe.

Most times a team and its members do not possess an *FSP* to achieve their shared goal. In this case, the concept of *FSP* puts limits on the SharedPlans theory. However, SharedPlans uses the concept of *PSP* as a snapshot of the team's mental states in different situations, which further leads to communication and planning to fulfill the conditions of an *FSP*. The idea behind *PSP* is enabling the agents to modify the shared plan over the course of planning without impairing the achievement of the shared goals. Notice that for the same reason recipes also can be partial [101, 103].

Communicating Intentions

In SharedPlans theory Grosz and Sidner are interested in the type of plans that underlie a discourse in which the agents collaborate to achieve a shared goal. Here we present their view of discourse structure, since it is directly related to the intentions behind collaborators' actions. In [103], Grosz and Sidner argue that the SharedPlans theory recognises three interrelated levels of discourse structure, and the components of the discourse structure are a trichotomy of linguistic structure, intentions structure and the attention state. In their work, the linguistic structure of a discourse is a sequence of utterances aggregating into discourse segments just as the words in a single sentence form constituent phrases. They also discuss the idea of the discourse purpose as the intention that underlies engagement in the particular discourse. They believe this intention is the reason behind performing a discourse rather than some other actions, and also the reason behind conveying a particular content of the discourse rather than some other contents. They describe mechanisms for plan analysis looking at Discourse Segment Purposes (DSPs). In fact, the DSPs specify how the discourse segments contribute to achieving the overall discourse purpose. Finally, the third component in their theory, the attentional state, provides an abstraction of the agent's focus of attention as the discourse unfolds. The focusing structure contains DSPs and the stacking of focus spaces reflects the relative salience of the entities in each space during the discourse. In short, the

focusing structure is the central repository for the contextual content required for processing utterances during the discourse [103]. Using discourse plans can help to encode the knowledge about conversation.

Intention-to and Intention-that

In Grosz and Sidner's SharedPlans theory [103], two intentional attitudes are employed: *intending to* (do an action) and *intending that* (a proposition will hold). The notion of *intention to*, as an individual-oriented intention, models the intention of an agent to do any single-agent action while the agent not only believes that it is able to execute that action, but it also commits to doing so. In short, it is an intention to perform an action, similar to Bratman's view of intention. In contrast with *intention to*, an *intention that*, as an intention directed toward group activity, does not directly imply an action. In fact, an individual agent's *intention that* is directed towards its collaborators' action or towards a group's joint action. *Intention that* guides an agent to take actions (including communication), that enable or facilitate other collaborators to perform assigned tasks. This leads an agent to behave collaboratively. Therefore, agents will adopt intentions to communicate about the plan [101]. As another difference, *Intention to* commits an agent to means-end reasoning and acting [26] while *Intention that* does not necessarily entail this commitment. The key point about *Intention to* and *intention that* is that both commit an agent not to adopt conflicting intentions, and constrain replanning in case of failure. Further, an agent can *intention that* another agent achieve the specified proposition.

2.2.2 Joint Intentions Theory

Following Bratman's guidelines, Cohen and Levesque propose a formal approach to building artificial collaborative agents. The Joint Intentions theory of Cohen and Levesque [54, 55, 56, 57, 146] represents one of the first attempts to establish a

formal theory of collaboration, and due to its clarity and expression, is a widely used teamwork theory.

The basic idea of Joint Intentions theory is based on individual and joint intentions (as well as commitments) to act as a team member. Their notion of joint intention is viewed not only as a persistent commitment of the team to a shared goal, but also implies a commitment on part of all its members to a mutual belief about the state of the goal. In other words, Joint Intentions theory describes how a team of agents can jointly act together by sharing mental states about their actions while an intention is viewed as a commitment to perform an action. A joint intention is a shared commitment to perform an action while in a group mental state [55].

In [54] Cohen and Levesque establish that joint intention cannot be defined simply as individual intention with the team regarded as an individual. The reason is that after the initial formation of an intention, team members may diverge in their beliefs and their attitudes towards the intention. Instead, Cohen and Levesque generalize their own definition of intention. First, they present a definition of individual persistent goal (see Section 2.2.2) and individual intention (see Section 2.2.2). Then, they define analogues of these concepts by presenting mutual belief in place of individual belief. The definition of joint persistent goal (see Section 2.2.2) requires team members to commit to informing other members, if it comes to believe that the shared goal is in its terminal status. As a result, in Cohen and Levesque's theory, a team with a joint intention is a group that shares a common objective and a certain shared mental state [121].

In this theory, once an agent entered into a joint commitment with other agents, the agent should communicate its private beliefs with other team members if the agent believes that the joint goal is in its terminal status, i.e., either the joint goal is achieved, or it is unachievable, or irrelevant [264]. Thus, as we mentioned above, team members are committed to inform other team members when they reach the conclusion that a goal is achievable, impossible, or irrelevant. For instance, if a robot and an astronaut are collaborating to install a solar panel, and the robot reaches

the conclusion that the welding tool has a deficiency, it is essential for the robot to have an intention to communicate with the astronaut and make this knowledge common. Therefore, according to this theory, in a collaboration, agents can count on the commitment of other members, first to the goal and then to the mutual belief of the status of the goal.

Individual Commitment

As we mentioned earlier, intentions and commitments are the basic ideas of Joint Intentions theory. Here, we provide the definition of "individual commitment" (also called *persistent goal*) by Cohen et. al. in [53]. According to their definition an agent has a persistent goal relative to q to achieve p only when:

1. agent believes that p is currently false;
2. agent wants p to be true;
3. it is true (and agent knows it) that (2) will continue to hold until the agent comes to believe either that p is true, or that it will never be true, or that q is false.

Note that the condition q is an "escape" clause, which can be omitted for brevity, or it can be used as a reason for the agent to drop a commitment, even though it could be quite vague.

Individual Intention

As we mentioned above, Joint Intention theory adopts Bratman's view of future-directed properties of intention. In this theory, an intention is defined to be a commitment to act in a certain mental state. In other words, an agent intends relative to some condition to do an action when it has a persistent goal or commitment (relative to that condition) of having done the action and, moreover, believing throughout that it is doing that action [54].

Intention inherits all the properties of commitment (e.g., consistency with mental states). Typically, an agent uses an intention as a decision within a subgoal-supergoal hierarchy to do a particular action. For instance, initially, the agent commits to p becoming true without having any concern about who or how p is going to be accomplished. Then, the agent commits to x or y as a mean to accomplish p. Lastly, the agent selects one of the actions (e.g., x) and forms an intention to do it. This intention will be given up when for whatever reason p is accomplished.

Joint Commitment

Before discussing joint commitment, we provide the definition of the *Weak Achievement Goal* (WAG) concept in Joint Intentions theory which shows the state of a team member nominally working on a goal. The concept of WAG is used to provide the definition of the Joint Commitment in this theory.

An agent has a WAG relative to q and with respect to a team to bring about p if either of the following conditions holds:

- The agent has a normal achievement goal to bring about p; that is, the agent does not yet believe that p is true and wants p to be true as a goal.
- The agent believes that p is true, will never be true, or is irrelevant, but has as a goal that the status of p be mutually believed by all the team members.

Joint commitment – A joint intention of a team Θ is based on its joint commitment, which is defined as a *Joint Persistent Goal* (JPG). A JPG to achieve a team action p, denoted JPG(Θ, p) requires all team members to mutually believe that p is currently false and want p to eventually be true. A JPG guarantees that team members cannot decommit until p is mutually known to be *achieved, unachievable* or *irrelevant*. Basically, JPG(Θ, p) requires team members to each hold p as a *Weak Achievement Goal* (WAG), where WAG(μ, p, Θ) in which μ is a team member in

Θ, requires μ to achieve p if it is false. However, if μ privately believes that p is either achieved, unachievable or irrelevant, JPG(Θ, p) is dissolved, but p is left with a commitment to have this belief become Θ's mutual belief. Such a commitment is required to establish mutual belief in Θ; this commitment typically makes an agent communicate with its teammates [54].

An important consequence of achieving joint commitment in a team is that it predicts future communication which is critical within the course of a collaboration. Thus, this communication leads team members to attain mutual beliefs which is a fundamental concept in teamwork activities. Notice that the minimum mutual belief for team members to attain is the achievement or failure of the shared goal which terminates collaboration.

Joint Intention

Joint intention is defined to be a joint commitment to the team members trying to do a joint action. Based on Cohen and Levesque's definition of joint intention, a team of agents jointly intends (relative to some escape condition) to do an action if and only if the members have a JPG (relative to that condition) of them having the action completed, and having it completed mutually believing throughout that they are doing it (knowingly) [54].

Teamwork & Communication

In summary, according to Joint Intentions theory, the notion of teamwork is characterized by joint commitment, also known as joint persistent goal (see Section 2.2.2). The definition of JPG states that the agents mutually believe they have the appropriate goal, and that they mutually believe a persistent weak achievement goal (which represents the one-way commitment of one agent directed towards another) to achieve it persists until the agents mutually believe that the goal has either been achieved, or become impossible, or irrelevant.

Joint Intentions theory claims that an efficient collaboration requires communication. Sharing information through communication is critical given that collaborators have different capabilities, and each individual often has only partial knowledge relevant to solving the problem, and sometimes diverging beliefs about the state of the collaborative activity. Communication is important in coordinating team members' roles and actions to accomplish their goal. For instance, it can help team members to establish and maintain a set of mutual beliefs regarding the current state of the collaboration, and the respective roles and capabilities of each member.

2.2.3 STEAM – A Hybrid Approach

Tambe in [250] argues that teamwork in complex, dynamic, multi-agent domains requires the agents to obtain flexibility and reusability by using integrated capabilities. Tambe created STEAM (simply, **S**hell **TEAM**work) based on this idea. STEAM's operationalization in complex, real-world domains is the key in its development to addressing important teamwork issues, some of which are discussed in Section 2.2.6. STEAM is founded on the Joint Intentions theory and it uses joint intentions as the basic building block of teamwork while it is informed by key concepts from SharedPlans theory.

Building on the well developed theory of joint intentions [54] and SharedPlans [101, 103], the STEAM teamwork model [250] was operationalized as a set of domain-independent rules that describe how teams should work together. According to Tambe, there are several advantages due to this use of Joint Intentions theory, such as achieving a principled framework for reasoning about coordination and communication in a team, which the joint intention can provide. Another advantage is the guidance for monitoring and maintenance of a team activity which the joint commitment in joint intention again provides. And lastly, Tambe believes the joint intention in a team can facilitate reasoning about team activity and team members' contribution to that activity.

However, he also believes that for a high level team goal, one single joint intention

is not sufficient to achieve all these advantages. Thus, STEAM borrows some of the concepts of SharedPlans theory. First, STEAM uses the concept of "intention that" (see Section 2.2.1) towards an activity as well as the fact that SharedPlans theory mandates team members' mutual belief in a common recipe and shared plans for individual steps in the common recipe. Thus, in this case, SharedPlans helps STEAM to achieve coherency within the teamwork. Besides, STEAM uses joint intentions to ensure the teamwork coherency to build the mental attitudes of team members. In other words, as the recipe evolves, STEAM requires all team members to agree on the execution of a step and form joint intentions to execute it while other joint intentions are formed, leading to a hierarchy. A second concept STEAM borrows from SharedPlans is the amount of information that a team member needs to know to perform an action. According to SharedPlans, team members require to know only that a recipe exists to enable them to perform actions (not recipe details – see Section 2.2.1). Similarly in STEAM, team members only track the subteam or individual team member responsible to perform a specific step; this tracking does not need detailed plan recognition. The third issue is parallel to what is called an unreconciled case in SharedPlans theory, which in STEAM is handled by replanning and communication between team members assigning the unassigned or unachieved task. The last issue is communication between team members which also borrows the concept of "intention that" from SharedPlans theory, to help the generalization of STEAM's communication capabilities beside what Joint Intentions theory offers.

In summary, STEAM builds on both Joint Intention theory and SharedPlans theory and tries to overcome their shortcomings. Based on joint intentions, STEAM builds up hierarchical structures that parallel the SharedPlans theory. Hence, STEAM formalizes commitments by building and maintaining Joint Intentions, and uses SharedPlans to formulate the team's attitudes in complex tasks.

In [250] Tambe argues that the novel aspects of STEAM relate to its teamwork capabilities. The key novelty in STEAM has team operators beside individual team member operators. In STEAM when agents select a team operator for execution,

they instantiate a team's joint intentions. Team operators explicitly express a team's joint activities, unlike the regular individual operators which express an agent's own activities. Hence, STEAM agents maintain their own private (to apply individual operators) and team states, e.g., mutual belief about the world (to apply team operators).

However, Tambe added more practical concepts into STEAM's architecture. For instance, STEAM has a team synchronization protocol to establish joint intention (see JPG in Section 2.2.2), or it has constructs for monitoring joint intentions which helps the agent to be able to monitor team performance. STEAM facilitates this monitoring by exploiting its explicit representation of team goals and plans. In particular, STEAM allows an explicit specification of monitoring conditions to determine achievement, unachievability or irrelevancy conditions of team operators. Finally, in STEAM, communication is driven by commitments embodied in the Joint Intentions theory, i.e., team members may communicate to obtain mutual belief while building and disbanding joint intentions. Thus, joint intentions provide STEAM with a principled framework for reasoning about communication. Also, STEAM addresses some practical issues, not addressed in other teamwork theories. One of these issues is STEAM's detailed attention to communication overheads and risks, which can be significant [249]. Furthermore, operationalization of STEAM is based on enhancements to the Soar architecture [137], plus a set of about 300 domain-independent Soar rules.

2.2.4 Other Approaches

There are other frameworks, approaches, and models focusing on teamwork and collaborative agents. For instance, Jennings provides the Joint Responsibility framework which is specified formally using modal, temporal logic. Joint Responsibility stresses the role of joint intentions (based on Joint Intentions theory) specifying how both individuals and teams should behave whilst engaged in collaborative problem solving [122, 123, 124, 125]. Jennings has developed *Generic Rules and Agent model*

Testbed Environment (GRATE) as a prototype system based on the Joint Responsibility framework. In [131] Kinny et. al. elaborate the concept of Planned Team Activity and introduce a language for representing joint plans for teams of agents and describe how agents can organize the formation of a skilled team to achieve a joint goal. They use joint intentions to capture the mental properties which characterize team activity.

2.2.5 Similarities and Differences

There are some similarities between SharedPlans and Joint Intentions theories. Here, we specify some of these similarities:

1. Similar to SharedPlans theory, Joint Intentions theory specifies what it means for agents to execute actions as a team [247].

2. Both theories follow Bratman's basic ideas about the roles of intention in relational actions which prevent the collaborative agents from adopting conflicting intentions. Besides, these two theories also follow Bratman's BDI model.

3. Just as SharedPlans theory, Joint Intentions theory also states that a joint action could not be seen as a collection of individual actions but as agents working together who need to share beliefs.

4. Both theories in their latest articles show that the agents are required to communicate to maintain collaboration. SharedPlans theory requires collaborators to communicate to establish and maintain the shared plan which is crucial especially when collaborators only have a partial shared plan. Similarly in Joint Intentions theory, communication is an explicit requirement of collaborative agents until the shared goal is achieved, unachievable or irrelevant.

5. Both Joint Intentions and SharedPlans theories are concerned about commitment to the joint activity. Although, these two theories use different concepts to fulfill the requirements of commitment during collaboration.

There are also differences between SharedPlans and Joint Intention theories; we address some of them here:

1. The crucial components of the SharedPlans theory (see Section 2.2.1) lack the notion of a joint intention, which is the most significant notion within the Joint Intentions theory; Grosz and Sidner do not believe that such a phenomenon (joint intention) exists in a collaboration. They believe their notion of "intention that" and mutual beliefs about states of the collaboration can provide similar functionalities as described in Joint Intentions theory (see Section 2.2.2).

2. In SharedPlans theory teammates agree on the shared plan, whereas in Joint Intentions theory teammates agree on intentions.

3. In contrast to Joint Intentions, the SharedPlans theory employs hierarchical structures over intentions, thus it overcomes the shortcoming of a single joint intention for complex team tasks.

4. The SharedPlans theory describes the way to achieve a common goal through the hierarchy of plans, whereas the Joint Intentions theory describes only this common goal [240].

5. Joint Intentions theory assumes that knowledge about the teammates is always available, whereas SharedPlans theory uses the concept of partial plan/recipe to make the process of dynamically achieving information possible through out the collaboration.

6. Communication requirements are derived from "intention that" in SharedPlans theory, as opposed to being "hard-wired" in Joint Intentions theory.

A critique to Joint Intention theory – Castelfranchi criticizes the necessary and sufficient conditions (see Section 2.2.2) for the joint persistent goal which plays a crucial role in the Joint Intentions theory. According to his example, if a French

scientist and an American scientist are both working on an AIDS vaccine and both have the final goal of p "vaccine anti-AIDS be found" relative to the belief q that "if vaccine is found, AIDS is wiped out", they both share the mental attitudes described in Joint Intentions theory. It means that they mutually believe that p is currently false, and they mutually know they both want p to be true, and it is true that until they come to believe either that p is true, that p will never be true, or that q is false, they will continue to mutually believe that they each have a weak achievement goal (see Section 2.2.2) relative to q and with respect to the team (i.e., the WAG with respect to the team has been defined as "a goal that the status of p be mutually believed by all the team members"). The problem is that we can not claim the French and American professors are working as a team. In fact, given their personal goals of finding the vaccine, they might come to strongly compete with each other [45].

2.2.6 Applications of Collaboration Theories

There are many research focusing on different aspects of collaboration based on different collaboration theories, i.e., SharedPlans, Joint Intentions, and hybrid theories of collaboration. In this section, we provide some examples of homogeneous and heterogeneous agent/robot and human collaborations.

There are some works focusing on the concepts of robot assistants [49], or teamwork and its challenges at cognitive and behavioral levels [178, 215]. Some researchers have an overall look at a collaboration concept at the architectural level. In [82] authors present a collaborative architecture, COCHI, to support the concept of emotional awareness. In [75] authors present the integration of emotional competence into a cognitive architecture which runs on a robot, MEXI. In [245] authors discuss the challenges of integrating natural language, gesture understanding and spatial reasoning of a collaborative humanoid robot situated in space. The importance of communication during collaboration has also been considered by some researchers from human-computer interaction and human-robot collaboration [48, 166, 205] to

theories describing collaborative negotiation, and discourse planning and structures [5, 102, 237]. There are other concepts such as joint actions and commitments [99], dynamics of intentions during collaboration [146], and task-based planning providing more depth in the context of collaboration [39, 203]. The concept of collaboration has also received attention in the industry and in research in robotic laboratories [93].

Applications of SharedPlans Theory – COLLAGEN [204, 205] is the first implemented system based on the SharedPlans theory. It incorporates certain algorithms for discourse generation and interpretation, and is able to maintain a segmented interaction history, which facilitates the discourse between the human user and the intelligent agent. The model includes two main parts: (1) a representation of a discourse state and (2) a discourse interpretation algorithm for the utterances of the user and agent [206]. In [108] Heeman presents a computational model of how a conversational participant collaborates in order to make a referring action successful. The model is based on the view of language as goal-directed behaviour, and in his work, he refers to SharedPlans as part of the planning and conversation literature. In [153], Lochbaum and Sidner modify and expand the SharedPlan model of collaborative behavior [103]. They present an algorithm for updating an agents beliefs about a partial shared plan and describe an initial implementation of this algorithm in the domain of network management. Lochbaum, in [152], provides a computational model (based on the collaborative planning framework of SharedPlans [101]) for recognizing intentional structure and utilizing it in discourse processing. In short, she presents a SharedPlans model for recognizing Discourse Segment Purposes (DSPs) [103] [237] and their interrelationships. CAST (Collaborative Agents for Simulating Teamwork) [267] [268] is a teamwork framework based on the SharedPlans theory. CAST focuses on flexibility in dynamic environments and on proactive information exchange enabled by anticipating what information team members will need. Petri Nets are used to represent both the team structure and the teamwork process, i.e., the plans to be executed. Researchers in [114]

discuss developing an ontology of microsocial concepts for use in an instructional system for teaching cross-cultural communication. They believe being acquainted with one another is not a strong enough relationship from which to create a society. Hence, there is a need for commitment and shared plans (as the basis of social life) to achieve a shared goal. In this work, Grosz and Sidner's SharedPlans theory [103] is used to explain the concept of shared plans within the interpersonal relationships of societies in an industrial environment. In [119] Hunsberger and Grosz discuss the idea of whether the rational, utility-maximizing agents should determine commitment to a group activity when there is an opportunity to collaborate. They call this problem the "initial-commitment decision problem" (ICDP) and provide a mechanism that agents can use to solve the ICDP. They use the representation of action, act-types and recipes in the SharedPlans theory. In [270] an integrated agent-based model for Group Decision Support Systems is proposed and discussed. The decisional model that authors outline in this paper is based on the Shared-Plans theory. Rauenbusch and Grosz in [198] formally define a search problem with search operators that correspond to the team planning decisions. They provide an algorithm for making the three types of interrelated decisions by recasting the problem as a search problem. Their model respects the constraints on mental states specified by the SharedPlans theory of collaboration. Babaian et. al. in [12] describe Writer's Aid, a system that deploys AI planning techniques to enable it to serve as an author's collaborative assistant. While an author writes a document, Writer's Aid helps in identifying and inserting citation keys and by autonomously finding and caching potentially relevant papers and their associated bibliographic information from various on-line sources. They believe the underlying concepts of SharedPlans is relevent since in collaborative interfaces like Writers Aid, the users establish shared goals with the system and user and the system both take initiative in satisfying them. In [171] researchers address high-level robot planning issues for an interactive cognitive robot that acts in the presence of or in collaboration with a human partner. They describe a Human Aware Task Planner (HATP) which is

designed to provide socially acceptable plans to achieve collaborative tasks. They use notions of plans based on SharedPlans theory. In [238] Sidner and Dzikovska argue that robots, in order to participate in conversations with humans, need to make use of conventions of conversation and the means to be connected to their human counterparts. They provide an initial research on engagement in human-human interaction and applications to stationary robots in hosting activities. They believe hosting activities are collaborative because neither party completely determines the goals to be undertaken nor the means of reaching the goal. To build a robot host, they rely on an agent built using COLLAGEN which is implemented based on the SharedPlans theory.

Applications of Joint Intentions Theory – In [131] authors introduce a language for representing joint plans for teams of agents. They describe how agents can organize the formation of a suitably skilled team to achieve a joint goal, and they explain how such a team can execute these plans to generate complex, synchronized team activity. In this paper, authors adopt the underlying concepts of the Joint Intentions theory as the structure of their collaborative agents. Breazeal et. al. in [34] present an overview of their work towards building socially intelligent, cooperative humanoid robots, such as Leonardo, that can collaborate and learn in partnership with humans. They employ the Joint Intentions theory of collaboration to implement the collaborative behaviors while performing a task in collaboration with humans. In [247] the researchers' goal is to develop an architecture (based on the concepts of Joint Intentions theory) that can guide an agent during collaborative teamwork. They describe how a joint intention interpreter that is integrated with a reasoner over beliefs and communicative acts can form the core of a dialogue engine. Ultimately, the system engages in dialogue through the planning and execution of communicative acts necessary to attain the collaborative task at hand. Mutlu et. al. in [175] discuss key mechanisms for effective coordination toward informing the design of communication and coordination mechanisms for robots. They present two

illustrative studies that explore how robot behavior might be designed to employ these mechanisms (particularly joint attention and action observation) to improve measure of task performance in human-robot collaboration. Their work uses Joint Intentions theory to develop shared task representations and strategies for task decomposition. The system GRATE* by Jennings [124] is based on the Joint Intention theory. GRATE* provides a rule-based modelling approach to cooperation using the notion of Joint Responsibilities, which in turn is based on Join Intentions. GRATE* is geared towards industrial settings in which both agents and the communication between them can be considered to be reliable.

Applications of Hybrid Theories – This domain independent teamwork model, STEAM, has been successfully applied to a variety of domains. From combat air missions [112] to robot soccer [134] to teams supporting human organizations [196] to rescue response [216], applying the same set of STEAM rules has resulted in successful coordination between heterogeneous agents. The successful use of the same teamwork model in a wide variety of diverse domains provides compelling evidence that it is the principles of team-work, rather than exploitation of specific domain phenomena, that underlies the success of teamwork based approaches. In [159] authors provide their RoboCup (robotics soccer testbed) in which their focus is on teamwork and learning challenges. Their research investigation in RobotCup is based on ISI Synthetic, a team of synthetic soccer-players. They also investigate the use of STEAM as their model of teamwork which is influenced by the Joint Intentions and SharedPlans theories. In [127] researchers propose a behavioral architecture C^2BDI that allows the enhancement of the knowledge sharing using natural language communication between team members. They define collaborative conversation protocols that provide proactive behavior to agents for the coordination between team members. Their agent architecture provides deliberative and conversational behaviors for collaboration, and it is based on both of the SharedPlans and Joint Intentions theories.

2.3 Emotions and Affective Computing

According to Picard [188], the term affective computing encapsulates a new approach in Artificial Intelligence, to build computers that show human affection. Studies show that the decision making of humans is not always logical [95], and in fact, not only is pure logic not enough to model human intelligence, but it also shows failures when applied in artificial intelligence systems [69].

If we want robots and virtual agents to be more believable and efficient partners for humans, we must consider the personal and social functionalities and characteristics of emotions; this will enable our robots to coexist with humans, who are emotional beings. To have a better understanding of applications of affective computing, we can categorize the whole existing literature of computational emotion modeling and their applications into four major categories of: a) detecting and recognizing human emotions, b) interpreting and understanding human emotions, c) generating artificial emotions and applying the underlying processes to exploit emotion functions, and d) expressing human-perceivable emotions during interaction.

There are some major emotion theories including *appraisal, dimensional* and *discrete (basic)* theories, some of which have corresponding computational models, e.g., EMA [162] and WASABI [22, 23]. These models have been used in different domains including AI and robotics. Modeling and applying these theories can help robots and virtual agents to achieve communicative, evaluative, interpretive, and regulatory aspects of emotions in some or all of the four application domains we mentioned above.

This section provides descriptions of major computational emotion theories, their comparison, and their applications in AI and robotics. It includes the existing influential computational emotion theories as well as the underlying psychological theories; it majorly focuses on appraisal and dimensional theories, although it briefly mentions other approaches, e.g. discrete (basic) emotions.

2.3.1 Affect and Emotions

Emotion affects not only what people do, but also the way they do it [60]. Aristotle in *The Nicomachean Ethics* reveals his idea about emotions. He says "Anyone can become angry–that is easy. But to be angry with the right person, to the right degree, at the right time, for the right purpose, and in the right way–this is not easy [7]."

Intelligence is the process that humans use to explain the different degrees of adaptive success in one's behavior. It is a set of mental abilities that enables a human to comprehend, reason and adapt in the environment, and as a result, act effectively and purposefully in that environment. Emotions play a crucial role in humans' explanation of intelligent behaviors. Emotions significantly impact the procedures of action generation, execution, control, and interpretation [273] in different environments. Emotions are conceptualized as ongoing processes rooted in dynamic social contexts, which can shape both implicit and explicit emotional responses [156]. An emotion is a dynamic episode that not only makes changes in cognitive states, but also produces a sequence of response patterns on body movements, posture, voice and face [223]. Emotions typically occur in response to an event, usually a social event, real, remembered, anticipated, or imagined. They are associated with distinctive relational meanings [186]. These relations can be with the individual's past experience, the individual's surrounding objects and environment, or the other individuals with or without mutual beliefs in a dyadic or a group setting. Emotions are evaluative and responsive patterns that serve the function of providing appraisal about whether the ongoing event is harmful, threatening or beneficial for the well-being of an individual [273]. Consequently, reasoning and emotional processes have an integral and a supportive relationship, rather than an antagonistic and a conflicting one.

A better question than what emotions are, is the question of what they can do, and how they impact humans' life. Emotions impact fundamental parts of cognition

including perception, memory, attention and reasoning [50]. This impact is caused by the information emotions carry about the environment and event values. The influence of emotions depends on an individual's focus of attention. For instance, a positive affect can cause a positive attitude towards an object if the individual's focus is on the object, whereas the same positive affect can be interpreted as a positive feedback towards one's partner during the course of a collaboration. As another example, a positive feedback can promote certain cognitive processes, or it can inhibit other cognitive processes according to the conditions in the environment [51]. In both cases, emotions play a regulatory role for cognitive processes [94]. Some of the effects flow from underlying shifts in the way people perceive and think under the influence of emotion.

2.3.2 Emotion in Social Context

In this section, we discuss the importance of studying emotions within a social context. This perspective is important in our research because our work is focused on collaboration as a particular social setting between individuals. Understanding the dynamics of collaboration requires one to understand influential underlying components. We have chosen to study emotion as a crucial underlying component in humans' social life which will be discussed in detail throughout this section.

Emotions are involved in developing social contexts. Humans are social and most of the causations and constitutions of their emotions are social. Brian Parkinson in [184] argues that many of the causes of emotions are interpersonal and communicative rather than internal and reactive phenomena. There are different social aspects of emotions influenced by various factors such as social context and social relationship type. For instance, a dominant-submissive social relationship can cause and contain different emotions with different intensities compared to a reciprocal or a friendship social relationship type. As another example, an emotion can be interpreted in a certain way when an individual is situated in an environment with other people who are expressing a particular emotion.

As mentioned earlier, the social context is an important factor influencing one's emotions. A dyadic interaction is one type of a setting in a social context. Dyadic interaction tasks allow us to study emotion in a social setting [52]. Dyadic interaction tasks make it possible to examine how individuals experience and express emotions during social interactions and how emotions shape and are shaped by the reciprocal interactions between individuals. In addition, eliciting and monitoring emotional processes yields useful information about the role emotion plays in interpersonal relationships. Compared with other emotion-eliciting events, events in a dyadic interaction can better help us study an ongoing emotional relationship between two individuals in addition to their internal emotional and cognitive processes. Dyadic interaction tasks are ideal for studying a range of emotional responses because of the fairly unstructured conversations between the individuals. Thus, dyadic interaction tasks will generate a wide range of emotions in comparison with the controlled emotion-eliciting events.

There are numerous ways that emotions can be social [255]. There is a consensus on the fact that social events and entities surrounding the individual play an essential role in the generation of emotion. There are several ways in which other people elicit emotional responses in us. One is that we feel the emotions of those around us. Also, we have emotions about actions of those people around us. Another is we have emotions about the things that happen to other people. Yet another is our concern about our relationship with others that elicits emotion in us. The groups to which we belong can also elicit our emotions. Moreover, we can feel emotion about the success and failure of our own group or of other groups. In addition, groups or individuals may make salient cultural concerns or societal expectations that can elicit our emotions.

Beside the fact that social context can elicit emotions in individuals, social context provides information about what emotion should be expressed, by whom, and in what situations. For instance, people are well aware of the inappropriateness of expressing too much emotion to acquaintances [255]. However, the social knowl-

edge of emotion expression is only partially delivered in an explicit fashion. There are studies on the regulatory role of society and social relationships on emotions, showing that people's emotions become socialized in implicit and unconscious ways. From this perspective, social context can control and direct our attention toward certain types of events and away from others.

Humans are emotional and social beings. Their emotions and the social context in which they are involved have mutual impacts on each other. But, what if humans can share their emotions with others just as they share their thoughts, resources and their environment. Sharing an emotion with others may alter the experience of an event. For instance, according to the nature of the relationship between the individuals, the expression of emotions can either restrain them from further interactions or improve their relationship. Furthermore, individuals sharing emotions might possess a shared understanding of their environment. Socially shared and regulated emotions also provide social meanings to the events happening in the environment [265]. For instance, people are likely to make social inferences based on the presence or absence of particular emotions in their social environment. Moreover, emotions can provide a basis for judgment depending on the individual's relationships with others. In other words, emotions can associate or disassociate an individual, therefore, they can change or maintain the individual's social relationships [255].

Emotions can also play the role of a motivator in a social context. There is a subset of social emotions delineated as role-taking emotions in [234]. Shott provides two categories of *reflexive* (e.g. shame or pride) and *empathic* (e.g., empathy or pity) role-taking emotions. The reflexive emotions can motivate the individual's self-control which depends on the anticipated reactions of others to the individual's behaviors. For instance, guilt might lead the individual to behave altruistically to restore a positive social stance for that individual. Empathic or vicarious emotions are based on an individual mentally placing himself in other's situation to understand how the other feels in that situation. These emotions motivate prosocial behaviors to maintain an individual's internal well-being [253].

2.3.3 Communicating Emotions

Humans need to communicate their emotions within the social context for different reasons. In [85] Goffman argues that human behaviors around others are performative which is often intended to convey information to others. When human's actions are visible in the social context, they behave differently in the presence of the others [269]. The social life of an individual is comprised of the individual's internal cognitive competencies and his interactions in the society. Lazarus says, if society is a fabric, then emotion is its color [140]. Although emotions undeniably have personal aspects, they are usually experienced in a social context and acquire their significance in relation to this context [156].

There are several events that can elicit emotions in social contexts. For instance, during the interaction the cause of an emotion can be verbal (an utterance during conversation), nonverbal (someone's gesture), personal thoughts (interpretation of an event), or even emotions themselves (e.g., happiness for a partner's sense of pride). An utterance can include content and relational meaning. The content carries the information about the topic or the subject of the interaction, and the relational meaning reveals the meaning between the speaker and the hearer. An emotion might seem to be elicited by the content of the utterance, but in fact it is an individual's response to the relational meaning [189].

The interpretation of these relational meanings are handled by the appraisal of the events. Appraisal processes (see Section 4.3) also give us a way of viewing emotion as social [259]. Meaning is created by an individual's social relationships and experiences in the social world. Individuals communicate these meanings through utterances. Utterances in emotionally charged conversations, by their very nature, are supposed to inform the others about something novel. Novelty is an essential component of an event for appraisal. Conversations also possess the concept of consistency because the utterances with consistent meaning constitute the individual's underlying beliefs. Relevancy is another component of an event that can be assessed

by appraisal. The degree to which the individual's personal and mutual beliefs are strong and related controls emotionally rich social contexts. In other words, the more divergent the individual's beliefs, the more effort is required to converge (to be understood) which leads to more emotional responses in individuals. From another point of view, human speech carries emotional information in the semantics and in the speech prosody. The semantics or the content of what an individual says includes obvious expression of emotion. However, the prosody holds more detailed emotional information by combining non-semantic cues in spoken language (e.g., rhythm and intonation) [154].

Interpretation of the events in the social context requires a baseline for the individual's assessment process. Goals as the pillar of collaborative interactions can provide this baseline for an individual. Goals are crucial in relational meanings of the events in a social context. The facilitation, interference and inhibition of goals are each correlated with certain type of emotions. In most conversations during collaboration goals can be categorized into three different groups: goals related to accomplishing a task, goals to reveal one's personal beliefs, and goals to regulate one's social relationships [189]. For instance, for task-related goals, utterances related to accomplished tasks reveal joyful relational meaning; utterances related to impeded tasks reveal disappointing relational meanings which can lead to anger, and utterances related to tasks with no or little progress reveal the frustration of the individuals. Lastly, all these emotional responses in a social context will not only regulate or maintain individual's actions to reveal or hinder an intention, but also can control the way that action should be taken.

A successful and effective emotional communication necessitates ongoing reciprocal adjustments between interactants that can happen by interpreting each other's behaviors [156]. It not only requires proper interpretation of the other's expressions, but also correct assessment of the extent to which others can read an individual's expressions. In emotional communication, individuals are constantly exchanging messages about their mental states, and modifying each other's emotional responses

as they occur. Individuals perceive other's emotional states through verbal and nonverbal responses during the interaction by processing relevant messages. Communication dynamics represent the temporal relationship between these communicative messages. The verbal and nonverbal messages from one participant are better interpreted inside the correct context including the history and the ongoing messages from the other individuals. Interpersonal dynamics (also known as micro-dynamics in sociology) represent this influence of relationships between individuals [172].

2.3.4 Social Functions of Emotions

Humans are able to communnicate their emotions in a social context. The social functions of emotions are the reason behind why humans try to communicate their emotions. Ekman in [70] asserts that the primary function of emotions is to mobilize the organism to deal with important interpersonal encounters. Darwin in [62] argues the significance of social communicative functions of emotions. Emotions describe interpersonal dynamics in a way that they can constitute individuals' relationships [184, 255]. One aspect of expressing and communicating emotion in a social context is to express one's social motives and intentions [110]. Another aspect of communicating emotions is to reveal the underlying mental states of an individual [185]. In other words, emotions constitute two different functionalities of expressing communicative signals associated with one's social motives and intentions as well as expressing one's internal states and how one feels about something. In [135] Van Kleef has discussed the idea of inferential processes with which individuals can infer information about others' feelings, relational orientations and behavioral intentions based on their emotional expressions. He also argues that emotional expressions can impact social interactions by eliciting others' affective responses.

Functional accounts vary according to the kind of system being analyzed. Therefore, functional approaches to the emotions should vary by level of analysis. Social functions of emotions can be analyzed in *individual, dyadic, group* and *cultural* levels. The focus of this research is on social functions in dyadic interaction (more

specifically collaboration); these functions are also considered at the individual's level especially when interpreting the other collaborator's behaviors. Studies in all these levels share a few assumptions about social accounts of emotions. They assume a) individuals are social by nature and pursue solutions to survival problems in social relationships, b) individuals apply their emotions to coordinate their social interactions and relationships to address these survival problems, c) emotions are processes mediating the individuals' relations to their dynamic environment [129]. In dyadic interactions, studies focus on how emotions impact the interactions of individuals in meaningful relationships. In [129] Keltner and Haidt discuss that in a dyadic setting, researchers mostly focus on communication of emotion (e.g. Scherer [218], DePaulo [66]), properties (e.g. emotion contingency, emotion synchrony) of dyadic emotions (e.g. Levenson & Gottman [144]), discourse (e.g. Bretherton [36]), and attachments (e.g. Hazan & Shaver [107]).

Examples of Social Emotions:

There are many different types of emotions, some of which are considered social, since they appear and provide meaning in social context. Here, we provide four examples of these emotions as well as their social functions to show how social functions of emotions impact individuals and the groups they belong to, and what causes them to be expressed by an individual.

Guilt – The function of guilt is to positively direct our behavior toward our group. We feel guilt when we hurt someone in our group, or when we fail to reciprocate care or kindness. Guilt motivates us to not hurt people in our group and to give back to others who have given to us, and in this way we strengthen the survival prospects of both the group and ourselves.

Shame – The function of shame is twofold. On the one hand, it keeps us within the rules and norms of society by informing us when we have done something dishonorable, disgraceful, or in some way condemned by our group. On the other hand,

it informs the other members of our group that we know that we have dishonored ourselves. The main difference between guilt and shame is that guilt is focused on a behavior, whereas shame is focused on ourselves.

Embarrassment – Embarrassment is related to shame, but includes some important differences. Embarrassment can only happen in public, whereas shame can happen when we are alone. We can feel embarrassment about very minor issues that have no moral implications, such as body odor, whereas shame typically concerns more grave issues with moral implications.

Pride – The function of pride is to reinforce when we or another person has done or represented something the group finds excellent. In this way, group values are reinforced and incentivized, which again helps the group to function better and motivates us to do things the group values. There is a negative form of pride in which our internal appraisal of our worth is inflated compared to the opinions of others, which is more correctly called hubris.

2.3.5 Artificial Emotions

Emotions, as an integral part of rational behavior, provide adaptive values for an artificial creature. They can control an agent's *attention* to focus on the most salient and relevant stimulus to solve the immediate problem. They can also help an agent to *monitor its own performance* so that the agent can make alterations on goals and plans. Emotions can act as a *memory filter* allowing a better recall of the events that are congruent with current cognitive and emotional states [30]. *Assisting the reasoning process* is another role of emotions; they assist the reasoning process by directing the cognitive information processes to the perceptual cues. Emotions impact the transformation of the agent's *decision-making behavior* [84] leading to a particular type of actions in a certain type of environment [273]. Emotions can *govern behavior tendencies* by providing immediate emotional responses, e.g., avoidance of elaborate reasoning because of lack of time or an unconcerned situation. Furthermore, emotions *provide support for social interactions* by helping the agent to

understand others' behaviors as well as making expressions of the agent's internal states more perceivable during the interaction [81].

The importance of these values of emotions for designing social agents having artificial emotions is prominent. However, the question is what problems are we facing in designing an effective social agent? In [63] authors discuss some of these problems and provide references speculating on the nature, function and mechanisms of emotions. Also, the importance of emotions and the incorporation of emotions in intelligent systems as well as implementation of emotions in several multi-agent systems are presented in [164]. Scheutz discusses the role of emotions in artificial intelligence and how we can determine the utility of emotions for the design of an artificial agent [226]. In [25] authors present a definition and theory of artificial emotions viewed as a sequential process comprising the appraisal of the agent's global state; they also show how emotions are generated, represented and used in the Salt and Pepper architecture for autonomous agents. From the behavior perspective, appropriately timed and clearly expressed emotions are a central requirement for believable agents [19].

There are several architectures modelling emotions for the purpose of enhancing the believability and effectiveness of the agents and robots. But the question is how do we model emotions? Hudlika in [117] deconstructs the concept of emotion modelling into: (a) fundamental categories of processes for emotion generation and emotion effects, and (b) identification of some of the fundamental computational tasks required to implement these processes. These building blocks can be helpful as a guideline for the systematic development of new computational models, or for the assessment of existing computational models of emotions as discussed in [148] and [163]. There are also logical formalizations of emotions and emotional attitudes (including speech acts) and corresponding mental states to provide a systematic analysis of computational models of emotions [1, 88, 104].

From another perspective, the necessity of employing emotions in robotics and more specifically social robots has been argued in [187] and [243]. Social robotics

and cognitive robotics have many overlapping concepts, especially when they focus on interaction between a robot and a human. The relationship between cognition and emotion receives more attention due to the mutual influences they have on each other [173, 229]. For instance, in [81] authors employ emotions in the learning procedure of a robot, and in [38] and [225] authors discuss the importance of emotions in the action selection procedure of an agent or a robot, impacting the behavior arbitration and self-adaptation mechanisms. Ultimately, employing artificial emotions will impact the context of human-robot/computer interaction [115] and how humans and robots understand each other's emotions in a social environment [132, 176]. In [208] authors selected twelve autonomous agents that incorporate an emotion mechanism into the action selection procedure to compare. They introduced a framework based on correlations between emotion roles performed and aspects of emotion mechanisms used to perform those roles. Gratch and Marsella also present one method to evaluate a computational model of emotion in [90] which compares behavior of the model against human behavior.

2.3.6 Cognitive Architectures

There are several integrated cognitive architectures trying to produce all aspects of behavior as a single system while remaining constant across different domains [126, 137]. The comparison of underlying philosophy and functional description of the most prominent cognitive architectures have been surveyed; several criteria are provided to evaluate such architectures [47, 139, 254]. The necessity of integrating these cognitive architectures with robots has been discussed from the perspective of developmental psychology [10, 67, 128]. There are also many examples emphasizing the importance of cognitive robotics from this perspective. Some of these cognitive architectures are biologically inspired, e.g., *eBICA* [212], or [20] and [200], while some others are inspired by psychological theories, e.g., $ACT - R\Phi$ [61], or [170] and [68], while some of them also incorporate the concept of affect in their design [40].

2.4 Computational Models of Emotions

There are different types of computational theories of emotion. These theories differ in the type of relationships between their components and whether a particular component plays a crucial role in an individual emotion. For instance, the basic component of an emotion can be the behavioral tendencies, the cognitive elements, or the somatic processes. Emotion theories can also differ based on their representationial distinction.

2.4.1 Appraisal Theory

Appraisal theories of emotion were first formulated by Arnold [9] and Lazarus [140] and then were actively developed in the early 80s by Ellsworth and Scherer and their students [207] [213] [217] [222] [224]. The emotional experience is the experience of a particular situation [79]. Appraisal theory describes the cognitive process by which an individual evaluates the situation in the environment with respect to the individual's well-being and triggers emotions to control internal changes and external actions.

Componential Approach

This approach emphasizes the distinct components of emotions, and is often called the *componential* approach [145]. The "components" referred to in this approach are the components of the cognitive appraisal process. These are referred to as *appraisal variables*, and include *novelty, valence, goal relevance, goal congruence*, and *coping abilities* (further on, in this section, some of the appraisal variables used in computational models are introduced) [217, 224]. A stimulus, whether real or imagined, is analyzed in terms of its meaning and consequences for the agent, to determine the affective reaction. The analysis involves assigning specific values to the appraisal variables. Once the appraisal variable values are determined by the organisms evaluative processes, the resulting vector is mapped onto a particular emotion, within

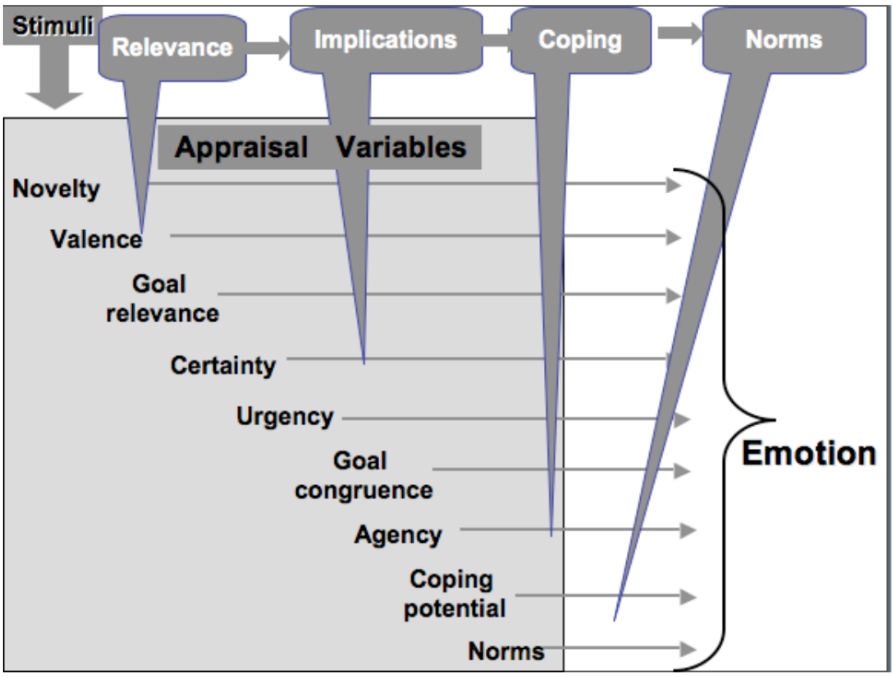

Figure 2.2: Schematic view of the componential theory of emotion [118].

the n-dimensional space defined by the n appraisal variables. The semantic primitives for representing emotions within this model are thus these individual appraisal variables. Figure 2.2 shows the relationship of the individual appraisal dimensions to the broader categories of evaluations taking place during appraisal (Relevance, Implications, etc.).

Component Process Model

The Component Process Model (CPM) is Scherer's influential and major theory of emotions [219, 224]. This theory focuses on the dynamic unfolding of emotions. The CPM suggests that an event and its consequences are appraised with a set of criteria on multiple levels of processing (the appraisal component). The result of the appraisal will generally have a motivational effect, often changing or modifying the motivational state before the occurrence of the event. Based on the appraisal results and the motivational changes, some effects will occur in the autonomic and somatic nervous system. The CPM considers emotions as the synchronisation of many dif-

ferent cognitive and physiological components. Emotions are identified with the overall process whereby low level cognitive appraisals, in particular the processing of relevance, trigger bodily reactions, behaviours and subjective feelings. The model suggests that there are four major appraisal objectives required to adaptively react to a salient event [221]:

a) **Relevance:** How relevant is this event for the agent? Does it directly affect the agent or its social reference group?

b) **Implications:** What are the implications or consequences of this event and how do they affect the agent's well-being and its immediate or long-term goals?

c) **Coping Potential:** How well can the agent cope with or adjust to these consequences?

d) **Normative Significance:** What is the significance of this event for the agent's self-concept and for social norms and values?

To attain these objectives, the agent evaluates the event and its consequences on a number of criteria or *Stimulus Evaluation Checks* (SECs), with the results reflecting the agents subjective assessment of consequences and implications on a background of personal needs, goals, and values [224]. Figure 2.3 shows the postulated sequence, the cognitive and motivational inputs and the effects on response systems. Also, the bidirectional effects between appraisal and other cognitive functions are illustrated by the arrows in the upper part of Figure 2.3.

Appraisal Process

According to this theory, appraisals are separable antecedents of emotion, that is, the individual first evaluates the environment and then feels an appropriate emotion [224]. The appraisal procedure begins with the evaluation of the environment according to the internalized goals and is based on systematic assessment of several elements [219]. The outcome of this process triggers the appropriate emotions.

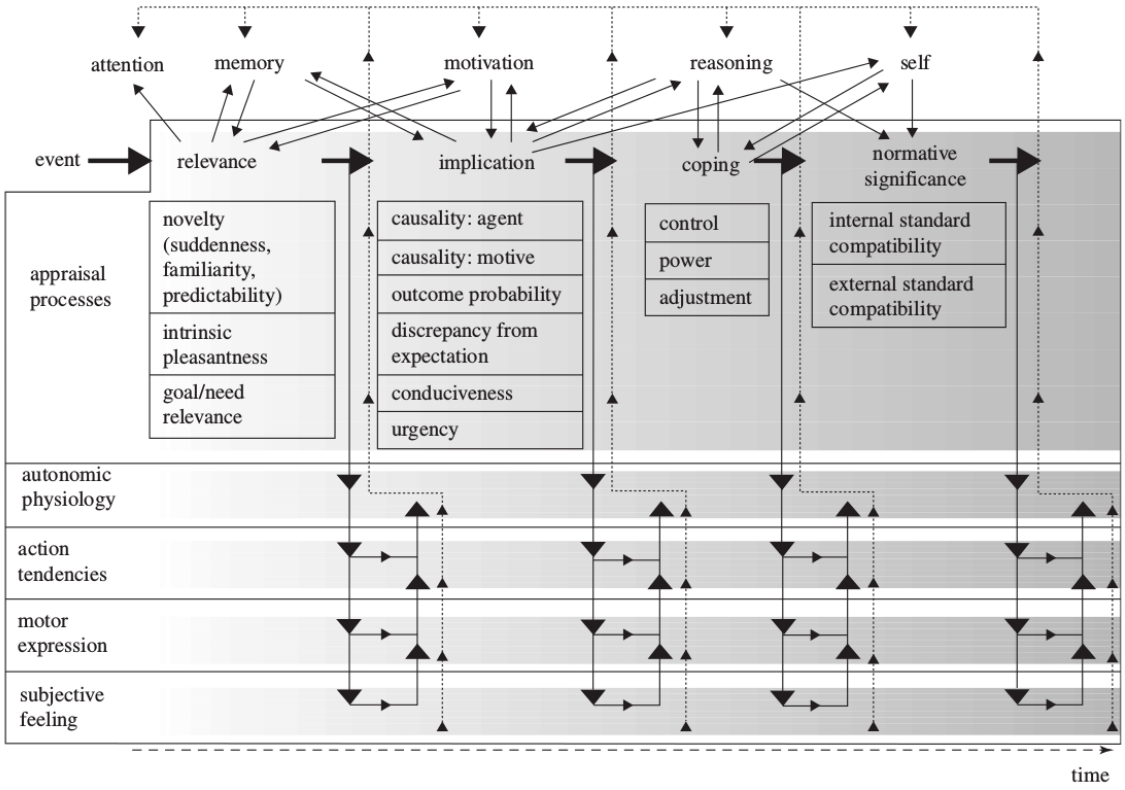

Figure 2.3: Comprehensive illustration of the CPM of emotion [221, 224].

In many versions of appraisal theory, appraisals also trigger cognitive responses often called *coping strategies*. In fact, the coping mechanism manages the individual's action with respect to the individual's emotional state and the existing internal and/or external demands [77]. The large majority of computational models of emotions are based on this theory. An individual can also use knowledge about the emotional reactions of others to make inferences about them. According to the appraisal patterns, different emotions can be experienced and expressed. Since expression of emotions reflects one's intentions through the appraisal process, the *reverse appraisal* mechanism helps one to infer others' mental states based on their expressions. [64, 106].

Appraisal process is typically viewed as the cause of emotion and the cognitive and behavioral changes associated with emotion. For instance, a particular pattern of the appraisal variables (i.e., individual judgements) will elicit a certain emotion

or emotional expressions. These appraisal variables include [162]:

- **Relevance:** A relevant event has non-zero utility for an agent. This relevancy can either be based on a negative influence of an event on the agent or a positive one.
- **Perspective:** The point of view in which an event will be judged, e.g. self or other.
- **Desirability:** A desirable event advances a state of the utility for an agent whose perspective is being taken, or if it is an undesirable event, inhibits that.
- **Likelihood:** A measure of likelihood of the outcome.
- **Expectedness:** The extent to which the truth value of a state could have been predicted from causal interpretation.
- **Causal Attribution:** The agent who deserves the credit/blame.
- **Controllability:** Whether the outcome can be altered by the agent whose perspective is taken (this variable is related to the coping process).
- **Changeability:** Whether the outcome can be altered by some other causal agent (this variable is related to coping process).

Coping Process

Another key process involved in appraisal is the coping process. This process determines whether and how the agent shoud respond with respect to the outcome of appraising the events. There are several coping strategies that computational models like EMA [92] use as control signals. These control signals enable or suppress the cognitive processes that operate on the causal interpretation of the appraisal patterns. The coping process controls the congruency of the actions according to these

patterns. As it is shown below, in [92] coping strategies are organized into two categories: *problem-focused* and *emotion-focused*. Problem-focused coping strategies can be applied when the agent must do something with respect to the problem, whereas Emotion-focused coping works by changing one's interpretation of circumstances. The following is a short list of a broad range of coping strategies [92]:

Problem-focused coping

- **Active coping:** Taking active steps to remove or circumvent the stressor,
- **Planning:** Coming up w/ action strategies,
- **Seeking social support for instrumental reasons:** Seeking advice, assistance, or information.

Emotion-focused coping

- **Seeking social support for instrumental reasons:** Getting sympathy, moral support or understanding,
- **Acceptance:** Accepting the stressor and learning to live with it,
- **Restraint coping:** Waiting till the appropriate opportunity (holding back).

OCC, a Structural Appraisal Theory of Emotion

OCC (Ortony, Clore and Collins) model, similar to Lazarus' [141] and Scherer's [217] cognitive views, considers emotions to arise from affective or valenced reactions subsequent to the appraisal of a stimulus as being beneficial or harmful to one's concern [179]. The model categorizes emotions based on their underlying appraisal patterns. These patterns are fundamental criteria a person employs for evaluating a situation. They involve the person's focus of attention, her concern, and her appraisal preceding an affective reaction. Figure 2.4 shows main building blocks of OCC model.

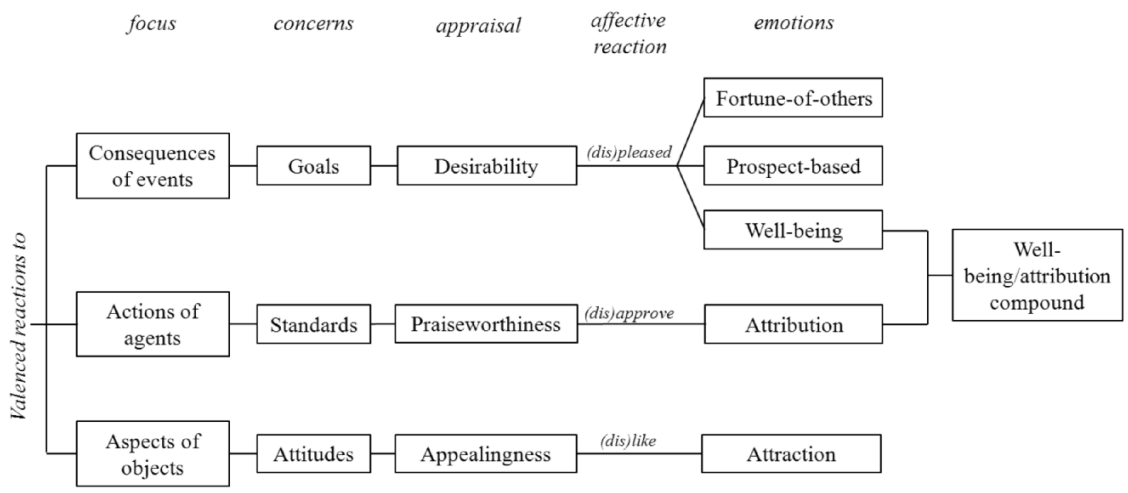

Figure 2.4: A simple visualization of OCC model [179].

As shown in Figure 2.4, a person could alternatively have three types of focuses. These types of focuses are consequence of events, actions of agents, and aspects of objects. The person evaluates the significance of causes behind these three types of focuses based on her personal concern. As a result, an affective reaction will be elicited resulting in an emotion. Various combinations of the elements depicted in Figure 2.4 create specific patterns demonstrating six main groups of emotions in which all emotion types in a group share the same cognitive pattern. Emotion groups are *fortune-of-others, prospect-based, well-being, attribution, well-being/attribution-compound*, and *attraction*. The OCC model introduces 22 emotion types. These emotions are introduced each as representative of a family of similar emotions with various intensities (since relying on a list of discrete emotions that is understood by everyone equally is impossible due to people's language barriers and various interpretations of the actual words). For instance, happiness can be referred to by other emotion terms such as joy, cheerfulness, gladness, delighted while they all share the same eliciting conditions. Thus the emotion types used in the model (e.g., relief, love, pride, and shame) are meant to represent an emotional experience rather than a lexical taxonomy.

For instance, as shown in Figure 2.4, the appraisal criterion for consequences

of events is their *desirability* for achieving one's goals. This generates the affective reaction of being *pleased* in positive cases, or *displeased* in negative ones. Figure 2.5 shows the resulting emotion groups in OCC model such as *fortune-of-others* (e.g., gloating, pity), *prospect- based* (e.g., satisfaction, relief), and *well-being* (e.g., joy, distress) [179]. The appraisal of the praiseworthiness of the actions of an agent against one's personal standards, as well as the appealing aspects of objects happens in the same way as shown in Figure 2.4.

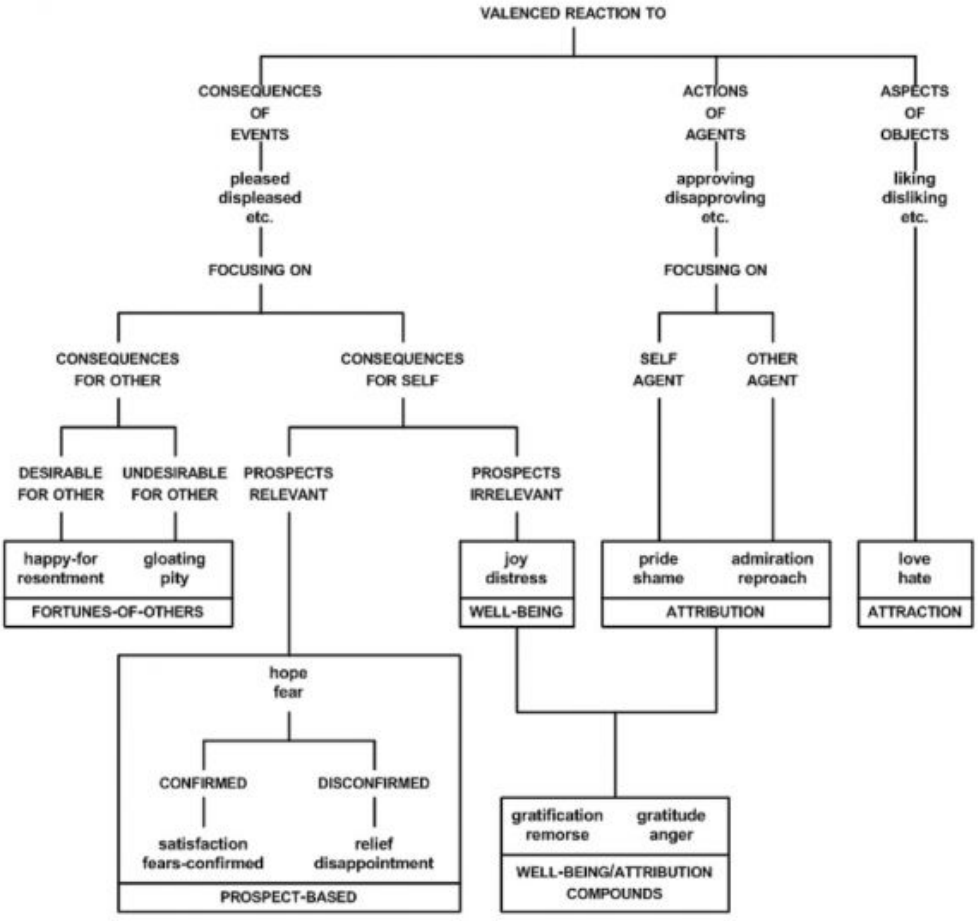

Figure 2.5: OCC taxonomy of emotion triggers and emotions [179].

Finally, the OCC model introduces some global variables of an emotion's in-

tensity to distinguish all types of emotions that a person could experience when encountering events, agents or objects. These variables are as follows

1. Sense of reality (representing the degree to which the event, agent or object in focus appear real to the person),

2. Proximity variable (representing the psychological proximity of an event, agent or object),

3. Unexpectedness (representing how surprising an event is for one, either positive or negative),

4. Arousal (representing how arousing an event, agent or object is).

2.4.2 Other Computational Models

Constructivist (Dimensional) Emotion Theories

The components and dimensions of emotions were the subject of much speculation since the 19th century. Dimensional models of emotion attempt to conceptualize human emotions by defining where they lie in two or three dimensions. Dimensional theories of emotion argue that emotion should be conceptualized, as points in a continuous (typically two or three) dimensional space rather than looking at them as discrete entities [43] [168] [210] [262].

Two dimensions that are commonly proposed to describe emotions are valence and physiological arousal [9] [141] [209]. Models based on dimensional theories contrast theories of basic emotion, which propose that different emotions arise from separate neural systems [193]. Many dimensional theories argue that discrete emotion categories (e.g., sadness, fear and anger) have no "reality" in that there are no specific brain regions or functions that correspond to specific emotions [18]. Dimensional theories do not emphasize the term emotion.

One of the most prominent two-dimensional models is Russell's circumplex model [209]. Russell suggested that affective states are all related to each other systemati-

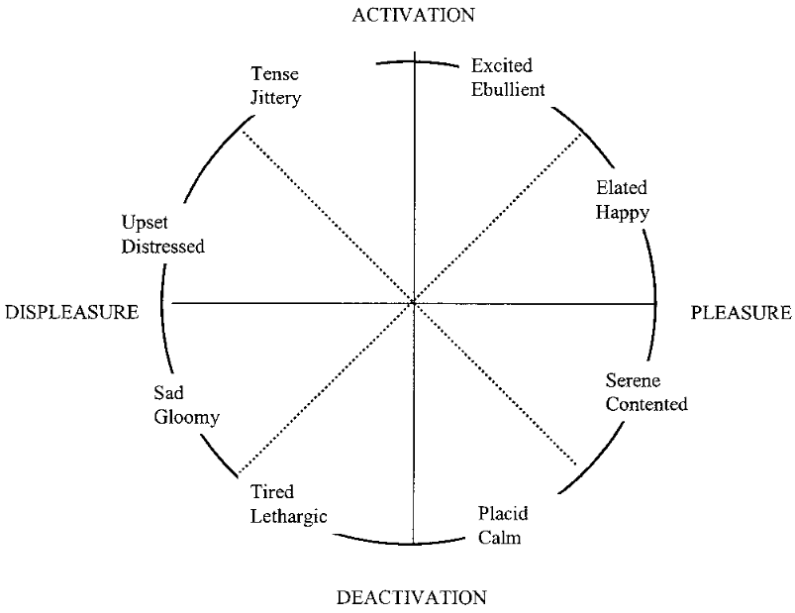

Figure 2.6: Russell's suggested affective states based on core affect [210].

cally through what is called core affect [209, 210] (see Figure 2.6) and emotions are best described as a change in core affect which, in turn, is describable as a point in a space between two bipolar dimensions. One dimension is *valence* or how good or bad objects and events are for a being ranging from pleasant to unpleasant. The other dimension is *arousal*, ranging from calm to excited. Russell put a number of affective states around a circular space between those two dimensions (see Figure 2.6) which is also known as *circumplex*, representing the variety of core affects [209, 210]. Since sometimes two-dimensional space cannot easily differentiate among emotions that share the same values of arousal and valence, e.g., anger and fear (both characterized by high arousal and negative valence), some of the dimensional models incorporate valence and arousal as well as *intensity*, or *dominance* or *stance* dimensions. Many computational dimensional models build on the three dimensional PAD model of Mehrabian and Russell [168] where these dimensions correspond to pleasure (a measure of valence), arousal (indicating the level of affective activation) and dominance (a measure of power or control). Figure 2.7 shows these three dimensions.

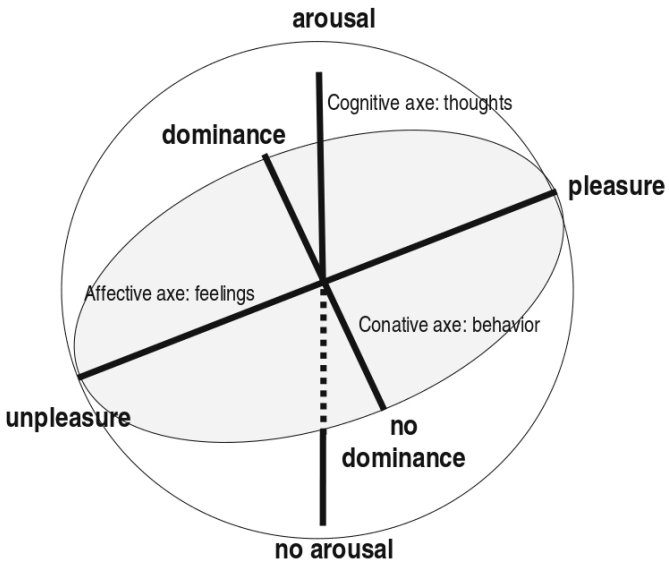

Figure 2.7: Three dimensional model of pleasure, arousal and dominance as tripartite view of experience [17].

Basic (Discrete) Emotion Theories

Basic emotion theories are inspired by Tomkins' [256] rediscovery of Darwin's work [62, 110] which were later developed by Ekman [70] and Izard [120]. These theories emphasize a small set of discrete and fundamental emotions. The underlying assumption of this approach is that these emotions are mediated by associated neural circuitry, with a hardwired component [70]. Different emotions are then characterized by stable patterns of triggers, behavioral expression, and associated distinct subjective experiences. The emotions addressed by these theories are typically called the *basic* emotions. Emotions including happiness, sadness, fear, anger, surprise, and disgust are often considered to comprise the most prototypical basic emotions [70]. The theory of basic emotions holds that there is a set of emotions shared by all humans that evolved to deal with ancestral life challenges [70]. For instance, disgust evolved to address the challenge of avoiding noxious stimuli, and fear evolved to address the challenge of avoiding dangers. Because of the emphasis on discrete categories of states, this approach is also termed the *categorical* approach [183]. Much of the supporting evidence offered for the theory comes from experiments that show

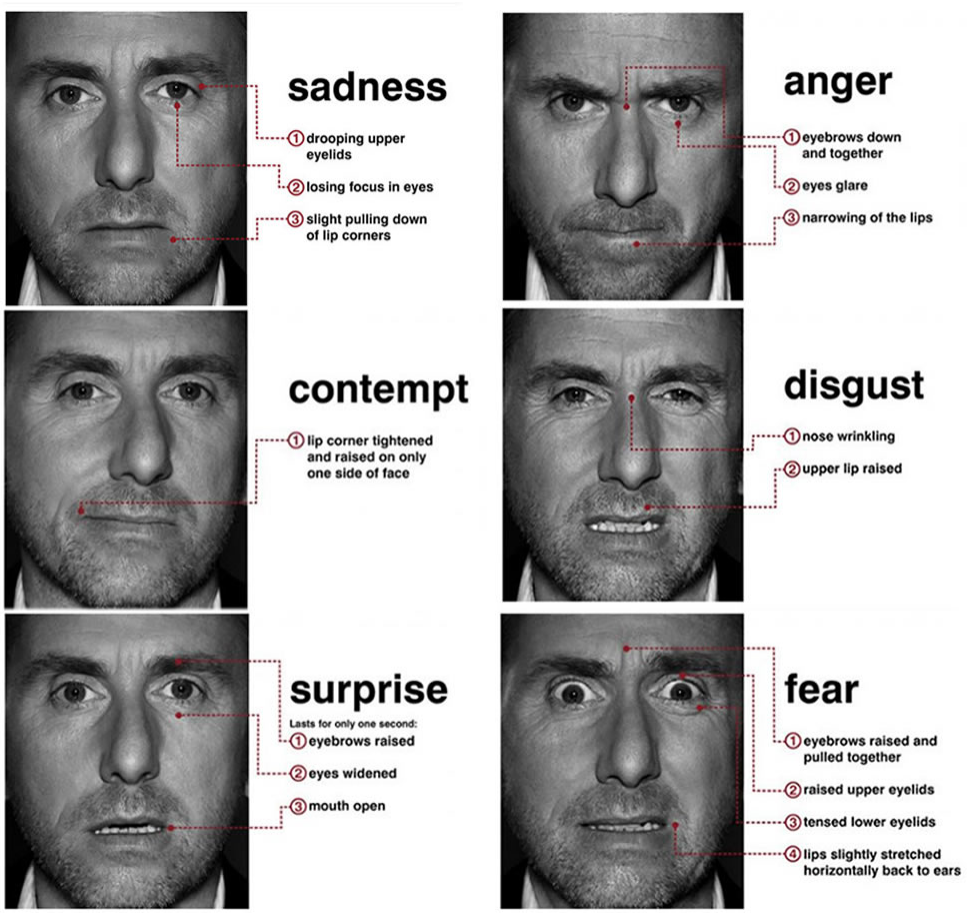

Figure 2.8: Basic emotions and corresponding expressions.

how certain facial expressions are universally associated with specific basic emotions, regardless of the observer's cultural background. This universality has a production side and a recognition side. On the production side, a particular emotional state is said to elicit a facial expression comprised of a fixed set of facial muscles. On the recognition side, observers are able to infer the emotional state of the person who expresses an emotion, due to the direct correspondence between emotional states and the facial expressions they cause. Computational models inspired by the basic emotions or discrete approach often focus on low-level perceptual-motor tasks and encode a two-process view of emotion that argues for a fast, automatic, undifferentiated emotional response and a slower, more differentiated response that relies on higher level reasoning processes (e.g., [8]).

There are other approaches that different researchers take based on their emphasis on the applicability of emotions in their systems.

Rational Approaches

Rational approaches start from the question of what adaptive functions emotions serve and then attempt to incorporate these functions into a model of intelligence. Emotion, within this approach, is simply another set of processes and constraints that have adaptive value. Models of this sort are most naturally directed towards the goal of improving theories of machine intelligence [4] [228] [239].

Communicative Approaches

Communicative theories of emotion argue that emotion processes function as a communicative system. They can function first, as a mechanism for informing other individuals of one's mental state (thereby facilitating social coordination), and second, as a mechanism for requesting/demanding changes in the behavior of others. Communicative theories emphasize the social-communicative function of expressions [89]. Computational models inspired by communicative theories focus on machinery that decides when an emotional expression can have a desirable effect on a human counterpart.

2.4.3 Similarities and Differences

Different theoretical perspectives should not be viewed as competing for a single truth. They should be seen as distinct perspectives, each arising from a particular research area (e.g., biological vs. social psychology), focusing on different sets of affective phenomena, considering distinct levels of resolution and fundamental components (e.g., emotions vs. appraisal variables as the distinct primitives). These different perspectives also provide different degrees of support for the distinct processes

of emotion, e.g., the componential theories provide extensive details about cognitive appraisals [118]. Therefore, this section provides a pairwise comparison between these fundamental theories. Note that a distinct pairwise comparison will not be provided for appraisal vs. discrete (basic) emotion theories as important points are adequately covered in the comparisons presented below.

Dimensional Vs. Discrete (Basic) Emotion Theories

The fundamental assumption of the basic emotion theory is that a specific type of event triggers a specific affect program corresponding to one of the basic emotions and producing characteristic expression patterns and physiological response configurations [222]. Dimensional theory's main criticism of basic emotions theory is based on the observation that affective phenomena appear to be both qualitatively and quantitatively diverse.

Russell in [210] argues the labels such as "fear", "anger", "happiness" do not capture this diversity. For instance, one might say: a) a person being chased by an assailant brandishing a knife, b) a person who retreats from an insect moving across the floor, and c) a person who is concerned they will never find a fulfilling career are all in a state of fear. On the basic emotions account, an emotional episode involves fixed patterns of neurophysiological and facial expression changes in response to an eliciting stimulus that are distinct between emotions, but are the same within the same emotional category [70]. If this were the case, one would expect that the three individuals described above would respond to their eliciting stimuli in the same way, yet a similarity of behavioral responses between these three cases seems unlikely. Dimensional theorists, in contrast, would argue that the individuals in the above three cases are applying the concept of fear to experience, despite the fact that each individual has a unique core affect. While basic emotion theorists would hold that since all three individuals are experiencing fear, they would perform the same behavioral responses to the stimuli, dimensional theorists would argue this is not the case, as each individual bears a core affective state that is distinguished from the

other two. For instance, the individual's arousal in response to an armed assailant should be higher than the individual in response to an insect, as the former case poses a threat to their life. As a result, the individual in the first case would likely make every effort to escape from the assailant, including trying to negotiate and plead with the assailant, while the individual in the second case would be relatively less dedicated to escaping the insect.

In sum, dimensional theory is compatible with the differences in the behavioral responses to eliciting stimuli, while basic emotions theory only allows for a single fixed behavior of responses to a given emotion. Furthermore, dimensional theories can represent instances of basic emotions (see Figure 2.9), for example, fear elicited by a snake (green rectangle), in terms of variation along affective dimensions, i.e., arousal and valence.

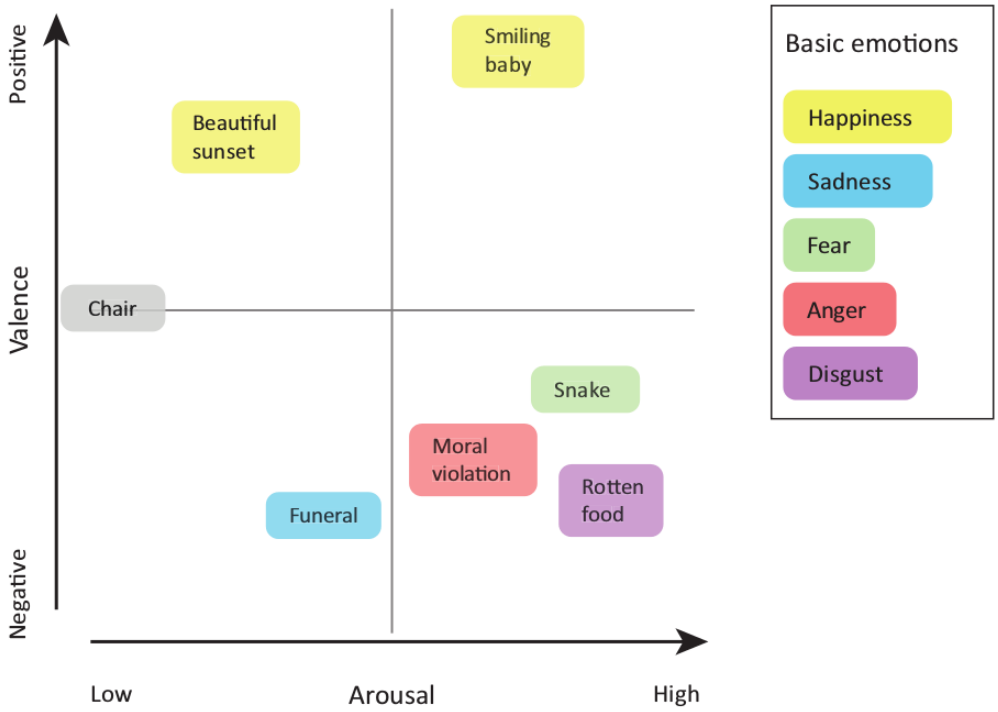

Figure 2.9: Representing basic emotions within a dimensional framework [105].

Also, basic emotion theory fails to account for affect that lacks object-directedness [210]. In the basic emotions approach, an emotion is supposed to have an intentional

object it is directed towards (e.g., being angry at someone, or being sad for someone). The dimensional theory argues that emotion may not necessarily be aimed at a particular object. For instance, an individual can experience a certain type of emotion (e.g., anger) without knowing of anything in particular that has offended her. Dimensional models of emotion are therefore capable of accounting for a wider range of affective phenomena than basic emotions theory.

Another difference between dimensional and basic emotion theories is that the basic emotion categorization of emotions captures facets of the experience of an emotion not conveyed by the dimensional description, such as elicitation of a facial expression of the emotion. In fact, this attribute of the basic emotions theory is one of the major differences with all other emotion theories. As it is argued in basic emotion theory, basic emotions are hard-wired to their corresponding facial expressions. Ekman who elaborated the concept of basic emotions, developed the *Facial Action Coding System* (FACS) which encodes movements of individual facial muscles and it is a common standard to systematically categorize the physical expression of emotions [71].

Appraisal Vs. Dimensional Emotions Theories

Dimensional theories might struggle to adequately distinguish emotions because of the existence of limited dimensions.

To compare the appraisal and dimensional theories of emotion, we can argue that there is a relationship between the dimensions in the constructivist or dimensional theory of emotion and appraisal dimensions. For instance, the pleasure dimension roughly maps onto appraisal dimensions that characterize the valence of an appraisal-eliciting event (e.g., intrinsic pleasantness –desirability–, or goal congruence), dominance roughly maps onto the appraisal dimension of coping potential, and arousal can be considered as a measure of intensity. However, they also have quite different meanings. Appraisal (as mentioned earlier) is a relational construct characterizing the relationship between some specific object/event in the environ-

ment and the individual's mental constructs including beliefs, motives and intentions and several appraisals may be simultaneously active; whereas emotions in dimensional emotion theory are non-relational constructs, each summarizing a unique overall state of the individual.

Furthermore, dimensional emotion theories emphasize different components of emotion than appraisal theories and link these components quite differently. In contrast to appraisal theories, dimensional emotion theories do not address affects antecedents in detail. However, dimensional theorists question the tight causal linkage between appraisal and emotion that is central to appraisal accounts. As mentioned earlier, dimensional theorists believe that the emotion is not necessarily about some object (as in "I am angry at him"). In such theories, many factors may contribute to a change in emotion including intentional judgments (e.g., appraisal). However, in dimensional emotion theories the link between any preceding intentional meaning and emotion is broken and most of the time can not be recovered correctly. For example, Russell argues for the following sequence of emotional components: some external event occurs (e.g., a bear walks out of the forest), it is perceived in terms of its affective quality; this perception results in a crucial change in core affect; this change is attributed to some "object" (e.g., the bear); and only then is the object cognitively appraised in terms of its goal relevance, causal antecedents and future prospects [161].

We can also compare the dimensional emotion theories to OCC model as a cognitive appraisal model. The major similarity between these two models is that they both consider emotions to descend from valenced reactions to the stimuli. Furthermore, they acknowledge the role of arousal in determining emotional reactions. As we mentioned in Section 2.4.2 Russell considered arousal as one of the two key dimensions of emotions which could be used to partially discriminate emotional states [209]. In a different manner, the OCC model recognizes arousal as a necessary condition for eliciting emotions, and regards the arousal as a major determinant of the elicited emotion's intensity which distinguishes among various emotions of a par-

ticular type (e.g., fearful and scared). In [220] Scherer speculates that the arousal dimension in dimensional models gives little information about the underlying appraisal of the elicited emotion and he proposes to replace it with coping potential which is an appraisal dimension referring to the individual's perceived control in a given situation.

Furthermore, models based on dimensional emotions theory pursue the idea of eliciting an emotion according to the joint features in circumplex space (2D or 3D – see Section 2.4.2) while OCC or other models of appraisal theory are based on patterns of antecedents of emotions. This is the fundamental difference between OCC, or appraisal theories in general, and the circumplex approach of Russell [209] or Mehrabian's PAD model [17, 168]. Also, models based on appraisal theory of emotions employ causation, attribution and eliciting conditions in order to distinguish emotions while the eliciting conditions are not directly accessible from a dimensional approach. A dimensional model might fall short in establishing why certain emotions are elicited. However, when the objective is to identify the generated emotions and their level of pleasantness and intensity, a circumplex model presents an excellent opportunity [3].

Finally, here, we discuss how a model based on dimensional emotions theory (i.e., Russell's 2D circumplex) relates to a cognitive model based on appraisal theory (i.e., OCC model). Figure 2.10 shows the relationship between Russell's circumplex and OCC model in terms of categorization of the actual emotions. The number of emotions in a section of Russell's circumplex that fall into an emotion group of OCC are shown in parentheses (see Figure 2.10). For instance, all three emotions in the top section (highly excited, neutrally valenced emotions) fall into prospect based emotion group, hence number (3) is indicated. Or, as another example, emotions in the left section (neutral arousal value, negative valenced emotions) make a one to one relationship between disappointment and the prospect based emotion group, contempt and attribution emotion group, and jealousy and fortune of others emotion group, hence number (1) is indicated in front of each.

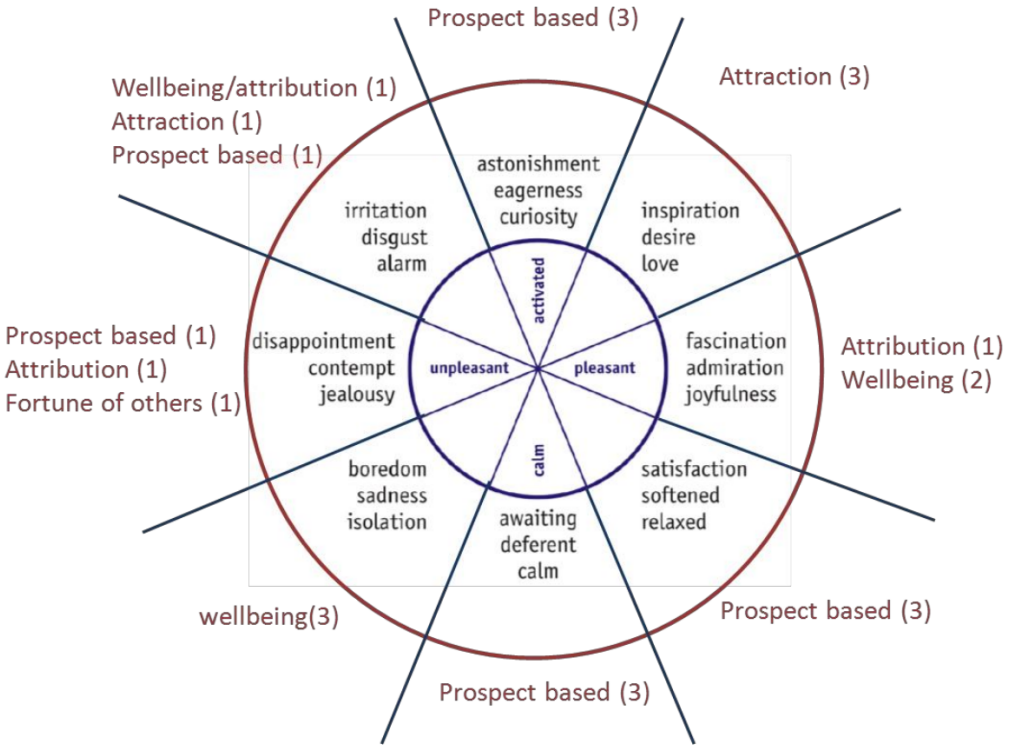

Figure 2.10: A rough projection of emotion groups of OCC on the circumplex of affect [3].

2.4.4 Applications in Autonomous Agents and Robots

There are many research areas, including robotics and autonomous agents, that employ the structure and/or functions of emotions in their work with a variety of motivations behind modeling emotions [263]. Some of these works are inspired by specific psychological theories, some are freely using the concept of emotion without using the theoretical background in social sciences, and some are using a combination of concepts from the psychological theories. For instance, in PECS [258] which is designed for modeling human behaviors, the agent's architecture is not based on a certain kind of social or psychological emotion theory. In fact, it is intentionally designed and described in a way which enables the integration of a variety of theories. The PECS' design enables an integrative modeling of physical, emotional, cognitive and social influences within a component-oriented agent architecture. Also, in

[165] the computational architecture which is designed to provide information about the possible overall behavior of a work team is not based any specific theory. As mentioned earlier, some researchers apply combinations of emotion theories in their work [133]. For instance, in [41] Cañamero shows how an agent can use emotions for activity selection while taking into account both dimensional and discrete approaches in an action selection mechanism. Throughout this section, we provide different examples of works using major emotion theories in robots and autonomous agents.

We can also see the application of emotion theories in designing companion robots, robots capable of expressing emotions and social behaviors, as well as robots which can convey certain types of emotion products, e.g., empathy [33] [142] [182] [233]. Robots also use emotions theories for automatic affect recognition using different modalities [109] [271]. Moreover, in some works, researchers have explored the user's affective state as a mechanism to adapt the robot's behaviors during the interaction [32] [151].

Applications of Appraisal Theory – The emphasis of models derived from appraisal theories of emotion is on making appraisal the central process. Computational appraisal models often exploit elaborate mechanisms for deriving appraisal variables such as decision-theoretic plans [92] [162], reactive plans [197] [202] [246], Markov-decision processes [72] [235], or detailed cognitive models [158]. However, emotion itself is sometimes treated less elaborately, and simply as a label to which behavior can be attached [74]. Appraisal is usually modeled as the cause of emotion being derived via simple rules on a set of appraisal variables.

Computational appraisal models have been applied to a variety of uses including contributions to psychology, robotics, AI, and HCI. For instance, Marsella and Gratch have used EMA [162] to generate specific predictions about how human subjects will appraise and cope with emotional situations and argue that empirical tests of these predictions have implications for psychological appraisal theory [91]

[160]. There are several examples in artificial intelligence and robotics of applying appraisal theory [2] [130] [162]. In robotics, appraisal theory has been used to establish and maintain a better interaction between a robot and a human. For instance in [130] researchers provide their computational model of emotion generation based on appraisal theory to have a positive human-robot interaction experience. In [213] authors describe a system approach to appraisal processes based on Scherer's work on appraisal and the Component Process Model [217]. They show how the temporal unfolding of emotions can be experimentally tested. They also lay out a general domain-independent computational model of appraisal and coping. In [261] researchers consider their robot's (INDIGO) emotion, speech and facial expressions as a key point to establish effective communication between the robot and a human during their interaction. They apply concepts of appraisal theory in INDIGO's emotion modeling. MAGGIE, a sociable robot, also applies the appraisal theory of emotions to consider fear in its decision making system [86]. Velasquez developed Cathexis which is a distributed computational model for generation of emotions and their influence in the behavior of the autonomous agents [260]. The emotion model in this work is based on Roseman's work on appraisal theory. Marinier and Laird in [157] focus on the functional benefits of emotion in a cognitive system. In this work, they integrate their emotion theory (which is based on appraisal theory) with Soar cognitive architecture, and use emotional feedback to drive reinforcement learning. In [116] Hudlicka provides a model of a generic mechanism mediating the affective influences on cognition based on cognitive appraisal. This model is implemented within a domain-independent cognitive-affective architecture (MAMAID).

In the virtual agents community, empathy is a research topic that has received much attention in the last decade [30] [167] [180] [194] [251]. In [192] researchers developed an agent with capability of affective decision-making based on appraisal theory to establish an affective relationship with its users. Then, they compared the performance of their agent with a human (based on a WoZ study) in a speed-dating experiment. In HCI, appraisal theory has been primarily used for the creation

of interactive characters that exhibit emotions in order to make characters more believable [201], more realistic [155] [257], more capable of understanding human motivational states [58] or more able to induce desirable social effects in human users [181].

Applications of Dimensional Theory – The emphasis of models influenced by dimensional theories is on processes associated with core affect which is usually represented as a continuous time-varying process, and it can be determined at a given time by a point in a 2D or 3D-space as a response to the eliciting events. Generally, there are detailed mechanisms in computational dimensional models which determine how this point changes over time, e.g., decay to some resting state, and incorporating the impact of dispositional tendencies such as personality or temperament [83] [161]. Models based on dimensional theories have also been used in robotics. For instance, researchers in [147] apply PAD's three-dimensional space to rate the pleasure, arousal and dominance of their Multimodal Emotional Intelligence robot (MEI) in each interaction with human subjects. Their goal is to introduce the first steps in MEI which can understand and express emotions in voice, gesture and gait. In [272] researchers want to understand the effect of different interface features for a service robot. They use valence and arousal dimensions in their questionnaires to assess the perceived anthropomorphism of their own service robot by their subjects. In [136] researchers introduce the implementation of a dynamic personality for a robot based on a dimensional emotion model. They use WASABI's architecture [22, 23] as their emotional model. In [149] the author describes an affective knowledge representation scheme to be used in the design of a socially intelligent artificial agent. Lisetti uses the valence-arousal two dimensional model of emotion in this work. This model has been applied in an emotion-based architecture of Lisetti's autonomous robots as well as a multimodal affective user interface agent. ROMAN, an expressive robotic head, uses a behavior-based emotional control architecture. The approach to the emotional component of the architecture is based on the dimensional emotion theory [113].

Comparison of Applications of Emotion Theories – Researchers often use computational dimensional models for behavior generation of animated characters. The reason might be because it is easier for emotion translation to a limited number of dimensions that can be readily mapped to continuous features of behavior such as the spatial extent of a facial expression. For example, PAD models describe all behavior in terms of only three dimensions of pleasure, arousal and dominance, whereas researchers using appraisal models should either associate each behavior with a large number of appraisal variables [223] [244], or try to map appraisal variables into a limited and small number of discrete expressions [74]. For a similar reason, dimensional models also frequently used as a good representational framework for systems that attempt to recognize human emotional behavior and there is some evidence that they may better discriminate user affective states than approaches that rely on discrete labels [18].

There is also a relationship between dimensional and appraisal theories. Some of the computational models of emotion that incorporate dimensional theories have viewed appraisal as the mechanism that initiates changes to core affect. For instance, ALMA [83] includes OCC inspired appraisal rules [179], and WASABI [22] includes appraisal processes inspired by Scherer's sequential-checking theory into a PAD-based emotion model. Moreover, some computational models explore how core affect in dimensional models can influence cognitive processes. For example, HOTCO 2 [252] allows explanations to be biased by dimensional affect [161].

2.5 Affect and Motives

Motives are essential mental components in decision-making procedures and applying them in an affect-driven collaborative agent is part of this thesis' contribution. In this section, we provide related works on computational models of motivation and discuss the nature of motives. We also explain three of the important social motives which will be used in our work. Finally, we discuss that humans' beliefs,

emotions and motives are related and influence each other.

Motives' principles and mechanisms, as the reasons behind one's intentions and actions, and the influences of motives on cognition have been discussed in philosophy, neuroscience, psychology and artificial intelligence [15, 24, 37, 239, 241]. There are several examples in AI providing computational models for different psychological theories of motivation. Bach's MicroPsi agent architecture describes the interaction of emotion, motivation and cognition of agents based on Dietrich D*ö*rner's Psi theory [13, 14, 15, 16]. Merrick and Shafi provide a computational model for motivation based on Henry Murray's theory [174] describing the three important social motivations of *achievement, affiliation* and *power*. They focus on the role of motivation in a goal-selection mechanism [169]. There are other examples focusing on the impact of motives on different cognitive processes in robots and artificial agents [31, 46, 65, 232, 260, 266]. The motivation mechanism in our work is inspired by Murray's theory and Bach's approach on D*ö*rner's theory. It is focused on the role of motives in cognitive processes, e.g., intention formation in coping, during collaboration, which will be discussed in Chapters 3 and 4.

2.5.1 Motives

A motive consists of an urge (that is, a need indicating a demand) and a goal that is related to this urge [14]. Motives shape cognition and behavior [230]. To be motivated means to be moved to do something [211]. Motives direct behaviors towards particular goals, which makes the agent more persistent in actions it takes. They also affect cognitive processes by increasing level of attention. Motive, as the outcome of the motivation process, initiates, directs and maintains goal-oriented behaviors.

Motives are goal-driven and they move the agent towards the attainment of corresponding sets of intentions. In other words, motives as an essential part of affect can lead the agent to empower an intention. They are essentially mechanisms that in light of beliefs tend to produce, modify or select between actions and their reciprocal

intentions. Some of the motives are transient, like helping the Astronaut to hold the panel, while some are long term, like reaching to the shared goal during collaboration which in our example is installing solar panels and satisfying the Astronaut's needs in the field (see Section 3.1).

2.5.2 Motivation Theory

There are several motivation theories in psychology [21, 87, 138], some of which have received little attention as the basis for computational models. In [174], Murray described and studied 20 different human motives, of which three have received attention in psychology and artificial intelligence as social motives [169, 274]. The following is a brief description of these three social motives, *achievement, affiliation* and *power* [11, 274] which will be used in this thesis:

- **Achievement motivation:** Achievement motivation drives humans to strive for excellence by improving on personal and societal standards of performance. It involves a concern for excellence, for doing one's best. In artificial agents, achievement motivation has potential roles in focusing agent behavior and driving the acquisition of competence.

- **Affiliation motivation:** Affiliation refers to a class of social interactions that seek contact with formerly unknown or little known individuals and maintain contact with those individuals in a manner that both parties experience as satisfying, stimulating, and enriching. It involves a concern with developing friendly connections with others through the two contrasting emotional components of hope of affiliation and fear of rejection. These two components become more crucial in the collaboration domain due to the importance of social emotions and their impact on beliefs and intentions.

- **Power motivation:** Power can be described as a domain specific relationship between two individuals, characterized by the asymmetric distribution of

social competence, access to resources, or social status. It involves concern with having an impact on other people or on the world at large. There are different aspects of fear or avoidance of power which channel and moderate the expression of power into socially acceptable behavior, working as inhibitions to unseemly tendencies. Power motivation can be considered with respect to the probability of success which makes it relevant to the cognitive appraisal of emotions during collaboration.

In [274] it is shown that success of a power goal is associated with anger, confusion and disgust; success at an affiliation goal is associated with interest, happiness and feeling loved; and success at an achievement goal is associated with interest, surprise, happiness, excitement and a sense of focus. In other words, succeeding at a particular motive is associated with experiencing particular emotions.

2.6 Theory of Mind

Theory of mind, as a crucial component in human's social interaction, plays an important role in our computational model. It discusses one's beliefs about others as intentional agents. Beside the immediate effect, an individual's action also depends on the beliefs about other's perception of that action as well as the reaction they take. In this thesis, we use this concept whenever the agent needs to anticipate the human's mental states. We will also use the concept of *user model* as a standard collection of properties to describe others.

The concept of theory of mind has received attention in social psychology and artificial intelligence. Eligio et al. explore what collaborators understand about each other's emotions and conclude being aware of each other's emotions helps collaborators to improve their performance [73]. Fussell and Kraus discuss the importance of perspective taking in a successful communication in a social setting [80]. Scassellati discusses the importance of attribution of beliefs, goals and desires to others. He presents two psychological theories on the development of theory of mind in humans

and their potential application in building robots with similar capabilities [214]. Hiatt and Trafton present a cognitive model which borrows mechanisms from three different postulates of theory of mind and show that their model produces behaviors in accordance with various theories of experiences [111]. Si, Marsella and Pynadath discuss PsychSim, an implemented multi-agent-based simulation tool for modeling social interaction, which has its own beliefs about its environment and a recursive model of other agents [195]. They also investigate the computational modeling of appraisal in a multi-agent decision-theoretic framework using POMDP based agents [236, 235]. Since applying the concept of theory of mind is crucial in social interaction and collaboration, this thesis employs a simplified mechanism inspired by the existing works for our agent.

2.7 Conclusion

In this chapter, we started by defining the concept of collaboration based on Grosz and Sidner's work [103], and listed a number of collaboration properties. Then, we provided the background of two prominent collaboration theories which helped develop a better understanding of the actual theories and how they relate to each other. Next, we presented the SharedPlans theory and its major properties, e.g., partial shared plan, recipe, and two notions of intention. Afterwards, we delivered key concepts of the Joint Intentions theory including joint commitment and joint intention. Then, we continued with the hybrid approach of modeling collaboration and provided one of the most well-established models, STEAM. We also briefly mentioned some other approaches. Later, we presented two different lists to compare similarities and differences between SharedPlans and Joint Intentions collaboration theories. We ended this document with different categories of applications of these theories in agent/robot and human collaboration areas.

We believe the SharedPlans and Joint Intentions collaboration theories are the most well-defined and well-established theories in computer science. We found

SharedPlans theory more convincing than the other major and subordinate approaches, with respect to its inclusive explanation of the collaboration structure and its association to discourse analysis which directly improves the communicative aspects of a collaboration theory. We also understand the value of Joint Intentions theory due to its clarity and closeness to the foundations of collaboration concepts. These specifications of the Joint Intentions theory can make it applicable in multi-agent system designs and human-robot collaboration. We also consider hybrid approaches valuable, such as STEAM, if they clearly understand drawbacks with existing theories and successfully achieve better collaborative agents by infusing different concepts from different theories. Although all these theories are well-defined and properly introduce collaboration concepts, they mostly explain the structure of a collaboration and they lack the underlying domain-independent processes with which collaborative procedures could be defined more systematically and effectively in different applications.

Later, we looked at the description of affective computing and the importance of the concept of emotion in general and in social context. We also discussed the importance of communicating emotions as well as emotions' social functions. Then, we provided some examples of agents and robots using artificial emotions in their decision making process. We also briefly provided a few examples of cognitive architectures producing different aspects of behaviors in robotics.

There are major theories of emotions explaining the concept of emotion. We discussed these major theories in detail separately, providing their psychological background and underlying concepts. Following the explanation of these theories, we were able to discuss the similarities and differences between these major theories. Finally, we provided applications of these theories in robotics and AI.

We believe to develop or work based on computational models of emotions, it is good to follow well-established (in comparison with others) theoretical foundations. These theories can be a guideline for our computational models, and they can explain more details of the structure or the processes involved in affective situations.

However, we do not necessarily think that the computational models must exactly follow only one theory and its descriptions. Meaning, different aspects of models can represent different theories. For instance, appraisal theory is a good representation of the interpretive aspect of emotions and basic emotion theories provide detailed systematic methods for expressive application. More importantly, we believe the interpersonal functions of emotions should be our first concern and we should try to relate them to the structure of our domain, i.e., collaboration. In conclusion, we can see the importance of interpretive, communicative and regulatory aspects of emotion functions in this proposed work.