Emotional Reasoning in an Action Language for Emotion-Aware Planning

Andreas Brännström[0000−0001−9379−4281] and Juan Carlos Nieves[0000−0003−4072−8795]

Umeå University, Department of Computing Science, SE-901 87, Umeå, Sweden
{andreasb, jcnieves}@cs.umu.se

Abstract. This paper introduces formal models for emotional reasoning, expressing emotional states and emotional causality, using action reasoning and transition systems. A general framework is defined, comprised of two main components: 1) a model for emotions based on the Appraisal theory of Emotion (AE), and 2) a model for emotional change based on Hedonic Emotion Regulation (HER). A particular transition system is modelled in which states correspond to human emotional states and transitions correspond to restrictive (safe) ways to influence emotions while reducing negative emotional side-effects. The introduced emotional reasoning can be applied to guide a software agent's actions for dealing with emotions while estimating and planning future interactions with humans.

Keywords: Emotional reasoning · Human-aware planning · Action languages · Appraisal theory.

1 Introduction

An aim in the area of Human-Agent Interaction (HAI) is to develop interactive cognitive systems that are *human-aware*, providing a proactive and personalized interaction. Human-Aware Planning (HAP) [11] regards a scenario where an intelligent system is situated in an environment populated by humans, in which the system must plan its actions by meeting the requirements of human plans and goals. In order for a software agent to execute suitable actions in interactions with humans, the agent must consider the mental states of its human interlocutors in its internal reasoning and decision-making. This is an ability, referred to as Theory of Mind (ToM) [12], to infer another agent's beliefs, such as emotions, motivations, goals and intentions. We need to develop dynamic ways for systems to compute a ToM of their users, making systems aware of human mental properties and their causes. "Emotions" play a fundamental role in human behavior and interactions [2]. By providing a system mathematical computational models for emotional reasoning [9] when planning its actions, interaction capabilities of an agent can be greatly improved. The software agent must be aware of what emotions that are present in the mind of the human and what emotions that can be triggered, in each state of the interaction.

Challenges when building *emotion-aware* [4] interactive systems include to provide capabilities of: 1) backward reasoning, e.g., recognizing and reasoning about causes of emotions, 2) context reasoning, e.g., evaluating timely and appropriate emotional states, 3) forward reasoning, e.g., predicting the effects of their actions on emotions of humans, and 4) action reasoning for adapting their behaviors accordingly, promoting appropriate emotions while avoiding unintended emotional side-effects. To achieve such capabilities, intelligent systems require models that capture explanations to why and how emotions arise and change. Previous approaches to computational emotional reasoning [1, 9, 12, 14] mainly focus on recognizing emotional context, e.g., by simulating emotional behavior [12] or to model expected human behavior in response to emotions [9], and do not capture explanations for emotional change as state transitions. In order to predict the effects of an agent's actions on emotions of humans, the agent needs a way to reason backward and forward using models that specify how human emotions are caused and change, in terms of states and transitions. Given the challenges of emotional reasoning in the setting of HAI, the following research question arises: — How to track emotional states of human agents in a goal-oriented interaction between humans and software agents?

We introduce a methodology to model emotional state transitions by formalizing two emotion theories, the Appraisal theory of Emotion (AE) [6] and Hedonic Emotion Regulation (HER) [17], capturing links between human emotions and their underlying beliefs, using transition systems and action reasoning [7]. To this end, a set of action specifications is introduced, $\mathcal{C}_{AE}$, that captures transitions between human emotions. The proposed emotional reasoning framework regards two main components: 1) a model for emotion representation, following the psychological theory of AE, through which a set of 16 basic human emotions is explained, and 2) a model for emotional change, following the theory of HER, aiming to increase positive emotion and decrease negative emotion.

This paper is organized as follows. In Section 2, the state-of-the-art in emotional reasoning is presented. In Section 3, the theoretical (computational and psychological) background is presented. In Section 4, syntax and semantics of the proposed emotional framework reasoning are presented. Finally, in Section 5-6, the paper is concluded by discussing potential applications, limitations, and directions for future work.

2 Related Work

There is a diverse body of research related to the ideas presented in the present work. In the area of affective agents and computational theory of mind [1, 12], agent models have been developed to reason about emotion and behavior. For instance, agents based on Partially Observable Markov Decision Processes (POMDP) [12] have been used to (similar to the present study) model appraisal and emotion. Their models show potential in simulating human emotional behavior. They have, however, lacked to capture human emotional change to deliberate about emotion regulation, future interactions and emotional effects of actions.

A variety of Emotion BDI (Belief, Desire, Intention) frameworks [9, 14, 15] have been introduced. These approaches have aimed, e.g., to model behaviors which are expected from agents under the influence of emotions [14], or to provide modular generic interfaces for emotional agents [9] to enable emotion theory-based models as filters for emotional reasoning. While these works define generic architectures for emotional agents, they need to be coupled with emotion theory-based models to enable reasoning about emotional change, and human-aware reasoning to avoid unintended emotional side-effects in their interactions.

3 Theoretical Background

This section presents the emotion theories of AE and HER, the theoretical base of the proposed emotional reasoning framework. The section then presents action reasoning languages and transition systems, serving as a platform on which emotional reasoning is formalized and characterized.

3.1 Emotion Theories: AE and HER

AE [6] proposes that emotions are caused by an appraisal of a situation in terms of 1) being consistent or inconsistent with needs, 2) being consistent or inconsistent with goals, 3) the accountability of a situation, which can be the environment, others, or oneself, and 4) as being easy or difficult to control. According to AE, the difference between goal consistency and need consistency determines negative, stable and positive emotions. More intense negative emotions (e.g., Anger or Fear) arise when the need consistency is greater than the goal consistency, while less intense negative emotions can arise when both the need consistency and goal consistency are low. On the other hand, positive emotions (e.g., Joy or Liking) arise when the goal consistency is greater than the need consistency, or when both are high. By ranking consistency values as Low < Undecided < High and by looking at the difference between need and goal consistency, positive and negative emotions can be distinguished.

HER [17] is a theory for regulating emotions, guided by the goals to 1) increase positive emotion and 2) decrease negative emotion. According to HER, both of these emotion regulation goals are associated with improved well-being, where decreasing of negative emotion has been most effective [13]. The principles of HER can be applied in the framework of AE to reason about emotional change.

3.2 Action Reasoning and Transition Systems

A transition system is a directed graph, whose nodes correspond to states (configurations of variables) and edges correspond to valid transitions between states. A transition system has an initial state (the current observation) and a set of goal states (which it aims to reach). Action reasoning [7] regards logical descriptions of actions that result in transitions between states. As a platform for our emotional reasoning specification, we build on the action language $\mathcal{C}_{TAID}$ [5].

The alphabet of $\mathcal{C}_{TAID}$ consists of two nonempty disjoint sets of symbols $\mathbf{F}$ and $\mathbf{A}$. They are called the set of fluents $\mathbf{F}$ and the set of actions $\mathbf{A}$. A *fluent* expresses a property of an object in a world, and forms part of the description of states of the world. A *fluent literal* is a fluent or a fluent preceded by $\neg$. A *state* σ is a collection of fluents. A fluent f holds in a state σ if $f \in \sigma$. A fluent literal $\neg f$ holds in σ if $f \notin \sigma$.

4 Emotional Reasoning

The contribution of the paper starts in this section, which presents an emotional reasoning specification, $\mathcal{C}_{AE}$. Components of AE are formalized as a particular transition system, called an emotion decision-graph (EDG), to reason about emotional states and (safe) emotional change to reduce unintended emotional side-effects. The EDG specifies transitions between emotional states (in terms of HER), which serve as safety restrictions for emotion-influencing actions.

Recall that AE defines emotions as a composition of an individual's appraisal of a situation, in terms of consistency with needs, consistency with goals, accountability and control potential. By following this definition of emotional causes, we specify states with emotion fluents and values of the following form:

- need_consistency(ne), $ne \in$ {low=l, high=h, undecided=u},
- goal_consistency(go), $go \in$ {low=l, high=h, undecided=u},
- accountability(ac), $ac \in$ {environment=e, others=o, self=s, undecided=u},
- control_potential(co), $co \in$ {low=l, high=h, undecided=u}

By defining a set of emotions following AE in this way, and by utilizing principles of hedonic emotion regulation, we can specify preferable (safe) transitions between emotional states. In the following subsection, we specify an EDG to reason about emotional transitions.

4.1 Emotion decision-graph (EDG)

Following AE, 16 emotional states are specified, one for each basic emotion explained by AE theory, i.e., {Anger, Dislike, Disgust, Sadness, Hope, Frustration, Fear, Distress, Joy, Liking, Pride, Surprise, Relief, Regret, Shame, Guilt}. We can model these states and transitions as a graph, an EDG, that represents a prioritized focus of emotional change given a recognized emotional state (see Fig. 1).

Definition 1. *An emotion decision-graph EDG is a transition system that is a tuple of the form $EDG = (E, Act, T, O)$ where E is a non-empty set of states such that each state contains emotion fluents in terms of AE, Act is a set of actions, $T \subseteq E \times E$ is a non-empty set of transition relations between emotional states, O is a set of initial observations.*

The emotion decision-graph is formalized by the semantics of the action language specification $\mathcal{C}_{AE}$, serving as restrictions for safe emotional change, presented in the following section.

Anger	**Hope**	**Joy**	**Relief**
ne:h go:l ac:o co:h	ne:u go:h ac:e co:l	ne:h go:h ac:e co:u	ne:h go:h ac:e co:u
Dislike	**Frustration**	**Liking**	**Regret**
ne:u go:l ac:o co:l	ne:h go:l ac:e co:h	ne:u go:h ac:o co:u	ne:u go:l ac:s co:l
Disgust	**Fear**	**Pride**	**Shame**
ne:l go:l ac:e co:h	ne:u go:l ac:e co:l	ne:u go:h ac:s co:u	ne:l go:l ac:s co:h
Sadness	**Distress**	**Surprise**	**Guilt**
ne:h go:l ac:e co:l	ne:l go:l ac:e co:l	ne:u go:u ac:e co:u	ne:h go:h ac:s co:h

Fig. 1. Emotional states following Appraisal theory of Emotion [6].

4.2 Action language specifications

$\mathcal{C}_{AE}$ is comprised of sets of symbols to represent emotional appraisals, which define an emotion-aware alphabet as follows:

Definition 2 (Emotion-aware alphabet). *Let* $\mathbf{A}$ *be a non-empty set of actions and* $\mathbf{F}$ *be a non-empty set of fluents.*

- $\mathbf{F} = \mathbf{F}^E \cup \mathbf{F}^H$ *such that* $\mathbf{F}^E$ *is a non-empty set of fluent literals describing observable items in an environment and* $\mathbf{F}^H$ *is a non-empty set of fluent literals describing the emotional-states of humans.* $\mathbf{F}^E$ *and* $\mathbf{F}^H$ *are pairwise disjoint.*
- $\mathbf{F}^H = \mathbf{F}^N \cup\ \mathbf{F}^G \cup\ \mathbf{F}^A \cup\ \mathbf{F}^C$ *such that* $\mathbf{F}^N, \mathbf{F}^G,\ \mathbf{F}^A$*and* $\mathbf{F}^C$ *are non-empty pairwise disjoint sets of fluent literals describing a human agent's need consistency, goal consistency, accountability and control potential, respectively.*
- $\mathbf{A} = \mathbf{A}^E \cup \mathbf{A}^H$ *such that* $\mathbf{A}^E$ *is a non-empty set of actions that can be performed by a software agent and* $\mathbf{A}^H$ *is non-empty set of actions that can be performed by a human agent.* $\mathbf{A}^E$ *and* $\mathbf{A}^H$ *are pairwise disjoint.*

Definition 3 (Emotion fluent). *An emotion fluent is a predicate f(X,Y,Z) of arity 3 such that* $X \in$ *{ne,go,ac,co},* $Y \in$ *{l,h,u,e,o,s} and* $Z \in \mathbb{N} \cup \{0\}$*. An emotion fluent f(X,Y,Z) is well-formed if the following conditions hold true:*

1. *if* $X \in$ *ne,go,co, then* $Y \in$ *{l,h,u}*
2. *if* $X = ac$*, then* $Y \in$ *{e,o,s,u}*

where ne represents need consistency, go represents goal consistency, ac represents accountability and co represents control potential; l represents low, h represents high, u represents undecided, e represents environment, o represents other and s represents self; and Z represents a point in time.

$\mathcal{C}_{AE}$ defines a set of static and dynamic causal laws of actions. These laws specify emotional influences, either as effects of actions or as indirect causal effects. Laws for emotional change work by influencing appraisal of a situation in the human agent while complying with the constraints of the EDG.

Definition 4 (Emotion-aware domain description language). *An emotion-aware domain description language $D^{ae}(\mathbf{A}, \mathbf{F})$ consists of static and dynamic causal laws of the following form:*

$\mathcal{C}_{TAID}$ *domain description language:*
$(a\ \textbf{causes}\ f_1, \ldots, f_n\ \textbf{if}\ g_1, \ldots g_m)$ (1)
$(f_1, \ldots, f_n\ \textbf{if}\ g_1, \ldots g_m)$ (2)
$(f_1, \ldots, f_n\ \textbf{triggers}\ a)$ (3)
$(f_1, \ldots, f_n\ \textbf{allows}\ a)$ (4)
$(f_1, \ldots, f_n\ \textbf{inhibits}\ a)$ (5)
$(\textbf{noconcurrency}\ a_1, \ldots, a_n)$ (6)
$(\textbf{default}\ g)$ (7)

$\mathcal{C}_{AE}$ *emotional reasoning extension:*
$(a\ \textbf{influences need consistency}\ f\ \textbf{if}\ f_1, \ldots f_n)$ (8)
$(a\ \textbf{influences goal consistency}\ f\ \textbf{if}\ f_1, \ldots f_n)$ (9)
$(a\ \textbf{influences accountability}\ f\ \textbf{if}\ f_1, \ldots f_n)$ (10)
$(a\ \textbf{influences control potential}\ f\ \textbf{if}\ f_1, \ldots f_n)$ (11)
$(f_1, \ldots, f_n\ \textbf{influences need consistency}\ f)$ (12)
$(f_1, \ldots, f_n\ \textbf{influences goal consistency}\ f)$ (13)
$(f_1, \ldots, f_n\ \textbf{influences accountability}\ f)$ (14)
$(f_1, \ldots, f_n\ \textbf{influences control potential}\ f)$ (15)
$(f_1, \ldots, f_n\ \textbf{intervenes action tendency}\ a)$ (16)
$(f_1, \ldots, f_n\ \textbf{facilitates action tendency}\ a)$ (17)

where $a \in \mathbf{A}$ and $a_i \in \mathbf{A}$ $(0 \leq i \leq n)$ and $f_j \in \mathbf{F}$, $(0 \leq j \leq n)$ and $g_j \in \mathbf{F}$, $(0 \leq j \leq n)$, and $f \in \mathbf{F}$ is a well-formed emotion fluent.

The semantics of $\mathcal{C}_{AE}$ is characterized by the constraints of the EDG, captured by the definition of emotional state, specified through a set of static causal laws. In this way, we can restrict states and state-transitions to comply with safe emotional change.

Definition 5 (Emotional state). *An emotional state $s \in S$ of the domain description $D^{ae}(\mathbf{A}, \mathbf{F})$ is an interpretation over F such that*

1. *for every static causal law $(f_1, \ldots, f_n\ \textbf{if}\ g_1, \ldots g_m) \in D^{ae}(\mathbf{A}, \mathbf{F})$, we have $\{f_1, \ldots, f_n\} \subseteq s$ whenever $\{g_1, \ldots g_m\} \subseteq s$.*
2. *for every static causal law $(f_1, \ldots, f_n\ \textbf{influences need consistency}\ f) \in D^{ae}(\mathbf{A}, \mathbf{F})$, we have $\{f\} \subset s$ whenever $\{f_1, \ldots, f_n\} \subseteq s$, and $f \in F^N$.*
3. *for every static causal law $(f_1, \ldots, f_n\ \textbf{influences goal consistency}\ f) \in D^{ae}(\mathbf{A}, \mathbf{F})$, we have $\{f\} \subset s$ whenever $\{f_1, \ldots, f_n\} \subseteq s$, and $f \in F^G$.*
4. *for every static causal law $(f_1, \ldots, f_n\ \textbf{influences accountability}\ f) \in D^{ae}(\mathbf{A}, \mathbf{F})$, we have $\{f\} \subset s$ whenever $\{f_1, \ldots, f_n\} \subseteq s$, and $f \in F^A$.*
5. *for every static causal law $(f_1, \ldots, f_n\ \textbf{influences control potential}\ f) \in D^{ae}(\mathbf{A}, \mathbf{F})$, we have $\{f\} \subset s$ whenever $\{f_1, \ldots, f_n\} \subseteq s$, and $f \in F^C$.*

S denotes all the possible states of $D^{ae}(\mathbf{A}, \mathbf{F})$.

The general definition of emotional state captures a fully connected EDG transition system. For any particular application, we need to define an EDG that, based on application specific interaction goals and relevant theories for emotion regulation, avoids unintended emotional states. Here, we define a safe emotional state that follows principles of HER. Note that this specifies an EDG with a subset of transitions (in the fully connected graph) that is considered safe/valid.

Definition 6 (Safe emotional state). *A safe emotional state $s \in S$ of the domain description $D^{ae}(\mathbf{A},\mathbf{F})$ is an emotional state following principles of hedonic emotion regulation, where s is an interpretation over F such that*

1. *for every static causal law $(f_1,\dots,f_n$ **if** $g_1,\dots g_m) \in D^{ae}(\mathbf{A},\mathbf{F})$, we have $\{f_1,\dots,f_n\} \subseteq s$ whenever $\{g_1,\dots g_m\} \subseteq s$.*
2. *for every static causal law $(f_1,\dots,f_n$ **influences need consistency** $f) \in D^{ae}(\mathbf{A},\mathbf{F})$, we have $\{f\} \subset s$ whenever $\{f_1,\dots,f_n\} \subseteq s$, and $(f \in F^N \wedge f(ne,high,\_) \in s \ \wedge\ \exists f_i \in F^N(1 \le i \le n) \ \wedge\ f_i(ne,low,\_) \in s \ \wedge\ \exists f_j \in F^G(1 \le j \le n) \wedge f_j(go,high,\_) \in s) \vee\ (f \in F^N \wedge f(ne,undecided,\_) \in s \ \wedge\ \exists f_i \in F^N(1 \le i \le n) \ \wedge\ f_i(ne,low,\_) \in s \ \wedge\ \exists f_j \in F^G(1 \le j \le n) \ \wedge f_j(go,high,\_) \in s)$.*
3. *for every static causal law $(f_1,\dots,f_n$ **influences goal consistency** $f) \in D^{ae}(\mathbf{A},\mathbf{F})$, we have $\{f\} \subset s$ whenever $\{f_1,\dots,f_n\} \subseteq s$, and $(f \in F^G \wedge f(go,high,\_) \in s)$.*
4. *for every static causal law $(f_1,\dots,f_n$ **influences accountability** $f) \in D^{ae}(\mathbf{A},\mathbf{F})$, we have $\{f\} \subset s$ whenever $\{f_1,\dots,f_n\} \subseteq s$, and $(f \in F^A \wedge f(ac,other,\_) \in s \ \wedge\ (\exists f_j \in F^G(1 \le j \le n) \ \wedge\ f_j(go,high,\_) \in s)) \vee (f \in F^A \ \wedge\ f(ac,environment,\_) \in s \ \wedge\ (\exists f_j \in F^G(1 \le j \le n) \ \wedge f_j(go,high,\_) \in s)) \vee (f \in F^A \ \wedge\ f(ac,self,\_) \in s \ \wedge\ (\exists f_j \in F^G(1 \le j \le n) \ \wedge\ f_j(go,high,\_) \in s))$.*
5. *for every static causal law $(f_1,\dots,f_n$ **influences control potential** $f) \in D^{ae}(\mathbf{A},\mathbf{F})$, we have $\{f\} \subset s$ whenever $\{f_1,\dots,f_n\} \subseteq s$, and $((f \in F^C \wedge f(co,high,\_) \in s \vee f(co,undecided,\_) \in s) \wedge (\exists f_j \in F^G(1 \le j \le n) \ \wedge f_j(go,high,\_) \in s)) \vee\ (f \in F^C \wedge f(co,high,\_) \in s \wedge (\exists f_i \in F^N(1 \le i \le n) \wedge (f_i(ne,low,\_) \in s \vee f_i(ne,undecided,\_) \in s) \wedge (\exists f_j \in F^G(1 \le j \le n) \wedge f_j(go,low,\_) \in s) \wedge (\exists f_k \in F^A(1 \le k \le n) \wedge f_k(ac,environment,\_) \in s)))$.*

S denotes all the possible safe emotional states of $D^{ae}(\mathbf{A},\mathbf{F})$.

Definition 7. *Let $D^{ae}(\mathbf{A},\mathbf{F})$ be a domain description and s a state of $D^{ae}(\mathbf{A},\mathbf{F})$.*

1. *An inhibition rule $(f_1,\dots,f_n$ **inhibits** $a)$ is active in s, if $s \models f_1,\dots,f_n$, otherwise, passive. The set $A_I(s)$ is the set of actions for which there exists at least one active inhibition rule in s (as in $\mathcal{C}_{TAID}$ [5]).*
2. *A triggering rule $(f_1,\dots,f_n$ **triggers** $a)$ is active in s, if $s \models f_1,\dots,f_n$ and all inhibition rules of action a are passive in s, otherwise, the triggering rule is passive in s. The set $A_T(s)$ is the set of actions for which there exists at least one active triggering rule in s. The set $\overline{A}_T(s)$ is the set of actions for which there exists at least one triggering rule and all triggering rules are passive in s (as in $\mathcal{C}_{TAID}$ [5]).*

3. *An allowance rule* ($f_1, \ldots, f_n$ ***allows*** a) *is active in* s, *if* $s \models f_1, \ldots, f_n$ *and all inhibition rules of action* a *are passive in* s, *otherwise, the allowance rule is passive in* s. *The set* $A_A(s)$ *is the set of actions for which there exists at least one active allowance rule in* s. *The set* $\overline{A}_A(s)$ *is the set of actions for which there exists at least one allowance rule and all allowance rules are passive in* s *(as in* $\mathcal{C}_{TAID}$ *[5]).*
4. *A facilitating rule* ($f_1, \ldots, f_n$ ***facilitates action tendency*** a) *is active in* s, *if* $a \in \mathbf{A}^H$ *and* $s \models f_1, \ldots, f_n$ *and all inhibition rules and intervening rules of action* a *are passive in* s, *otherwise, the facilitating rule is passive in* s. *The set* $A_{FAC}(s)$ *is the set of actions for which there exists at least one active facilitating rule in* s. *The set* $\overline{A}_{FAC}(s)$ *is the set of actions for which there exists at least one facilitating rule and all facilitating rules are passive in* s.
5. *An intervening rule* ($f_1, \ldots, f_n$ ***intervenes action tendency*** a) *is active in* s, *if* $a \in \mathbf{A}^H$ *and* $s \models f_1, \ldots, f_n$ *and all inhibition rules and facilitating rules of action* a *are passive in* s, *otherwise, the intervening rule is passive in* s. *The set* $A_{INT}(s)$ *is the set of actions for which there exists at least one active intervening rule in* s. *The set* $\overline{A}_{INT}(s)$ *is the set of actions for which there exists at least one intervening rule and all intervening rules are passive in* s.
6. *A dynamic causal law* (a *causes* $f_1, \ldots, f_n$ *if* $g_1, \ldots, g_n$) *is applicable in* s, *if* $s \models g_1, \ldots, g_n$.
7. *A static causal law* ($f_1, \ldots, f_n$ *if* $g_1, \ldots, g_n$) *is applicable in* s, *if* $s \models g_1, \ldots, g_n$.
8. *A dynamic causal law* (a ***influences need consistency*** f *if* $f_1, \ldots, f_n$) *is applicable in* s, *if* $s \models f_1, \ldots, f_n$, *and* $f \in F^N$, *and* $\exists f_i \in F^N(1 \leq i \leq n)$, *and* $\exists f_j \in F^G(1 \leq j \leq n)$, *and* $\exists f_k \in F^A(1 \leq k \leq n)$, *and* $\exists f_m \in F^C(1 \leq m \leq n)$.
9. *A dynamic causal law* (a ***influences goal consistency*** f *if* $f_1, \ldots, f_n$) *is applicable in* s, *if* $s \models f_1, \ldots, f_n$, *and* $f \in F^G$, *and* $\exists f_i \in F^N(1 \leq i \leq n)$, *and* $\exists f_j \in F^G(1 \leq j \leq n)$, *and* $\exists f_k \in F^A(1 \leq k \leq n)$, *and* $\exists f_m \in F^C(1 \leq m \leq n)$.
10. *A dynamic causal law* (a ***influences accountability*** f *if* $f_1, \ldots, f_n$) *is applicable in* s, *if* $s \models f_1, \ldots, f_n$, *and* $f \in F^A$, *and* $\exists f_i \in F^N(1 \leq i \leq n)$, *and* $\exists f_j \in F^G(1 \leq j \leq n)$, *and* $\exists f_k \in F^A(1 \leq k \leq n)$, *and* $\exists f_m \in F^C(1 \leq m \leq n)$.
11. *A dynamic causal law* (a ***influences control potential*** f *if* $f_1, \ldots, f_n$) *is applicable in* s, *if* $s \models f_1, \ldots, f_n$, *and* $f \in F^C$, *and* $\exists f_i \in F^N(1 \leq i \leq n)$, *and* $\exists f_j \in F^G(1 \leq j \leq n)$, *and* $\exists f_k \in F^A(1 \leq k \leq n)$, *and* $\exists f_m \in F^C(1 \leq m \leq n)$.
12. *A static causal law* ($f_1, \ldots, f_n$ ***influences need consistency*** f) *is applicable in* s, *if* $s \models f_1, \ldots, f_n$, *and* $f \in F^N$, *and* $\exists f_i \in F^N(1 \leq i \leq n)$, *and* $\exists f_j \in F^G(1 \leq j \leq n)$, *and* $\exists f_k \in F^A(1 \leq k \leq n)$, *and* $\exists f_m \in F^C(1 \leq m \leq n)$.
13. *A static causal law* ($f_1, \ldots, f_n$ ***influences goal consistency*** f) *is applicable in* s, *if* $s \models f_1, \ldots, f_n$, *and* $f \in F^G$, *and* $\exists f_i \in F^N(1 \leq i \leq n)$, *and*

$\exists f_j \in F^G (1 \leq j \leq n)$, and $\exists f_k \in F^A (1 \leq k \leq n)$, and $\exists f_m \in F^C (1 \leq m \leq n)$.

14. *A static causal law ($f_1, \ldots, f_n$ **influences accountability** f) is applicable in s, if $s \models f_1, \ldots, f_n$, and $f \in F^A$, and $\exists f_i \in F^N (1 \leq i \leq n)$, and $\exists f_j \in F^G (1 \leq j \leq n)$, and $\exists f_k \in F^A (1 \leq k \leq n)$, and $\exists f_m \in F^C (1 \leq m \leq n)$.*
15. *A static causal law ($f_1, \ldots, f_n$ **influences control potential** f) is applicable in s, if $s \models f_1, \ldots, f_n$, and $f \in F^C$, and $\exists f_i \in F^N (1 \leq i \leq n)$, and $\exists f_j \in F^G (1 \leq j \leq n)$, and $\exists f_k \in F^A (1 \leq k \leq n)$, and $\exists f_m \in F^C (1 \leq m \leq n)$.*

Definition 8 (Trajectory). *Let $D^{ae}(\mathbf{A}, \mathbf{F})$ be a domain description. A trajectory $\langle s_0, A_1, s_1, A_2, \ldots, A_n, s_n \rangle$ of $D^{ae}(\mathbf{A}, \mathbf{F})$ is a sequence of sets of actions $A_i \subseteq A$ and states s_i of $D^{ae}(\mathbf{A}, \mathbf{F})$ satisfying the following conditions for 0 $\leq$ i $<$ n:*

1. $(s_i, A, s_{i+1}) \in S \times 2^A \backslash \{\} \times S$
2. $A_T(s_i) \subseteq A_{i+1}$
3. $A_{FAC}(s_i) \subseteq A_{i+1}$
4. $\underline{A}_{INT}(s_i) \subseteq A_{i+1}$
5. $\overline{\mathrm{A}}_T(s_i) \cap A_{i+1} = \emptyset$
6. $\overline{\mathrm{A}}_A(s_i) \cap A_{i+1} = \emptyset$
7. $\underline{A}_I(s_i) \cap A_{i+1} = \emptyset$
8. $\overline{\mathrm{A}}_{FAC}(s_i) \cap A_{i+1} = \emptyset$
9. $\overline{\mathrm{A}}_{INT}(s_i) \cap A_{i+1} = \emptyset$
10. $|A_i \cap B| \leq 1 \ for \ all \ (noconcurrency \ B) \ \in \ D^{ae}(\mathbf{A}, \mathbf{F})$.

Definition 9 (Action Observation Language). *The action observation language of $\mathcal{C}_{AE}$ (similar to $\mathcal{C}_{TAID}$) consists of expressions of the following form:*
$(f \ \ \mathbf{at} \ \ t_i) \quad (a \ \ \mathbf{occurs\_at} \ \ t_i)$ (8)
where $f \in \mathbf{F}$, a is an action and t_i is a point of time.

Definition 10 (Action Theory). *Let D be a domain description and O be a set of observations. The pair (D, O) is called an action theory.*

Definition 11 (Trajectory Model). *Let (D, O) be an action theory. A trajectory $\langle s_0, A_1, s_1, A_2, \ldots, A_n, s_n \rangle$ of D is a trajectory model of (D, O), if it satisfies all observations of O in the following way:*

1. *if (f at t) $\in$ O, then $f \in s_t$*
2. *if (a occurs_at t) $\in$ O, then $a \in A_{t+1}$.*

Definition 12 (Action Query Language). *The action query language of $\mathcal{C}_{AE}$ regards assertions about executing sequences of actions with expressions that constitute trajectories. A query is of the following form: ($f_1, \ldots, f_n$ **after** A_i **occurs_at** $t_i, \ldots, A_m$ **occurs_at** t_m) where $f_1, \ldots, f_n$ are fluent literals $\in \mathbf{F}$, $A_i, \ldots, A_m$ are subsets of $\mathbf{A}$, and $t_i, \ldots, t_m$ are points in time.*

We can observe that actions in a trajectory model can be actions executed by a rational agent, to influence appraisals of the situation, or action tendencies estimated to be executed by the human agent. Adjustments of appraisal must be done in a controlled and safe way to reduce unintended emotional side-effects. In the next section, we present a proof for safe emotional change.

4.3 Proving safe emotional change

We present a theorem and prove that trajectories generated by $\mathcal{C}_{AE}$ preserve a safety property in terms of avoiding unintended emotional side-effects. An invariance property is defined by following principles of hedonic emotion regulation, called an Emotional Invariant (EI), a state predicate which is preserved by the state conditions of the EDG. This is proven using the invariance principle [8]. To support readability of the proof, we define an emotion labeling.

Definition 13 (Emotion labeling). *For any trajectory $\langle s_0, A_1, s_1, A_2, \ldots, A_n, s_n\rangle$ of $D^{ae}(\mathbf{A}, \mathbf{F})$, there is a transition emotion labeling $\langle E_O, \ldots, E_n\rangle$ such that Labeling(s_i) = E_i $(0 \leq i \leq n)$, and E_i = [V_N,V_G,V_A,V_C, i], where V_N, V_G, V_A, V_C are values of well-formed emotion fluents e_N, e_G, e_A, $e_C \in s_i$, representing need consistency, goal consistency, accountability and control potential, respectively.*

Theorem 1 (Safe emotional change). *Let $(D^{ae}, O_{initial})$ be an action theory such that $O_{initial}$ are the fluent observations of the initial state, i.e., the fluents of the situation/interaction and the fluents of the estimated emotional state of the human agent. Let Q be a query according to Definition 12 and let*

$$A_Q = \{(a\ occurs\_at\ t_i) \mid a \in A_i, 1 \leq i \leq m\}.$$

If there is a trajectory model $M = \langle s_0, A_1, s_1, A_2, \ldots, A_n, s_m\rangle$ where $A_i \subseteq \mathbf{A}$ $(0 \leq i \leq m)$ of $\mathcal{C}_{AE}$ $(D^{ae}, O_{initial} \cup A_Q)$, then all states $s \in M$ at the time points $0 \leq t \leq m$ preserve a state predicate EI, where the goal consistency is equal or higher than the need consistency, denoted according to Definition 6 as [V_N,V_G,V_A,V_C,t] $\wedge$ $V_N \leq V_G$ and where V_N, $V_G \in$ {low, undecided, high} are ranked as low < undecided < high (following the intuition of AE in Section 3).

Proof. We must show that EI holds in each state condition (Definition 6) of the EDG. We do this by showing that an initial observation holds, which we specify as [undecided, undecided, undecided, undecided, 0] $\wedge$ $V_N \leq V_G$. We then show that any transition from time step t to t+1 preserves EI, such that

[V_N,V_G,V_A,V_C,t] $\wedge$ $V_N \leq V_G$ implies [V_N',V_G',V_A',V_C',t+1] $\wedge$ V_N' $\leq$ V_G'.

Looking at each transition rule, we can observe that

- it is clear that the emotional invariant [V_N,V_G,V_A,V_C,t] $\wedge$ $V_N \leq V_G$ holds in the initial observation [undecided, undecided, undecided, undecided, 0] $\wedge$ undecided $\leq$ undecided.
- for every static causal law $(f_1, \ldots, f_n$ **influences need consistency** $f) \in D^{ae}(\mathbf{A}, \mathbf{F})$, only changes of V_N to high or undecided are permitted, and require that V_G is high. It is clear that *undecided* $\vee$ *high* $\leq$ *high* preserves EI in a transition from t to t+1.
- for every static causal law $(f_1, \ldots, f_n$ **influences goal consistency** $f) \in D^{ae}(\mathbf{A}, \mathbf{F})$, only changes of V_G to high are permitted. It is clear that a condition $V_N \leq$ *high* preserves EI in a transition from t to t+1.

- for every static causal law ($f_1, \ldots, f_n$ **influences accountability** f) $\in D^{ae}(\mathbf{A}, \mathbf{F})$, no changes regard either V_N or V_G, which preserves EI in a transition from t to t+1.
- for every static causal law ($f_1, \ldots, f_n$ **influences control potential** f) $\in D^{ae}(\mathbf{A}, \mathbf{F})$, no changes regard either V_N or V_G, which preserves EI in a transition from t to t+1.

We can conclude, by looking at an initial observation, and all state conditions in any transition from time step t to t+1, that the emotional invariant is preserved in the EDG according to hedonic emotion regulation, and show that the system avoids unintended emotional side-effects.

4.4 Example scenario: Backward reasoning

Backward reasoning is a process of searching past states in the interaction to reason about why a certain emotional state was reached. In the case of AE, this is explained by changes in appraisal of a situation. For instance, in the past trajectory:
$\langle\ s_0 : \{Frustration[h, l, e, h, 0]\}$,
$A_1 : \{Influence\_accountability(o)\}$,
$s_1 : \{Anger[h, l, o, h, 1]\}\rangle$
In this example, the agent looks one state backward (s_0) to find that the emotional state of frustration led to the emotional state of anger in the initial state (s_1). In addition, the agent can find that the state of anger was promoted due to a change of accountability from *environment* (e) to *other* (o). Such inferences can be taken in consideration when planning future interactions.

4.5 Example scenario: Forward reasoning

Forward reasoning is a process of planning future interactions by considering emotional change in response to actions that adapt the human agent's appraisal. This is a process of generating a set of alternative trajectories for reaching the goal of the interaction while reasoning about emotions in each state of the interaction. For instance, an alternative trajectory can be:
$\langle\ s_0 : \{Anger[h, l, o, h, 0]\}$,
$A_1 : \{Influence\_accountability(e)\}$,
$s_1 : \{Frustration[h, l, e, h, 1]\}$,
$A_2 : \{Influence\_need(u),\ Influence\_goal(h),\ Influence\_control(l)\}$,
$s_2 : \{Hope[u, h, e, l, 2]\}$,
$A_3 : \{Influence\_need(h),\ Influence\_control(u)\}$,
$s_3 : \{Joy[h, h, e, u, 3]\}\rangle$.

In this example, starting in an emotional state of anger, the agent plans an interaction while managing the human agent's emotions to decrease frustration and maintain a pleasurable interaction. Following the specified transition system for safe emotional change (Definition 6), the agent filters alternative trajectories and selects actions to avoid negative emotional side-effects.

5 Discussion

In this paper, we introduce *emotion-aware planning*. Emotional reasoning has been formalized in a structure called $\mathcal{C}_{AE}$, in terms of action reasoning and transition systems, formalizing the emotion theories of AE and HER. This constitutes computational models for emotions and emotional change, which can provide emotion-aware planning and decision-making in human-agent interactions.

An emotional state, to be captured by an agent, needs a representation of the emotion. Through a set of variables, recognized by an aggregation of the appraisal theory of emotion, abstractions of emotions are given. The emotion decision graph (transition system) is a representation, and we expect human emotions to be represented there. In that respect, the agent creates a theory of the mind of the human as an abstraction based on appraisal theory of emotion. This is one of the main contributions of this paper; We take psychological (emotion) theories and transform them into tangible, computational and multi-dimensional models of emotion.

Limitations of the proposed framework can be inherited from the appraisal theory of emotion, where emotions are solely based on appraisal [6]. This can limit the expressiveness of the model, not accounting for other components of emotions which are not related to human conscious reasoning. There are many other emotion theories that can be applied to model emotional states. For instance, emotions can be defined in terms of Arousal and Valence [10]. However, the chosen theory is particularly interesting for the current work due to its way of capturing emotional causes.

6 Conclusion and Future Work

The proposed framework for emotional reasoning enables a software agent to acquire a particular theory of the mind of the human to deal with emotions in interaction. The formal specifications assure that generated plans comply with safe emotional change. The main contribution of this paper is a framework to enable: 1) backward reasoning, by modelling causes to emotions; 2) context reasoning, to infer emotional states; 3) forward reasoning, by modelling emotional change in terms of state transitions; and 4) emotion-aware planning, to plan an agent's actions to be in balance with emotions in each state of the interaction, aiming to avoid unintended emotional side-effects.

The specified EDG filters trajectories by capturing principles of AE and HER, aiming to reduce negative emotions and increase positive emotions. However, depending on the goal of the interaction (such as stress-management, coaching or therapy), different emotion regulation theories are suitable. In a generalization of the framework, we can replace AE and HER for other emotion theories (such as the Two-Factor Theory of Emotion [3] or the Cognitive-Mediational Theory [16]). In this way, the proposed emotional reasoning framework can provide a modular tool for integrating, evaluating and comparing different emotion theories (by analyzing filtered trajectories). This is a focus for future work.

Acknowledgements

This work was partially funded by the Knut and Alice Wallenberg Foundation. We thank the anonymous reviewers for their valuable and useful comments.

References

1. Belkaid, M., Sabouret, N.: A logical model of theory of mind for virtual agents in the context of job interview simulation. arXiv preprint arXiv:1402.5043 (2014)
2. Blanchette, I.: Emotion and reasoning. Psychology Press (2013)
3. Cornelius, R.R.: Gregorio marañon's two-factor theory of emotion. Personality and Social Psychology Bulletin **17**(1), 65–69 (1991)
4. Di Lascio, E.: Emotion-aware systems for promoting human well-being. In: Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers. pp. 529–534 (2018)
5. Dworschak, S., Grell, S., Nikiforova, V., Schaub, T., Selbig, J.: Modeling Biological Networks by Action Languages via Answer Set Programming. Constraints **13**(1), 21–65 (2008)
6. Ellsworth, P.C.: Appraisal theory: Old and new questions. Emotion Review **5**(2), 125–131 (2013)
7. Gelfond, M., Lifschitz, V.: Action languages. Computer and Information Science **3**(16) (1998)
8. Hansen, M.N., Schmidt, E.M.: Algorithms and data structures: transition systems. Datalogisk Institut, Aarhus Universitet (2003)
9. Jiang, H., Vidal, J.M., Huhns, M.N.: Ebdi: an architecture for emotional agents. In: Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems. pp. 1–3 (2007)
10. Knez, I., Hygge, S.: The circumplex structure of affect: A swedish version. Scandinavian Journal of Psychology **42**(5), 389–398 (2001)
11. Leonetti, M., Iocchi, L., Cohn, A.G., Nardi, D.: Adaptive human-aware task planning. In: ICAPS Workshop on Planning and Robotics (PlanRob) (2019)
12. Ong, D.C., Zaki, J., Goodman, N.D.: Computational models of emotion inference in theory of mind: A review and roadmap. Topics in cognitive science **11**(2), 338–357 (2019)
13. Ortner, C.N., Corno, D., Fung, T.Y., Rapinda, K.: The roles of hedonic and eudaimonic motives in emotion regulation. Personality and Individual Differences **120**, 209–212 (2018)
14. Pereira, D., Oliveira, E., Moreira, N.: Formal modelling of emotions in bdi agents. In: International Workshop on Computational Logic in Multi-Agent Systems. pp. 62–81. Springer (2007)
15. Sánchez-López, Y., Cerezo, E.: Designing emotional bdi agents: good practices and open questions. The Knowledge Engineering Review **34** (2019)
16. Schulz, M.S., Lazarus, R.S.: Regulating emotion in adolescence: A cognitive-mediational conceptualization. (2012)
17. Zaki, J.: Integrating empathy and interpersonal emotion regulation. Annual Review of Psychology **71**, 517–540 (2020)