A logic for reasoning about counterfactual emotions (IJCAI 2009)

Emiliano Lorini, François Schwarzentruber

A logic for reasoning about counterfactual emotions

Emiliano Lorini and **François Schwarzentruber**
IRIT, Toulouse, France

Abstract

The aim of this work is to propose a logical framework for the specification of cognitive emotions that are based on counterfactual reasoning about agents' choices. An example of this kind of emotions is regret. In order to meet this objective, we exploit the well-known STIT logic [Belnap *et al.*, 2001; Horty, 2001]. STIT logic has been proposed in the domain of formal philosophy in the nineties and, more recently, it has been imported into the field of theoretical computer science where its formal relationships with other logics for multi-agent systems such as ATL and Coalition Logic (CL) have been studied. STIT is a very suitable formalism to reason about choices and capabilities of agents and groups of agents. Unfortunately, the version of STIT with agents and groups has been recently proved to be undecidable. In this work we study a decidable fragment of STIT with agents and groups which is sufficiently expressive for our purpose of formalizing counterfactual emotions.

1 Introduction

A major objective of AI is to develop interactive cognitive systems that are more attractive and closer to the users and that can be considered as believable interlocutors. In this perspective, a challenge for AI is to build artificial agents which are capable: to reason about emotions, to predict and understand human emotions, and to process emotions in reasoning and during their interaction with a human user. With the aim of creating a new generation of emotional interaction systems, the study of affective phenomena has become a "hot" topic in AI where the domain of Affective Computing [Picard, 1997] has emerged in the last few years.

Recently, some researchers have been interested in developing logical frameworks for the formal specification of emotions (see [Meyer, 2006; Steunebrink *et al.*, 2007; El-Nasr *et al.*, 2000] for instance). Their main concern is to exploit logical methods in order to provide a rigorous specification of how emotions should be implemented in an artificial agent. The design of agent-based systems where agents are capable to reason about and to display some kind of emotions can indeed benefit from the accuracy of logical methods.

Although the application of logical methods to the formal specification of emotions has been quite successful, there is still much work to be done in the field of computational and logical modeling of 'counterfactual emotions'. In line with psychological theories of 'counterfactual emotions', we use this term to denote those emotions such as regret which arise during 'counterfactual thinking', that is, when "[...] reality is compared to an imagined view of what might have been." [Kahneman and Miller, 1986, p. 136]. In other terms, counterfactual emotions are based on an agent's *alteration* of a factual situation and in the agent's *imagination* of an alternative situation that could have realized if something different was done [Roese *et al.*, 2005].

The aim of our work is to advance the state of the art on computational modeling of emotions by providing a logic which supports reasoning about this kind of emotions. Our major concern here is to find a fair trade off between expressivity and complexity of the formalism. We want a logic which is sufficiently expressive to capture the fundamental constituents of counterfactual emotions and, at the same time, with good mathematical properties in terms of decidability and complexity. To this aim, we exploit a well-known logic called STIT [Belnap *et al.*, 2001; Horty, 2001]. STIT logic has been proposed in the domain of formal philosophy in the nineties and, more recently, it has been imported into the field of theoretical computer science where its formal relationships with other logics for multi-agent systems have been studied (see [Broersen *et al.*, 2006] for instance). It is a very suitable formalism to make counterfactual reasoning about choices of agents and of groups of agents. Unfortunately, the version of STIT with agents and groups proposed in [Horty, 2001] has been recently proved to be undecidable [Herzig and Schwarzentruber, 2008]. In this work we study a decidable fragment of this logic which is sufficiently expressive for our purpose of formalizing counterfactual emotions.

The paper is organized as follows. In Section 2 we introduce a fragment of the version of STIT with agents and groups proposed in [Horty, 2001]. Differently from Horty's logic, we prove that this fragment is decidable. Section 3 is devoted to characterize in our STIT fragment counterfactual statements of the form "group J (or agent i) *could have prevented* χ to be true". These are indeed fundamental constituents of counterfactual emotions. In Section 4 we provide an extension of our STIT fragment with knowledge opera-

This works has been supported by the Agence Nationale de la Recherche, contract No. ANR-08-CORD-009.

tors. This is a necessary step in order to capture the subjective dimension of the affective phenomena we intend to analyze. The last part of the paper (Section 5) is devoted to the formalization of two kinds of counterfactual emotions: *regret* and *rejoicing*.

2 A decidable fragment of STIT

STIT (the logic of "Seeing to it that") is a modal logic framework dealing with what agents and groups of agents do and can do. More precisely, STIT supports reasoning about the effects of actions of agents and groups, and about the capabilities of agents and groups to ensure certain outcomes. In [Belnap *et al.*, 2001] the language of STIT without groups is studied: a complete axiomatization of STIT without groups is provided and the logic is proved to be decidable. The extension of STIT with groups has been proposed in [Horty, 2001]. Unfortunately, in [Herzig and Schwarzentruber, 2008] it has been proved to be undecidable.

We here introduce a decidable fragment of STIT with agents and groups called *df*STIT which is sufficiently expressive to formalize counterfactual emotions.

2.1 Syntax

Let n be a strictly positive integer. Let ATM be a countable set of atomic propositions and let $AGT = \{1, \ldots, n\}$ be a countable set of agents. The language $\mathcal{L}_{\mathsf{STIT}}$ of the logic STIT with agents and groups proposed by Horty [Horty, 2001] is defined by the following BNF:

$$\varphi ::= p \mid \varphi \wedge \varphi \mid \neg\varphi \mid [J]\varphi$$

where p ranges over ATM and J over 2^{AGT}. $\langle J\rangle\varphi$ is an abbreviation of $\neg[J]\neg\varphi$. Operators of type $[J]$ are used to describe the effects of the action that has be chosen by J. If J is a singleton we refer to J as an *agent*, whereas if J has more than one element we refer to J as a *group*. We call *joint actions* the actions chosen by groups. If J has more than one element the construction $[J]\varphi$ means "group J sees to it that φ no matter what the other agents in $AGT \setminus J$ do". If J is a singleton $\{i\}$ the construction $[\{i\}]\varphi$ means "agent i sees to it that φ no matter what the other agents in $AGT \setminus \{i\}$ do". For notational convenience, we write $[i]$ instead of $[\{i\}]$. $[\emptyset]\varphi$ can be shorten to "φ is necessarily true". The dual expression $\langle\emptyset\rangle\varphi$ means "φ is possibly true". Note that the operators $\langle\emptyset\rangle$ and $[J]$ can be combined in order to express what agents and groups can do: $\langle\emptyset\rangle[J]\varphi$ means "J can see to it that φ whatever the other agents in $AGT \setminus J$ do".

The STIT fragment we are interested in here is called *df*STIT and is defined by the following BNF:

$$\chi ::= \bot \mid p \mid \chi \wedge \chi \mid \neg\chi \text{ (propositional formulas)}$$
$$\psi ::= [J]\chi \mid \psi \wedge \psi \text{ (see-to-it formulas)}$$
$$\varphi ::= \chi \mid \psi \mid \varphi \wedge \varphi \mid \neg\varphi \mid \langle\emptyset\rangle\psi \text{ (see-to-it and "can" formulas)}$$

where p ranges over ATM and J over $2^{AGT} \setminus \emptyset$.

2.2 Models

Here we give two semantics of STIT. It is proved in [Herzig and Schwarzentruber, 2008] that these two semantics are equivalent. The first one corresponds to the original semantics of STIT with agents and groups given in [Horty, 2001]. The other one is based on the product logic $S5^n$ [Gabbay *et al.*, 2003] and is used in the proof of decidability of the satisfiability problem of a *df*STIT -formula (Theorem 1). Let us give first the original semantics of STIT.

Definition 1. *A STIT-model is a tuple $\mathcal{M} = (W, \{R_J\}_{J\subseteq AGT}, V)$ where:*

- *W is a non-empty set of possible worlds or states;*
- *For all $J \subseteq AGT$, R_J is an equivalence relation over W such that:*
  1. *$R_J \subseteq R_\emptyset$;*
  2. *$R_J = \bigcap_{j\in J} R_{\{j\}}$;*
  3. *for all $w \in W$, for all $(w_j)_{j\in AGT} \in R_\emptyset(w)^n$, $\bigcap_{j\in AGT} R_{\{j\}}(w_j) \neq \emptyset$;*
  4. *$R_{AGT} = id_W$.*
- *V is a valuation function, that is, $V : W \to 2^{ATM}$.*

As in the previous Constraint 3, it is convenient to view relations on W as functions from W to 2^W, that is, for every $J \in 2^{AGT}$, $R_J(w) = \{v \in W \mid wR_Jv\}$. R_J represents the actual action chosen by J: if wR_Jv then v is an outcome of the action chosen by J at w. We recall that $R_\emptyset$ is the relation over all possible outcomes: if w is the current world and $wR_\emptyset v$ then v is a possible outcome at w. Thus, Constraint 1 on STIT models just means that all outcomes brought about by J are possible outcomes. Constraint 2 just says that the set of outcomes brought about by J at a given world w is equal to the pointwise intersection of the sets of outcomes brought about by the agents in J at w. Constraint 3 expresses a so-called *assumption of independence of agents*: if $w_1, \ldots, w_n$ are possible outcomes at w then the intersection of the set of outcomes that agent 1 brings about at w_1, and the set of outcomes that agent 2 brings about at w_2,..., and the set of outcomes that agent n brings about at w_n is not empty. More intuitively, this means that agents can never be deprived of choices due to the choices made by other agents. Constraint 4 expresses an assumption of determinism: the set of outcomes brought about by all agents is a singleton.

Truth conditions for atomic formulas and the boolean operators are entirely standard. For every $J \in 2^{AGT}$, the truth condition of the modal operator $[J]$ is:

$$\mathcal{M}, w \models [J]\varphi \text{ iff } \mathcal{M}, v \models \varphi \text{ for all } v \in W \text{ such that } wR_Jv.$$

The alternative semantics of STIT is based on the product logic $S5^n$. It is defined as follows:

Definition 2. *A product STIT-model is a tuple $\mathcal{M} = (W, V)$ where:*

- *$W = W_1 \times \cdots \times W_n$ where W_i are non-empty sets of worlds or states;*
- *V is a valuation function, that is, $V : W \to 2^{ATM}$.*

The truth conditions for the modal operators $[J]$ in product STIT-models are:

$$\mathcal{M}, (w_1, \ldots, w_n) \models [J]\varphi \text{ iff } \mathcal{M}, (v_1, \ldots, v_n) \models \varphi$$
$$\text{for all } (v_1, \ldots, v_n) \in W \text{ such that } v_j = w_j \text{ if } j \in J.$$

A formula φ is STIT-valid (noted $\models_{\mathsf{STIT}} \varphi$) iff φ is true in every world of every STIT-model (or product STIT-model).

2.3 Decidability

Our fragment *df* STIT of STIT with agents and groups has interesting computational properties. In particular:

Theorem 1. *The problem of satisfiability of df STIT is NP-complete.*

This is implied by the following fact:

Theorem 2. *If a df STIT -formula is satisfiable, then it is in a polynomial sized model.*

Sketch of proof. The main idea of the proof is a selection-of-points argument as in [Ladner, 1977].[1] Let φ a satisfiable formula: there exists a product STIT-model $\mathcal{M} = (W, V)$ and w such that $\mathcal{M}, w \models \varphi$. First, we construct product STIT-model $\mathcal{M}'$ satisfying φ with selected points of the initial model $\mathcal{M}$:

- In the construction, we take care to create a new point in $\mathcal{M}'$ for each subformula $\langle\emptyset\rangle\psi$ of φ true in $\mathcal{M}$.
- We also take care to construct enough points so that all subformulas $\langle\emptyset\rangle\psi$ and $[J]\chi$ of φ false in $\mathcal{M}$ are false in $\mathcal{M}'$ as well.

Secondly we make sure that $\mathcal{M}'$ is polynomial sized and that there is a point w' so that $\mathcal{M}', w' \models \varphi$. □

3 Counterfactual statements in STIT

The following counterfactual statement is a fundamental constituent of an analysis of counterfactual emotions:

(*) group J (or agent i) *could have prevented* a certain state of affairs χ to be true now.

Our STIT fragment enables a formal translation of it. We note $\texttt{CHP}_J\chi$ this translation, where $\texttt{CHP}_J\chi$ is defined as follows:

$$\texttt{CHP}_J\chi \stackrel{\text{def}}{=} \chi \wedge \neg[AGT \setminus J]\chi.$$

The expression $\neg[AGT \setminus J]\chi$ just means that: the complement of J with respect to AGT (i.e. $AGT \setminus J$) does not see to it that χ (no matter what the agents in J have chosen to do). This is the same thing as saying that: given what the agents in $AGT \setminus J$ have chosen, there exists an alternative joint action of the agents in J such that, if the agents in J did choose this action, χ would be false now. Thus, χ and $\neg[AGT \setminus J]\chi$ together correctly translate the previous counterfactual statement (*).

EXAMPLE. Imagine a typical coordination scenario with two agents $AGT = \{1, 2\}$. Agents 1 and 2 have to take care of a plant. Each agent has only two actions available: water the plant ($water$) or do nothing ($skip$). If either both agents water the plant or both agents do nothing, the plant will die ($dead$). In the former case the plant will die since it does not tolerate too much water. In the latter case it will die since it lacks water. If one agent waters the plant and the other does nothing, the plant will survive ($\neg dead$). The scenario is represented in the STIT model in Fig. 1. For instance both at world w_2 and w_4, formulas $\texttt{CHP}_1\, dead$ and $\texttt{CHP}_2\, dead$ are true: each agent could have prevented the plant to be dead. Indeed, at world w_2, $dead$ and $\neg[2]\, dead$ are true: given what agent 2 has chosen (i.e. $water$), there exists an alternative action of agent 1 (i.e. $skip$) such that, if 1 did choose this action, $dead$ would be false now. At world w_4, $dead$ and $\neg[2]\, dead$ are also true: given what agent 2 has chosen (i.e. $skip$), there exists an alternative action of agent 1 (i.e. $water$) such that, if 1 did choose this action, $dead$ would be false now. The case for agent 2 is completely symmetrical.

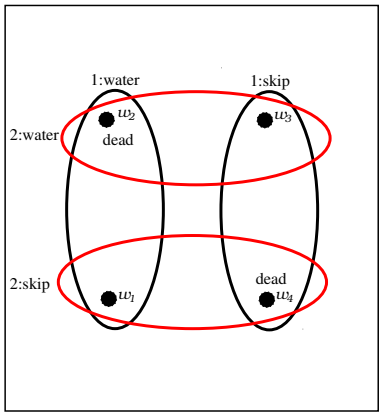

Figure 1: The four worlds w_1, w_2, w_3 and w_4 are in the equivalence class determined by $R_\emptyset$. Vertical circles represent the actions that agent 1 can choose, whereas horizontal circles represent the actions that agent 2 can choose. For example, w_1 is the world that results from agent 1 choosing the action $water$ and agent 2 choosing the action $skip$.

The following are some interesting properties of the operator $\texttt{CHP}_J$. If $J_1 \subseteq J_2$ then:

(1) $\models_{\text{STIT}} (\texttt{CHP}_{J_1}\chi_1 \vee \texttt{CHP}_{J_1}\chi_2) \leftrightarrow \texttt{CHP}_{J_1}(\chi_1 \vee \chi_2)$

(2) $\models_{\text{STIT}} \texttt{CHP}_{J_1}\chi \rightarrow \texttt{CHP}_{J_2}\chi$

(3) $\models_{\text{STIT}} \texttt{CHP}_{J_1}(\chi_1 \wedge \chi_2) \rightarrow (\texttt{CHP}_{J_1}\chi_1 \wedge \texttt{CHP}_{J_1}\chi_2)$

Proof. We give the proof of Validity 2 as an example. Let $\mathcal{M}$ be a STIT-model and $w \in W$ such that $\mathcal{M}, w \models \texttt{CHP}_{J_1}\chi$. We have $\mathcal{M}, w \models \chi$ and $\mathcal{M}, w \models \neg[AGT \setminus J_1]\chi$. As $R_{AGT\setminus J_1} \subseteq R_{AGT\setminus J_2}$, it implies that $\mathcal{M}, w \models \neg[AGT \setminus J_2]\chi$. That is why we have $\mathcal{M}, w \models \texttt{CHP}_{J_2}\chi$. □

According to Validity 1 J_1 could have prevented χ_1 or χ_2 to be true if and only if, J_1 could have prevented χ_1 or could have prevented χ_2. Validity 2 expresses a monotonicity property: if J_1 is a subset of J_2 and J_1 could have prevented χ then, J_2 could have prevented χ as well. Finally Validity 3 shows how the operator $\texttt{CHP}_J$ behaves over conjunction: if J_1 could have prevented χ_1 and χ_2 then, J_1 could have prevented χ_1 and could have prevented χ_2 separately.

REMARK. It is worth noting that counterfactual statements of the form "group J (or agent i) *could have prevented* χ to be true", which are expressible in STIT, are not expressible in other well-known logics of multi-agent interaction such as

[1] Space restrictions prevent from giving the extensive proof of the theorem. The interested reader might read it at the anonymous URL http://rapidshare.com/files/182384345/proofdfstitNP.pdf

Alternating-time temporal logic (ATL) [Alur and Henzinger, 2002] and Coalition Logic (CL) [Pauly, 2002]. This is due to the fact that STIT is more expressive than ATL and CL (this is formally proved in [Broersen *et al.*, 2006]) so that there are STIT formulas such as $[J]\chi$ and $\neg[J]\chi$ that cannot be translated into ATL and CL.

4 A STIT extension with knowledge

This section presents an extension of the fragment *df*STIT of STIT logic presented in section 2 with standard operators for knowledge of the form $\mathtt{K}_i$, where $\mathtt{K}_i\varphi$ means "agent i knows that φ is true". This is a necessary step for the formalization of regret and relief that will be presented in section 5.

The language $\mathcal{L}_{df\mathsf{KSTIT}}$ of logic *df*KSTIT is defined by the following BNF:

$\chi ::= \bot \mid p \mid \chi \wedge \chi \mid \neg\chi$ (propositional formulas)
$\psi ::= [J]\chi \mid \psi \wedge \psi$ (see-to-it formulas)
$\varphi ::= \chi \mid \psi \mid \varphi \wedge \varphi \mid \neg\varphi \mid \langle\emptyset\rangle\psi \mid \mathtt{K}_i\varphi$
(see-to-it, "can", knowledge formulas)

where p ranges over ATM, i ranges over *AGT* and J over $2^{AGT} \setminus \emptyset$.

Definition 3. *A KSTIT-model is a tuple* $\mathcal{M} = (W, \{R_J\}_{J\subseteq AGT}, \{E_i\}_{i\in AGT}, V)$ *where:*

- $(W, \{R_J\}_{J\subseteq AGT}, V)$ *is a STIT-model;*
- *For all* $i \in$ *AGT,* E_i *is an equivalence relation.*

Truth conditions for atomic formulas and the boolean operators are again entirely standard. Truth conditions for the STIT operators $[J]$ are given in section 2. Truth conditions for knowledge operators are defined in the standard way:

$$\mathcal{M}, w \models \mathtt{K}_i\varphi \text{ iff } \mathcal{M}, v \models \varphi \text{ for all } v \in W \text{ such that } wE_iv.$$

That is, agent i knows that φ at world w in model $\mathcal{M}$ if and only if φ is true at all worlds that are indistinguishable for agent i at world w.

A formula φ is KSTIT-valid (noted $\models_{\mathsf{KSTIT}} \varphi$) iff φ is true in every world of every KSTIT-model.

Theorem 3. *The satisfiability problem of df*KSTIT *is NP-complete if* card(*AGT*) $= 1$ *and PSPACE-complete if* card(*AGT*) ≥ 2.

Sketch of proof. If *card*(*AGT*) $= 1$, we can notice that there are only three operators: $[\emptyset]$, $[\{1\}]$, and $\mathtt{K}_1$. Nevertheless, the operator $[\{1\}]$ can be removed because we force $R_{AGT} = id_W$ in our models. As no $\mathtt{K}_1$-operator can appear after a $[\emptyset]$-operator, we can prove by a selected points argument (in [Ladner, 1977], it is done for $S5$) that if a *df*KSTIT -formula is KSTIT-satisfiable, then it is in a polynomial sized model.

If *card*(*AGT*) ≥ 2, we can create a tableau method to see if a *df*KSTIT -formula φ is KSTIT-satisfiable. The idea consists in using a classical tableau algorithm for the knowledge part considering the STIT-subformulas as propositions (see tableau method for $S5_n$ in [Halpern and Moses, 1992]). At each step, we take care of STIT-subformulas by choosing non-deterministically a polynomial sized STIT-model. We can prove correctness and soundness of this tableau method. This algorithm runs using a polynomial space memory so the satisfiability problem of *df*KSTIT is NPSPACE. As NPSPACE = PSPACE (Savitch's theorem [Papadimitriou, 1994]), it is PSPACE. It is PSPACE-hard because the logic $S5_n$ is embedded into *df*KSTIT . □

5 Regret and rejoicing: a formalization

In order to provide a logical characterization of counterfactual emotions such as regret, we need to introduce a concept of agent's preference. Modal operators for desires and goals have been widely studied (see e.g. [Cohen and Levesque, 1990; Meyer *et al.*, 1999]). The disadvantage of such approaches is that they complicate the underlying logical framework. An alternative, which we adopt in this paper is to label states with atoms that capture the "goodness" of these states for an agent. Our approach supposes a binary relation of preference between worlds.

Let us introduce a special atom $good_i$ for every agent $i \in AGT$. These atoms are used to specify those worlds which are positive for an agent.

We say that χ is good for agent i if and only if, necessarily if the current state is a good state for agent i then, χ is true in that state. Formally:

$$\mathtt{GOOD}_i\chi \stackrel{\text{def}}{=} [\emptyset](good_i \rightarrow \chi).$$

Now, we are in a position to define the concept of desirable state of affairs. We say that χ is desirable for agent i if and only if, i knows that χ is something good for him:

$$\mathtt{DES}_i\chi \stackrel{\text{def}}{=} \mathtt{K}_i\mathtt{GOOD}_i\chi.$$

As the following valid formulas highlight, every operator $\mathtt{DES}_i$ satisfies the principle K of normal modal logic, and the properties of positive and negative introspection: χ is (resp. is not) desirable for i if and only if i knows this.

(4) $\quad \models_{\mathsf{KSTIT}} (\mathtt{DES}_i\chi_1 \wedge \mathtt{DES}_i(\chi_1 \rightarrow \chi_2)) \rightarrow \mathtt{DES}_i\chi_2$

(5) $\quad \models_{\mathsf{KSTIT}} \mathtt{DES}_i\chi \leftrightarrow \mathtt{K}_i\mathtt{DES}_i\chi$

(6) $\quad \models_{\mathsf{KSTIT}} \neg\mathtt{DES}_i\chi \leftrightarrow \mathtt{K}_i\neg\mathtt{DES}_i\chi$

We have now all necessary and sufficient ingredients to define the cognitive structure of regret and to specify its counterfactual dimension. Such a dimension has been widely studied in the psychological literature on regret (see [Kahneman and Miller, 1986; Kahneman, 1995] for instance). Our aim here is to capture it formally. We say that an agent i regrets for χ if and only if $\neg\chi$ is desirable for i and i knows that it could have prevented χ to be true now. Formally:

$$\mathtt{REGRET}_i\chi \stackrel{\text{def}}{=} \mathtt{DES}_i\neg\chi \wedge \mathtt{K}_i\mathtt{CHP}_i\chi.$$

The following example is given in order to better clarify this definition.

EXAMPLE. Consider the popular two-person hand game "Rock-paper-scissors" (Roshambo). Each of the two players $AGT = \{1, 2\}$ has three available actions: play $rock$, play $paper$, play $scissors$. The goal of each player is to select an action which defeats that of the opponent. Combinations of actions are resolved as follows: rock wins against scissors, paper wins against rock; scissors wins against paper. If both players choose the same action, they both lose. The scenario is represented in the STIT model in Fig. 2. It is supposed

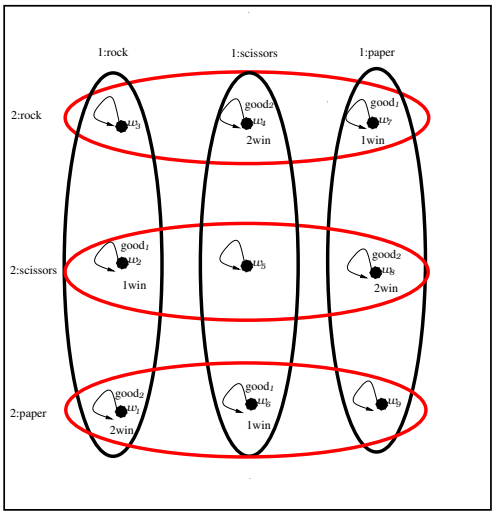

Figure 2: Again, vertical circles represent the actions that player 1 can choose, whereas horizontal circles represent the actions that player 2 can choose. For the sake of simplicity, we suppose that players 1 and 2 do not have uncertainty: everywhere in the model players 1 and 2 only consider possible the world in which they are (reflexive arrows represent indistinguishability relations for the two players).

winning is something good for each agent and each agent has the desire to win the game: $\mathtt{GOOD}_1\, 1Win$, $\mathtt{GOOD}_2\, 2Win$, $\mathtt{DES}_1\, 1Win$ and $\mathtt{DES}_2\, 2Win$ are true at worlds w_1-w_9. Suppose world w_1 is the actual world in which 1 plays $rock$ and 2 plays $paper$. In this world 1 loses the game ($\neg 1Win$), and 1 knows that (by playing $scissors$) it could have prevented $\neg 1Win$ to be true (i.e. $\mathtt{K}_1\mathtt{CHP}_1 \neg 1Win$ is true at w_1). It follows that at w_1 player 1 regrets for having lost the game, that is, $\mathtt{REGRET}_1 \neg 1Win$ is true at w_1.

As the following validity highlights, regret implies the frustration of an agent's desire:

(7) $\quad \models_{\mathsf{KSTIT}} \mathtt{REGRET}_i\chi \rightarrow (\mathtt{K}_i\chi \wedge \mathtt{DES}_i\neg\chi)$

More precisely, if agent i regrets for χ then, i knows that χ holds and $\neg\chi$ is something desirable for i (in this sense i feels frustrated for not having achieved $\neg\chi$). Moreover, regret satisfies the properties of positive and negative introspection:

(8) $\quad \models_{\mathsf{KSTIT}} \mathtt{REGRET}_i\chi \leftrightarrow \mathtt{K}_i\mathtt{REGRET}_i\chi$

(9) $\quad \models_{\mathsf{KSTIT}} \neg\mathtt{REGRET}_i\chi \leftrightarrow \mathtt{K}_i\neg\mathtt{REGRET}_i\chi$

That is, i regrets (resp. does not regret) for χ if and only if i knows this.

As emphasized by some psychological theories of counterfactual emotions (see [Zeelenberg *et al.*, 1996] for instance), the positive counterpart of regret is rejoicing: while regret has a *negative valence* (i.e. it is associated with the frustration of an agent's desire), rejoicing has a *positive valence* (i.e. it is associated with the satisfaction of an agent's desire). According to these theories, a person experiences regret when believing that the foregone outcome would have been better if she did a different action, whilst she rejoices when believing that the foregone outcome would have been worse if she did a different action. More precisely, an agent i rejoices for χ if and only if, χ is desirable for i and, i knows that it could have prevented χ to be true now by doing a different action. Formally:

$$\mathtt{REJOICE}_i\chi \stackrel{\text{def}}{=} \mathtt{DES}_i\chi \wedge \mathtt{K}_i\mathtt{CHP}_i\chi.$$

EXAMPLE. Consider again the game "Rock-paper-scissors" represented by the STIT-model in Fig. 2. Suppose world w_2 is the actual world in which player 1 plays $rock$ and player 2 plays $scissors$. In this world player 1 is the winner ($1Win$) and it knows that (by playing $paper$ or $scissors$) it could have prevented $1Win$ to be true (i.e. $\mathtt{K}_1\mathtt{CHP}_1\, 1Win$ is true at w_2). Since $\mathtt{DES}_1\, 1Win$ holds at w_2, it follows that at w_2 player 1 rejoices for having won the game, that is, $\mathtt{REJOICE}_1\, 1Win$ is true at w_2.

The following validity highlights that rejoicing implies desire satisfaction:

(10) $\quad \models_{\mathsf{KSTIT}} \mathtt{REJOICE}_i\chi \rightarrow (\mathtt{K}_i\chi \wedge \mathtt{DES}_i\chi)$

More precisely, if agent i rejoices for χ then, i knows that χ and χ is something desirable for i (in this sense i feels satisfied for having achieved χ). As regret, rejoicing satisfies the properties of positive and negative introspection:

(11) $\quad \models_{\mathsf{KSTIT}} \mathtt{REJOICE}_i\chi \leftrightarrow \mathtt{K}_i\mathtt{REJOICE}_i\chi$

(12) $\quad \models_{\mathsf{KSTIT}} \neg\mathtt{REJOICE}_i\chi \leftrightarrow \mathtt{K}_i\neg\mathtt{REJOICE}_i\chi$

6 Related works

As emphasized in the introduction, there are other researchers who have exploited logical methods in order to build formal models of emotions and affective agents.

In [Meyer, 2006; Steunebrink *et al.*, 2007] a logical approach to emotions based on the modal logical framework KARO [Meyer *et al.*, 1999] is proposed. KARO is a framework based on a blend of dynamic logic with epistemic logic, enriched with modal operators for motivational attitudes such as desires and goals. In Meyer et al.'s approach each instance of emotion is represented with a special predicate, or fluent, in the jargon of reasoning about action and change, to indicate that these predicates change over time. For every fluent a set of effects of the corresponding emotion on the agent's planning strategies and decision processes are specified, as well as the preconditions for triggering the emotion. Although Meyer et al. provide a very detailed formal analysis of emotions, they do not take into account counterfactual emotions. This is also due to some intrinsic limitations of the KARO framework in expressing counterfactual reasoning and statements of the form "agent i *could have prevented* χ to be true" which are fundamental constituents of this kind of emotions. Indeed, standard dynamic logic on the top of which KARO is built, is not suited to express such statements. On the contrary our STIT-based approach overcomes this limitation.

In [El-Nasr *et al.*, 2000] a formal approach to emotions based on fuzzy logic is proposed. The main contribution of this work is a quantification of emotional intensity based on appraisal variables like desirability of an event and its likelihood. For example, following [Ortony *et al.*, 1988], in

FLAME the variables affecting the intensity of hope with respect to the occurrence of a certain event are the degree to which the expected event is desirable, and the likelihood of the event. However, in FLAME only basic emotions like joy, sadness, fear and hope are considered and there is no formal analysis of counterfactual emotions as the ones analyzed in our work. Indeed, the formal language exploited in [El-Nasr *et al.*, 2000] is not sufficiently expressive to model counterfactual reasoning about agents' choices and actions.

7 Conclusion

Directions for our future research are manifold. An analysis of intensity of regret and rejoicing was beyond the objectives of the present work. However, we intend to investigate this issue in the future in order to complement our qualitative analysis of affective phenomena with a quantitative analysis. Moreover, we have focused in this paper on the logical characterization of two counterfactual emotions: regret and rejoicing. We intend to extend our analysis in the future by studying the counterfactual dimension of "moral" emotions such as guilt and shame. Several psychologists (see [Lazarus, 1991] for instance) have stressed that guilt involves the conviction of having injured someone or of having violated some norm or imperative, and the belief that this *could have been avoided*.

It has been proved in [Herzig and Schwarzentruber, 2008] that the logic STIT with agents and groups proposed by Horty [Horty, 2001] is not only undecidable but also not axiomatizable, i.e. there is no axiomatization Ax such that that for every formula φ in $\mathcal{L}_{\mathsf{STIT}}$, $\vdash_{Ax} \varphi$ if and only if $\models_{\mathsf{STIT}} \varphi$. In this work, we have dealt with the first problem by presenting a decidable fragment of Horty's logic which is sufficiently expressive for our purpose of formalizing counterfactual emotions. It is still an open question whether we can find a finite and complete axiomatization for our fragment dfSTIT of STIT with agents and groups. Our future work will also be devoted to solve this problem in order to come up with a complete and decidable fragment of STIT with agents and groups.

References

[Alur and Henzinger, 2002] R. Alur and T. Henzinger. Alternating-time temporal logic. *J. of the ACM*, 49:672–713, 2002.

[Belnap *et al.*, 2001] N. Belnap, M. Perloff, and M. Xu. *Facing the future: agents and choices in our indeterminist world*. Oxford Univ. Press, 2001.

[Broersen *et al.*, 2006] J. Broersen, A. Herzig, and N. Troquard. Embedding Alternating-time temporal logic in strategic STIT logic of agency. *J. of Logic and Computation*, 16(5):559–578, 2006.

[Cohen and Levesque, 1990] P. R. Cohen and H. J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42(2–3):213–261, 1990.

[El-Nasr *et al.*, 2000] M. S. El-Nasr, J. Yen, and T. R. Ioerger. FLAME: Fuzzy logic adaptive model of emotions. *Autonomous Agents and Multi-Agent Systems*, 3(3):219–257, 2000.

[Gabbay *et al.*, 2003] D. M. Gabbay, A. Kurucz, F. Wolter, and M. Zakharyaschev. *Many-dimensional modal logics: theory and applications*. Elsevier, 2003.

[Halpern and Moses, 1992] J. Y. Halpern and Y. Moses. A guide to completeness and complexity for modal logics of knowledge and belief. *Artificial Intelligence*, 54(2):319–379, 1992.

[Herzig and Schwarzentruber, 2008] A. Herzig and F. Schwarzentruber. Properties of logics of individual and group agency. In *Proc. of Advances in Modal Logic 2008*, pages 133–149. College Publ., 2008.

[Horty, 2001] J. F. Horty. *Agency and Deontic Logic*. Oxford Univ. Press, 2001.

[Kahneman and Miller, 1986] D. Kahneman and D. T. Miller. Norm theory: comparing reality to its alternatives. *Psychological Review*, 93(2):136–153, 1986.

[Kahneman, 1995] D. Kahneman. Varieties of counterfactual thinking. In N. J. Roese and J. M. Olson, editors, *What might have been: the social psychology of counterfactual thinking*. Erlbaum, 1995.

[Ladner, 1977] R. E. Ladner. The computational complexity of provability in systems of modal propositional logic. *SIAM Journal on Computing*, 6(3):467–480, 1977.

[Lazarus, 1991] R. S. Lazarus. *Emotion and adaptation*. Oxford Univ. Press, 1991.

[Meyer *et al.*, 1999] J. J. Ch. Meyer, W. van der Hoek, and B. van Linder. A logical approach to the dynamics of commitments. *Artificial Intelligence*, 113(1-2):1–40, 1999.

[Meyer, 2006] J.-J. Ch. Meyer. Reasoning about emotional agents. *Int. J. of Intelligent Systems*, 21(6):601–619, 2006.

[Ortony *et al.*, 1988] A. Ortony, G. L. Clore, and A. Collins. *The cognitive structure of emotions*. Cambridge Univ. Press, 1988.

[Papadimitriou, 1994] C. H. Papadimitriou. *Computational complexity*. Addison Wesley, 1994.

[Pauly, 2002] M. Pauly. A modal logic for coalitional power in games. *J. of Logic and Computation*, 12(1):149–166, 2002.

[Picard, 1997] R. W. Picard. *Affective Computing*. MIT Press, 1997.

[Roese *et al.*, 2005] N. J. Roese, L. J. Sanna, and A. D. Galinsky. The mechanics of imagination: automaticity and control in counterfactual thinking. In R. R. Hassin, J. S. Uleman, and J. A. Bargh, editors, *The new unconscious*. Oxford Univ. Press, 2005.

[Steunebrink *et al.*, 2007] B. R. Steunebrink, M. Dastani, and J.-J. Ch. Meyer. A logic of emotions for intelligent agents. In *Proc. of AAAI'07*, pages 142–147. AAAI Press, 2007.

[Zeelenberg *et al.*, 1996] M. Zeelenberg, J. Beattie, J. van der Pligt, and N. K. de Vries. Consequences of regret aversion: effects of expected feedback on risky decision making. *Organizational behavior and human decision processes*, 65(2):148–158, 1996.