A Logic for Reasoning about Counterfactual Emotions[☆]

Emiliano Lorini[a,*], François Schwarzentruber[a]

[a] *Institut de Recherche en Informatique de Toulouse (IRIT),*
118 route de Narbonne, 31062 Toulouse Cedex, France,
Tel: (+33)0561556447, Fax: (+33)561556258

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. The prototypical counterfactual emotion is regret. In order to meet this objective, we exploit the well-known STIT logic [9, 30, 31]. 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 and not finitely axiomatizable. In this work we study a decidable and finitely axiomatizable fragment of STIT with agents and groups which is sufficiently expressive for our purpose of formalizing counterfactual emotions. We call dfSTIT our STIT fragment. After having extended dfSTIT with knowledge modalities, in the second part of article, we exploit it in order to formalize four types of counterfactual emotions: regret, rejoicing, disappointment, and elation. At the end of the article we presents an application of our formalization of counterfactual emotions to a concrete example.

Keywords: modal logic, emotions, STIT

[☆] This work is an extended and improved version of the article "A Logic for Reasoning about Counterfactual Emotions" appeared in the Proceedings of the Twenty-first International Joint Conference on Artificial Intelligence (IJCAI'09), pp. 867-872. (Work supported by the project ANR-08-CORD-005 CECIL.)

[*] Corresponding author

Email addresses: lorini@irit.fr (Emiliano Lorini), schwarze@irit.fr (François Schwarzentruber)

Contents

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 [8]. In this perspective, a challenge for AI is to build artificial agents which are capable of: reasoning about emotions, showing their affective states and personalities, ascribing emotions to humans, predicting the effects of their actions on emotions of humans, and adapting their behaviors accordingly. 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 [44] has emerged in the last few years.

Recently, some researchers have been interested in developing logical frameworks for the formal analysis of emotions (see [39, 40, 58, 20] 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 of reasoning about and of displaying some kind of emotions can indeed benefit from the accuracy of logical methods. These logical frameworks for the specification of emotions are based on the so-called BDI logics (see e.g. [17, 41]). BDI logics allow to model agents' mental states such as beliefs, desires, intentions, ideals, values, etc. which are the cognitive constituents of emotions.

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." [33, 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 [49].

The aim of our work is to advance the state of the art on computational modeling of affective phenomena 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 [9, 30]. 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 [12] for instance). It is a very suitable formalism to reason about counterfactual choices of agents and of groups. Unfortunately, the version of STIT with agents and groups proposed by Horty [30] has been

recently proved to be undecidable and not finitely axiomatizable [29]. In this work we study a decidable and finitely axiomatizable 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 one of the most influential research approach to emotions: appraisal theory. We provide a general overview of existing models of emotions proposed in this area by devoting special attention to appraisal models of counterfactual emotions. We discuss how counterfactual emotions such as regret and disappointment are defined in these models.

Section 3 is the first step in developing a representation language for the formalization of counterfactual emotions. We introduce a fragment of the version of STIT logic with agents and groups proposed by Horty [30]. We call dfSTIT our STIT fragment. Differently from Horty's logic, we prove that our fragment is decidable and finitely axiomatizable.

In Section 4, we exploit the STIT fragment dfSTIT in order to formalize counterfactual statements of the form "group J (or agent i) *could have prevented* χ to be true". These statements are indeed basic constituents of counterfactual emotions and will be fundamental in the formalization of counterfactual emotions given in Section 6.

In Section 5, we extend the STIT fragment dfSTIT studied in Section 3 with knowledge operators. This is a necessary step in order to capture the subjective dimension of the affective phenomena we intend to analyze in our work. We provide decidability results and a complete axiomatization for our epistemic extension of dfSTIT. We decided to present first the STIT fragment without knowledge and then the extension with knowledge operators rather than to present a direct version of a STIT fragment with knowledge operators for several reasons. The first one is because the STIT fragment without knowledge studied in Section 3 is interesting in itself since it already allows to express counterfactual statements which are an interesting component of counterfactual emotions. The second one is because the proof of decidability and the proof of completeness of the STIT fragment with knowledge become much simpler after having studied the STIT fragment without knowledge.

In Section 6, the logical framework of Section 5, is finally applied to the formalization of counterfactual emotions. We provide a formalization of four types of counterfactual emotions: *regret* and its positive counterpart *rejoicing*, *disappointment* and its positive counterpart *elation*. The formal definitions of these four emotions will be based on the psychological models of counterfactual emotions discussed in Section 2. Section 7 presents an application of our logical formalization of counterfactual emotions to a concrete example. Before concluding we discuss in Section 8 some related works in the area of logical modeling of emotions and affective agents.

Proofs of the main theorems are collected in the annex at the end of the article.

2. Emotion theories

Our general objective in this work is to provide a formal model of emotions which can be used as an abstract specification for the design of artificial agents interacting with humans. To ensure the accuracy of a such a formal model, it is important to consider how emotions have been defined in the psychological literature. Indeed, in order to build artificial agents with the capability of recognizing the emotions of a human user, of anticipating the emotional effects of their actions on the human, of affecting the user's emotions by the performance of actions directed to his emotions (e.g. actions aimed at reducing the human's stress due to his negative emotions, actions aimed at inducing positive emotions in the human), we must endow such agents with an adequate model of human emotions.

There exist several theoretical approaches to emotions in psychology. We here consider one of the most influential called appraisal theory (see [53] for a broad introduction to the developments in appraisal theory).

In Section 2.1, we provide a general introduction to appraisal theory by reviewing some of the most popular models proposed in this area. Then, in Section 2.2, we will focus on appraisal models of counterfactual emotions and of regret in particular. This section will provide the conceptual basis for the formalization of counterfactual emotions proposed in Section 6.

2.1. Appraisal models of emotions

Appraisal theory has emphasized the strong relationship between emotion and cognition, by stating that each emotion can be related to specific patterns of evaluations and interpretations of events, situations or objects (appraisal patterns) based on a number of dimensions or criteria called *appraisal variables* (e.g. goal relevance, desirability, likelihood, causal attribution). Appraisal variables are directly related to the mental attitudes of the individual (e.g. beliefs, predictions, desires, goals, intentions). For instance, when prospecting the possibility of winning a lottery and considering 'I win the lottery' as a desirable event, an agent might feel an intense hope. When prospecting the possibility of catching the H1N1 flu and considering 'I catch the H1N1 flu' as an undesirable event, an agent might feel an intense fear.

It is worth noting that most appraisal models of emotions assume that explicit evaluations based on evaluative beliefs (i.e. the belief that a certain event is good or bad, pleasant or unpleasant, dangerous or frustrating) are a necessary constituent of emotional experience. On the other hand, there are some appraisal models mostly promoted by philosophers [55, 26] in which emotions are reduced to specific combinations of beliefs and desires, and in which the link between cognition and emotion is not necessarily mediated by evaluative beliefs. Reisenzein [47] calls *cognitive-evaluative* the former and *cognitive-motivational* the latter kind of models. For example, according to cognitive-motivational models of emotions, a person's happiness about a certain fact χ

can be reduced to the person's belief that χ obtains and the person's desire that χ obtains. On the contrary, according to cognitive-evaluative models, a person feels happy about a certain fact χ if she believes that χ obtains and she evaluates χ to be good (desirable) for her. In the present work, we stay closer to cognitive-evaluative models. In fact, we suppose that an agent's positive (resp. negative) emotion requires the agent's (evaluative) belief that a certain event, situation or object is good (resp. bad) for her. For example, according to the formalization of rejoicing we will propose in Section 6, if an agent rejoices for a certain event χ then he believes that χ is something good for him.

Now let us provide a more comprehensive overview of the research in appraisal theory by briefly discussing some of the most important models of emotions in this area.

Lazarus's model. Lazarus [56, 36] distinguishes *primary appraisal* from *secondary appraisal*. These two kinds of appraisal are not sequential: they can be executed in any order. During primary appraisal a person assesses the relevance and congruence of an event with respect to her desires and goals, that is, she evaluates whether an event helps or threatens the achievement of her goals and/or the satisfaction of her desires. During secondary appraisal, the person evaluates available capabilities and resources to cope with a certain event. For instance, after feeling an intense fear because of the belief that the undesirable event 'I catch the H1N1 flu' will probably occur, an agent might consider whether to get vaccinated against the H1N1 flu in order to reduce his risks. [1]

Scherer's model. In Scherer's model [52], the appraisal process is conceived as a sequence of processing levels of a given stimulus (Stimulus Evaluation Checks) which underlies the assessment of the significance of the stimulus for an individual. In particular, according to Scherer's model, an event is sequentially evaluated through the following four steps: *relevance detection* (i.e. whether the event is novel and important with respect to the momentary goals of the individual), *implication assessment* (i.e. whether the event will further or endanger the individual's attainment of his goals), *coping potential determination* (i.e. whether the individual can cope with the expected consequences of the event), *normative significance evaluation* (i.e. whether the event is significant with respect to the individual's ideals and values). Contrarily to Lazarus's model, in Scherer's model the different stages of the appraisal process are sequential.

[1] Lazarus also distinguishes *appraisal* from *coping*. Coping is the process of dealing with emotion, either externally by forming an intention to act in the world or internally by changing the agent's interpretation of the situation (e.g. by changing beliefs, shifting attention, shifting responsibility). Indeed, to discharge a certain emotion, an agent has to modify those mental attitudes that sustain her emotional state.

Roseman's model. Roseman's appraisal model [50, 51] distinguishes seven appraisal dimensions that were found to differentiate a large number of emotions: unexpectedness, situational state, motivational state, probability, control potential, problem source and agency. In Roseman's model, *unexpectedness* refers to whether an event is expected or unexpected by a person, and *situational state* refers to whether the event is wanted or unwanted by the person. *Motivational state* refers to whether the person assesses that the event has positive or negative implications on her goals, and *probability* refers to whether the person thinks that the occurrence of the event is merely possible/probable or is definite. *Control potential* refers to whether the person thinks she can cope with the event, and *problem source* refers to whether the event is unwanted by the person because she thinks that it blocks attainment of her goals or because of some inherent characteristic. Finally, *agency* refers to the person's evaluation of the cause of the event (i.e. whether it was caused by the self, by someone else, or by circumstances beyond anyone's control).

OCC model. According to Ortony, Clore and Collins's model (OCC model) [42], emotion arises from valenced (a dimension ranging from positive to negative) reactions to consequences of events, actions of agents, or aspects of objects. In the OCC model, the *consequence of an event* can be appraised as pleased or displeased. A person can be focused either on the consequences of an event for the self or on the consequences for another person. For example, if the person is focused on the self, she will feel hope when the consequences of the event are desirable for her, and she will feel fear when the consequences of the event are undesirable for her. In the OCC model the *action of an agent* can be approved or disapproved. A person can be focused either on her actions or on the actions of another agent. For example, if the person is focused on another agent's action, she will fell admiration when she approves this action, and she will feel reproach when she disapproves it. Finally, the *aspects of an object* can be liked or disliked. If a person likes the aspects of an object she will feel love. She will feel hate if she dislikes them.

Frijda's model. In Frijda's model [22] appraisal is defined as a sequence of evaluation steps determining the characteristics of a given stimulus: causes and consequences of the event, relevance and congruence with respect to current goals and interests, coping possibilities, and urgency. However, this model considers not only the appraisal patterns of different emotion types, but also the action tendencies induced by emotions. According to Frijda, actions tendencies are [22, pp. 75] "...states of readiness to achieve or maintain a given kind of relationship with the environment. They can be conceived of as plans or programs to achieve such ends, which are put in a state of readiness." For example, the action tendency associated to fear is escape. According to Frijda, after a stimulus has been evaluated according to the previous appraisal dimensions, an action tendency is then created that induces physiological changes, and finally an action is

selected and executed. [2]

2.2. Appraisal models of counterfactual emotions

Regret is the prototypical counterfactual emotion which has been widely investigated in psychology and in the field of decision theory in economics. Most authors (see, e.g., [37, 59, 48, 33, 32, 71]) agree in considering regret as "...a negative, cognitively determined emotion that we experience when realizing or imagining that our present situation would have been better, had we acted differently" [71, pp. 255]. In other words, regret stems from the comparison between the actual outcome deriving from a given choice and a *counterfactual* better outcome that might have been had one chosen a different action. Such a definition highlights the strong connection between decision-making and regret: broadly speaking, regret can be conceived as an emotion originating from a person's perception of her 'bad decision'. From this perspective, a sense responsibility for a bad outcome has been often considered a specific characteristic of the phenomenology of regret, that is, the more a decision maker perceives himself to be responsible for a negative outcome, the more regret he experiences [23]. [3]

This aspect clearly distinguishes regret from *disappointment*. According to some economists [38] and to some psychologists [19, 70], disappointment too is part of the family of counterfactual emotions. But, although regret and disappointment both originate from the comparison between the actual outcome and a counterfactual outcome that might have occurred, disappointment follows from the comparison between the actual outcome and a counterfactual better outcome that might have been had a different state of the world occurred. That is, while regret is related to a sense of responsibility and involves an internal attribution of the cause of a bad outcome (i.e. when feeling regret a person considers her own choices to be the cause of a bad outcome), disappointment is related to external attribution (i.e. when feeling disappointed a person considers external events to be the cause of a bad outcome).

The positive counterparts of regret and disappointment have also been considered in the psychological literature (see, e.g., [68, 69]). The former is called *rejoicing*, while the latter is called *elation*. Broadly speaking, one can say that while rejoicing stems from the comparison between the actual outcome deriving from a given choice and a counterfactual worst outcome that might have been had one chosen a different action,

[2] According to Lazarus [36], there is an important difference between action tendencies and coping strategies. While the former are innately programmed unconscious reflexes and routines, the latter are the product of a conscious deliberation process.

[3] Compared to the large number of authors relating regret with responsibility for a bad outcome, there are very few authors who separate the two concepts. According to [60, 57] for instance, one can be regretful also for events that are partially or totally beyond one's own control or for choices for which there was no alternative. However, we here adopt the definition of regret shared by the majority of authors emphasizing the link between regret and responsibility.

elation follows from the comparison between the actual outcome and a counterfactual worst outcome that might have been had a different state of the world occurred.

The next Section 3 is our first step in the development of a formal representation language for modeling counterfactual emotions. We study a decidable fragment of STIT logic with groups of agents proposed by Horty [30], and we give an axiomatization of it. STIT is indeed a suitable framework for expressing counterfactual statements about actions and choices of the form "group J (or agent i) could have prevented a certain state of affairs χ to be true now". Such statements are fundamental building blocks for an analysis of counterfactual emotions.

3. A decidable and finitely axiomatizable fragment of STIT

The logic STIT ("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 [9], the language of STIT with individuals but without groups is studied: Belnap et col. introduce constructions of the form $[istit : \varphi]$ to be read "agent i sees to it that φ" or "agent i brings it about that φ". They give a complete axiomatization of STIT without groups and prove that the logic is decidable. The extension of STIT with groups has been proposed by Horty in [30]: it deals with constructions of the form $[Jstit : \varphi]$ to be read "group J sees to it that φ". For notational convenience, we write here $[J]\varphi$ instead of $[Jstit : \varphi]$. Unfortunately, in [29] it has been proved to be undecidable and unaxiomatizable (with a finite number of axioms schemas, necessitation rules and modus ponens).

We here introduce a decidable and axiomatizable fragment of STIT with agents and groups called dfSTIT which is sufficiently expressive to formalize counterfactual emotions. First, in Subsection 3.1, we recall the syntax of STIT and define the syntactic fragment dfSTIT. In Subsection 3.2, we recall definition of models of the logic STIT. Then, in Subsection 3.3, we recall the logic NCL [5, 61, 54]. The logic NCL shares the same syntax with STIT and its semantics looks like the semantics of STIT. Nevertheless, NCL is axiomatizable. The logic NCL will be a key point to prove the decidability of the STIT fragment dfSTIT and to give a complete axiomatization of dfSTIT (Subsection 3.4).

3.1. 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 finite set of agents. The language $\mathcal{L}_{\mathsf{STIT}}$ of the logic STIT with agents and groups proposed by Horty [30] 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 been 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*. In Belnap et col.'s STIT, an agent i's action is described in terms of the result that agent i brings about by his acting. For example, i's action of killing another agent j is described by the fact that i sees to it that j is dead. In Horty's STIT with agents and groups we can make a distinction between *individual actions* of agents and *joint actions* of groups. The joint action of a group J is described in terms of the result that the agents in J bring about by acting together.

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 operator $[\emptyset]$ is exactly the *historic necessity operator* already present in the individual STIT logic [9]. 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 no matter what the other agents in $AGT\setminus J$ do".

Here we are interested in a fragment of $\mathcal{L}_{\mathsf{STIT}}$ we call $\mathcal{L}_{df\mathsf{STIT}}$. It 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)

$\varphi ::= \chi \mid \psi \mid \varphi\wedge\varphi \mid \neg\varphi \mid \langle\emptyset\rangle\psi$ (see-to-it and "can" formulas)

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

$\mathcal{L}_{df\mathsf{STIT}}$ is a syntactic restriction of $\mathcal{L}_{\mathsf{STIT}}$. We have $\mathcal{L}_{df\mathsf{STIT}}\subseteq\mathcal{L}_{\mathsf{STIT}}$ but $\mathcal{L}_{\mathsf{STIT}}\not\subseteq\mathcal{L}_{df\mathsf{STIT}}$. For instance, $[1][\{1,2\}]p$ is in $\mathcal{L}_{\mathsf{STIT}}$ but is not in $\mathcal{L}_{df\mathsf{STIT}}$.

3.2. Models

We give two semantics of STIT. It is proved in [29] that these two semantics are equivalent. The first one corresponds to the original semantics of STIT with agents and groups proposed by Horty [30]. The other one is based on the product logic $S5^n$ [24] and will be used in Section 3.4 in order to characterize the satisfiability of a dfSTIT -formula. Let us first recall the original semantics of STIT.

Definition 1 (STIT-model).
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 $J \in 2^{AGT}$, $R_J(w) = \{v \in W \mid wR_Jv\}$.

$R_J(w)$ is the set of outcomes that is forced by group J's action at world w, that is, at world w group J forces the world to be in some state of $R_J(w)$. Hence, if $v \in R_J(w)$ then v is an outcome that is *admitted* by group J's action at world w.

Note that if v is admitted by group J's action at world w (i.e. $v \in R_J(w)$) then this means that, given what the agents in J have chosen at w, there exists a joint action of the agents in $AGT \setminus J$ such that, if the agents in $AGT \setminus J$ did choose this joint action, v would be the actual outcome of the joint action of all agents.

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 the set of outcomes that is forced by J's action is a subset of the set of possible outcomes. Constraint 2 just says that the set of outcomes that is forced by J's joint action at a world w is equal to the pointwise intersection of the sets of outcomes that are forced by the individual actions of 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 is forced by agent 1's action at w_1, and the set of outcomes that is forced by agent 2's action at w_2,..., and the set of outcomes that is forced by agent n's action 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 that is forced by the joint action of all agents is a singleton that is to say we have $R_{AGT}(w) = \{w\}$ for all $w \in W$.

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

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

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

Definition 2 (product STIT-model).
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, $V : W \to 2^{ATM}$.

Truth conditions for atomic formulas and the boolean operators are also entirely standard. The truth conditions for the modal operators $[J]$ in product STIT-models are:

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

Now let us just recall the notion of validity and satisfiability in STIT. As there is an equivalence between a STIT-model and a product STIT-model as proved by [29], we can define those notions either with respect to STIT-models or with respect to STIT-models. A formula φ is STIT-valid (noted $\models_{\mathsf{STIT}} \varphi$) if and only if φ is true in every world of every STIT-model. Or, equivalently, a formula φ is STIT-valid if and only if φ is true in every world of every product STIT-model. A formula φ is STIT-satisfiable if and only if there exists a STIT-model $\mathcal{M} = (W, \{R_J\}_{J \subseteq AGT}, V)$ and a point $w \in W$ such that $\mathcal{M}, w \models \varphi$. Or, equivalently, a formula φ is STIT-satisfiable if and only if there exists a product STIT-model $\mathcal{M} = (W, V)$ and a point $(w_1, \ldots, w_n) \in W$ such that $\mathcal{M}, (w_1, \ldots, w_n) \models \varphi$.

3.3. The NCL logic

Unfortunately, STIT is not axiomatizable. Nevertheless, there exists an axiomatizable logic which is very close to STIT. This logic is the fragment of Normal Coalition Logic (NCL) [5, 61, 54, 13] in which we do not deal with the *next* operator. Normal Coalition Logic was originally proposed in order to embed non-normal Coalition Logic CL [43] into a *normal* modal logic. This embedding uses a general technique developed by [25]. The reader can find more details about this specific embedding in [5, 61, 13].Just as CL, NCL is also axiomatizable and decidable.

Below we show that the fragment of Normal Coalition Logic without time axiomatizes the set of validities in the fragment $\mathcal{L}_{df\mathsf{STIT}}$ of STIT. Moreover, we prove our central characterization theorem of STIT-satisfiable formula of the fragment $\mathcal{L}_{df\mathsf{STIT}}$ by using the Normal Coalition Logic without time. In rest of the paper, when we write NCL we refer to the fragment of Normal Coalition Logic with the operators of group action $[J]$ and without the *next* operator.

3.3.1. Definition

We start by giving the definition of the logic NCL. Concerning the syntax, as here we do not deal with the *next* operator, the language of NCL-formulas is the same as the language of STIT-formulas, that is to say, $\mathcal{L}_{\mathsf{NCL}} = \mathcal{L}_{\mathsf{STIT}}$. Concerning the semantics, here is the definition of a NCL-model:

Definition 3 (NCL-model).
A NCL-model is a tuple $\mathcal{M} = (W, R, V)$ where:

- W is a nonempty set of worlds or states;
- R is a collection of equivalence relations R_J (one for every coalition $J \subseteq AGT$) such that:
  1. $R_{J_1 \cup J_2} \subseteq R_{J_1} \cap R_{J_2}$;
  2. $R_\emptyset \subseteq R_J \circ R_{AGT \setminus J}$;
  3. $R_{AGT} = Id_W$.
- $V : W \to 2^{ATM}$ is a valuation function.

As in Definition 1, $R_J(w)$ represents the set of outcomes that is forced by group J's action at world w, and if $v \in R_J(w)$ then v is an outcome that is *admitted* by group J's action at world w. Hence, Constraint 1 says that if v is admitted by group $J_1 \cup J_2$'s action at w, then v is admitted by group J_1's action and by group J_2's action at w. Constraint 2 is close to the assumption of independence of agents of STIT logic. According to Constraint 2, if v is a possible outcome at w then, there exists a world u such that u is admitted by group J's action at w and v is admitted by group $AGT \setminus J$'s action at u. Constraint 3 expresses an assumption of determinism.

As usual truth conditions for atomic formulas and the boolean operators are entirely standard and the truth conditions of the operators $[J]$ are given in a traditional way by:

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

In the same way, we introduce notions of validity and satisfiability in NCL. A formula φ is NCL-valid (noted $\models_{\mathsf{NCL}} \varphi$) if and only if φ is true in every world of every NCL-model. A formula φ is NCL-satisfiable if and only if there exists a NCL-model $\mathcal{M} = (W, R, V)$ and a point $w \in W$ such that $\mathcal{M}, w \models \varphi$.

3.3.2. Axiomatization of NCL

Constraints 1, 2, 3 presented in the Definition 3 above directly correspond to Sahlqvist axiom schemas [10]. For instance Constraint 2 ($R_\emptyset \subseteq R_J \circ R_{AGT \setminus J}$) corresponds to the axiom schema $\langle\emptyset\rangle\varphi \to \langle J\rangle\langle AGT \setminus J\rangle\varphi$. This is the reason why NCL logic is axiomatizable unlike STIT logic. The following Theorem 1, which has been proved by [13], sums up this fact.

Theorem 1. *The logic NCL is complete with respect to the following axiomatization:*

(ProTau)	*all tautologies of propositional calculus*
S5($[J]$)	*all S5-theorems, for every* $[J]$
(Mon)	$[J_1]\varphi \vee [J_2]\varphi \to [J_1 \cup J_2]\varphi$

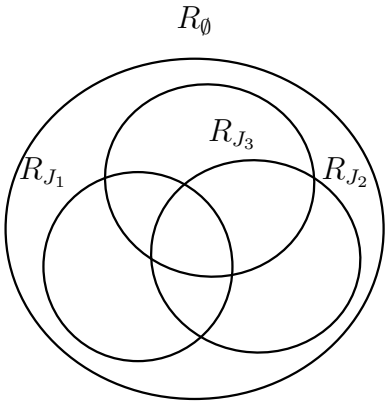

Figure 1: Independence of agents in NCL

Elim(∅) $\langle\emptyset\rangle\varphi \to \langle J\rangle\langle AGT \setminus J\rangle\varphi$

Triv(AGT) $\varphi \to [AGT]\varphi$

plus modus ponens and necessitation for all $[J]$.

As NCL is axiomatizable, we can introduce the symbol $\vdash_{\mathsf{NCL}}$ to deal with proofs. We write $\vdash_{\mathsf{NCL}} \varphi$ to say that φ is a theorem of the axiomatization given in Theorem 1.

3.3.3. Link between STIT and NCL

In the case of individual STIT logic, i.e. when the STIT language only has operator $[\emptyset]$ and operators $[i]$ with $i \in AGT$, the notion of satisfiability in STIT and the notion of satisfiability in NCL are equivalent [5, Theorem 14]. When we consider group STIT logic with operators of group action $[J]$ (with $J \subseteq AGT$), the two notions are different. The following Proposition 1 highlights the relationship between satisfiability in group STIT logic and satisfiability in NCL.

Proposition 1. *Let φ be a formula of $\mathcal{L}_{STIT}$.*

- *If* $\mathrm{card}(AGT) \leq 2$*: φ is STIT-satisfiable iff φ is NCL-satisfiable;*
- *If* $\mathrm{card}(AGT) \geq 3$*: if φ is STIT-satisfiable then φ is NCL-satisfiable. (the converse is false: there exists φ such that φ is NCL-satisfiable and $\neg\varphi$ is STIT-valid.)*

Although the two logics NCL and STIT are different, the property of independence of agents holds in NCL. This fact is stated in the following Lemma 1 and illustrated in Figure 1. Every NCL-model satisfies the constraint 3 (*assumption of independence of agents*) of Definition 1. This property will be important in the constructive proof of Theorem 2. More precisely, it will be used in the proof of Lemma 2 (see the Annex at the end of the paper).

Lemma 1. *Let $\mathcal{M} = (W, R, V)$ be a NCL-model. Let r be a positive integer*[4]*. Let $w_1, \ldots w_r \in W$ be such that for all $i, j \in \{1, \ldots r\}$, $w_i R_\emptyset w_j$. Let $J_1, \ldots J_r \subseteq AGT$ be such that $i \neq j$ implies $J_i \cap J_j = \emptyset$. We have:*

$$\bigcap_{i=1\ldots r} R_{J_i}(w_i) \neq \emptyset.$$

Our fragment dfSTIT of STIT logic with agents and groups has interesting computational properties. In the rest of this section, we are going to show that dfSTIT can be axiomatized by the axiomatics of the logic NCL, and that dfSTIT is decidable. To prove this, we are going to study the link between NCL and STIT when we restrict formulas to the fragment dfSTIT. Proposition 1 given above explains that in the general case, if a formula is STIT-satisfiable then it is NCL-satisfiable. The following Theorem 2 explains that the notion of satisfiability in STIT and in NCL is the same if we restrict formulas to the fragment dfSTIT.

Theorem 2. *Let $\varphi \in \mathcal{L}_{df\,STIT}$. Then, the following three propositions are equivalent:*

1. *φ is NCL-satisfiable;*
2. *φ is STIT-satisfiable;*
3. *φ is STIT-satisfiable in a polynomial sized product STIT-model.*

Figure 2 highlights the relation between STIT and NCL. If we consider the whole set of formulas $\mathcal{L}_{\mathsf{STIT}}$, then we have that all validities of NCL are validities of STIT but not the converse. But if we restrict formulas to the fragment $\mathcal{L}_{df\mathsf{STIT}}$, then the set of validities of NCL is equal to the set of validities of STIT.

3.4. Decidability and axiomatization

The result of Theorem 2 is close to the result of Pauly in [43]. In [43], Pauly compares strategic form games (like STIT-models) and CL standard models (like NCL-models). Theorem 2 provides two crucial results: one about complexity and another one about axiomatization of dfSTIT.

The following corollary follows from the equivalence between point 2 and 3 in the Theorem 2.

Corollary 1. *Deciding if a formula in $\mathcal{L}_{df\,STIT}$ is STIT-satisfiable is NP-complete.*

The following corollary follows from the equivalence between point 1 and 2 in the Theorem 2.

Corollary 2. *A formula φ in $\mathcal{L}_{df\,STIT}$ is STIT-valid iff we have $\vdash_{NCL} \varphi$.*

Of course, a proof of formula φ in $\mathcal{L}_{df\mathsf{STIT}}$ can contain formulas of $\mathcal{L}_{\mathsf{STIT}}$ that are not in $\mathcal{L}_{df\mathsf{STIT}}$.

[4]Note that Lemma 1 in the degenerated case $r = 0$, says that $\bigcap_{i=1\ldots 0} R_{J_i}(w_i) \neq \emptyset$. This is true because the intersection of zero subset is $W \times W$ by convention.

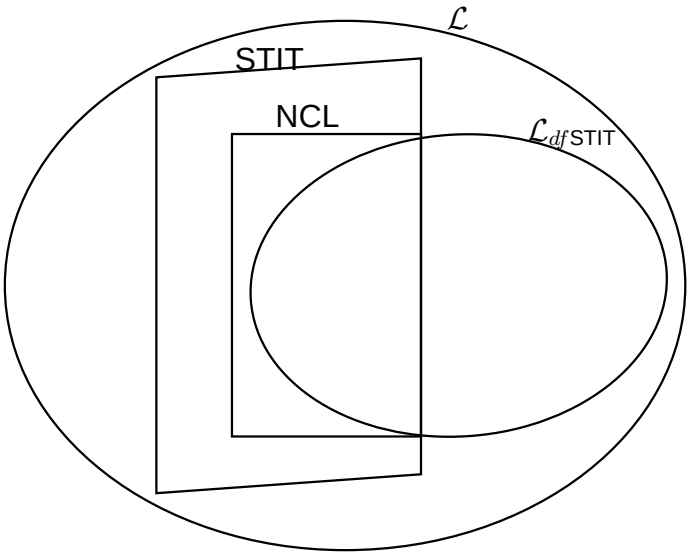

Figure 2: Overview of the languages $\mathcal{L}$ and $\mathcal{L}_{df\,\mathsf{STIT}}$ and of the logics STIT and NCL.

3.5. Discussion

Before concluding this section, let us explain why we decided to use $df\,\mathsf{STIT}$ instead of NCL for our logical analysis of counterfactual emotions.

The first reason is practical as the complexity of $df\,\mathsf{STIT}$ is lower than the complexity of NCL: the satisfiability problem for NCL is NEXPTIME-complete [5] while it is NP-complete for $df\,\mathsf{STIT}$ (Corollary 1). Moreover, as we will show in Section 5 (Theorem 3), the complexity of $df\,\mathsf{STIT}$ extended by epistemic modal operators is still lower than the complexity of NCL, in particular the satisfiability problem for the epistemic extension of $df\,\mathsf{STIT}$ is PSPACE-complete.

The second reason is theoretical. While STIT semantics has received several philosophical and conceptual justifications in works by Belnap, Horty and col. (see, e.g., [9, 30, 31]) and is nowdays widely accepted in the fields of philosophical logic and of logics for multi-agent systems, NCL semantics does not have such a robust conceptual and philosophical basis. Indeed, NCL was developed mainly in order to embed CL into a decidable normal modal logic. For instance, we have shown that in Horty's STIT logic, the set of outcomes that is forced by the joint action of a group J is equal to the pointwise intersection of the sets of outcomes that are forced by the individual actions of the agents in J (Constraint 2 in Definition 1). This is a natural way to define the notion of group action which is well-justified by Horty in [30]. But such a property of group action does not hold in the NCL semantics, and this is one the reason why the

notion of group action in NCL is not as clear as in STIT.[5]

It is worth noting that NCL and STIT already differs with a formula of modal depth 3. Indeed, the formula $\varphi = \neg[\langle\{2,3\}\rangle p \wedge \langle\{1,3\}\rangle q \wedge \langle\{1,2\}\rangle r \rightarrow \langle\emptyset\rangle[\langle\{2,3\}\rangle(\langle\{1,3\}\rangle p \wedge \langle\{2,3\}\rangle q) \wedge \langle\{1,3\}\rangle(\langle\{2,3\}\rangle r \wedge \langle\{1,2\}\rangle p) \wedge \langle\{2,3\}\rangle(\langle\{1,2\}\rangle q \wedge \langle\{1,3\}\rangle r)]]$ is NCL-satisfiable and $\neg\varphi$ is STIT-valid [24]. It is an open question whether NCL and STIT differs with a formula of modal depth 2.

4. Counterfactual statements in STIT

In this section we exploit the STIT fragment dfSTIT studied in Section 3 in order to formalize counterfactual statements of the form "group J (or agent i) could have prevented a certain state of affairs χ to be true now". Such statements are indeed basic constituents of the appraisal patters of counterfactual emotions such as regret. In particular, counterfactual emotions such as regret originate from reasoning about this kind of statements highlighting the connection between the actual state of the world and a counterfactual state of the world that might have been had one chosen a different action. The counterfactual statements formalized in this section will be fundamental in the formalization of counterfactual emotions we will give in Section 6.

4.1. J could have prevented χ

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

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

The statement just means that there is a *counterfactual dependence* between the state of affairs χ and group J (i.e. χ counterfactually depends on J's choice). The STIT fragment studied in Section 3 allows to represent it in a formal language. We write $\texttt{CHP}_J\chi$ this formal representation, which is defined as follows:

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

The expression $\neg[AGT \setminus J]\chi$ 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 joint action, χ would be false now. Thus, χ and $\neg[AGT \setminus J]\chi$ together correctly translate the previous counterfactual statement (*). If J is a singleton $\{i\}$, we write $\texttt{CHP}_i\chi$ instead of $\texttt{CHP}_{\{i\}}\chi$ which means "agent i could have prevented χ to be true".

[5] Note that in NCL semantics we only have $R_J \subseteq \bigcap_{j\in J} R_{\{j\}}$.

The following is the semantic counterpart of the operator $\texttt{CHP}_J$. We have that $\mathcal{M}, w \models \texttt{CHP}_J\chi$ if and only if, $\mathcal{M}, w \models \chi$ and there is $v \in R_{AGT\setminus J}(w)$ such that $\mathcal{M}, v \models \neg\chi$. That is, at world w of model $\mathcal{M}$, J could have prevented χ to be true if and only if, χ is true at w and, given what the agents in $AGT \setminus J$ have chosen at w, there exists a joint action of the agents in J such that, if the agents in J did choose this action, the actual outcome of the joint action of all agents would be a state in which χ is false.

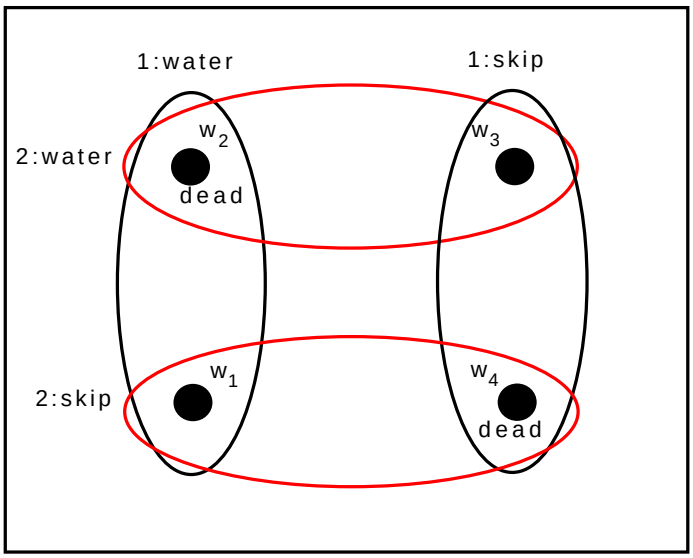

Figure 3: The four worlds w_1, w_2, w_3 and w_4 are in the equivalence class determined by $R_\emptyset$. Vertical circles represent the actions that agent 1 can choose, whereas horizontal circles represent the actions that agent 2 can choose. For example, w_1 is the world that results from agent 1 choosing the action $water$ and agent 2 choosing the action $skip$.

Example 1. *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. 3. 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.*

The following are some interesting properties of the operator $\texttt{CHP}_J$. For every J and for every J_1, J_2 such that $J_1 \subseteq J_2$:

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

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

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

$$\models_{\mathsf{STIT}} \neg\texttt{CHP}_J\top \quad (4)$$

$$\models_{\mathsf{STIT}} \neg\texttt{CHP}_J\bot \quad (5)$$

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. Validity 3 shows how the operator $\texttt{CHP}_J$ behaves over conjunction: if J_1 could have prevented χ_1 to be true and could have prevented χ_2 to be true separately then J_1 could have prevented χ_1 and χ_2 to be true. Finally, according to the Validities 4 and 5, tautologies and contradictions cannot counterfactually depend on the choice of a group: it is never the case that a coalition J could have prevented a tautology (resp. a contradiction).

4.2. Discussion

The following two sections discuss some aspects related to the analysis of counterfactual statements presented above. We first motivate why we chose STIT logic instead of concurrent logics such as Coalition Logic (CL) and ATL in order to provide a formal representation of such statements. Then, we make a brief excursus on the notion of "partial responsibility up to a certain degree".

4.2.1. Limitations of CL compared to STIT

In recent times several logics of group actions and group abilities have been proposed. Roughly, we can distinguish two families of such logics: those based on Coalition Logic (CL) [43], one for all Alternating-time temporal logic (ATL) [4] of which several variants and extensions have been studied (see, e.g., [2, 65, 2, 66]), and those based on STIT logic.

As shown in [12], STIT embeds CL, and STIT extended with strategies (so called strategic STIT) embeds ATL. The interesting point is that, while the statements:

1. "the group of agents J has a joint strategy that force χ" and,

2. "the group of agents J has not a joint strategy that force χ"

are expressible in STIT but also in CL and ATL, the statements:

3. "the group of agents J has chosen a joint strategy that force χ" and,

4. "the group of agents J did not choose a joint strategy that force χ"

are only expressible in STIT. More generally, while ATL and CL only support reasoning about what agents and coalitions of agents *can do* together, STIT also allows to express what agents and coalitions of agents *actually do* together (see also [11] for a discussion on this matter). In formal terms, the previous statements 1 and 2 are expressed in STIT by the formulas $\langle\emptyset\rangle[J]\chi$ and $\neg\langle\emptyset\rangle[J]\chi$, while the previous statements 3 and 4 are expressed in STIT by the formulas $[J]\chi$ and $\neg[J]\chi$.

As emphasized in Section 4.1, a logical analysis of counterfactual emotions is necessarily based on a logical analysis of counterfactual constructions of the form "agent i *could have prevented* χ to be true" which implies that:

5. "given what the agents in $AGT\backslash\{i\}$ have chosen, there exists an alternative action of agent i such that, if agent i did choose this action, the state of affairs χ would be false now".

We have shown that the previous statement 5 is expressed in STIT by the formula $\neg[AGT \setminus \{i\}]\chi$. As for statements 3 and 4 above, CL and ATL cannot express the previous statement 5. More generally, while STIT allows to express what agents and coalitions of agents *could have done* and *could have prevented*, this cannot be expressed in CL and ATL.

4.2.2. Partial responsibility up to a certain degree

We have given above a logical translation of the statement "agent i could have prevented χ to be true" noted $\texttt{CHP}_i\chi$ and expressing a counterfactual dependence between the state of affairs χ and agent i's choice.

It is worth noting that $\texttt{CHP}_i\chi$ does not cover situations in which agent i is partially responsible for χ *up to a certain degree* without being fully responsible for χ. The following voting example illustrates the difference between full responsibility and partial responsibility.

Example 2. *A and B are the two candidates for an election and* 1, 2, 3 *are the three voters. Suppose* w_7 *in the STIT model in Fig. 4 is the actual world. In this world, voter* 1 *and voter* 2 *vote for candidate A while voter* 3 *votes for candidate B so that A wins the election against B by a vote of 2-1. Formulas* $\texttt{CHP}_1 Awin$ *and* $\texttt{CHP}_2 Awin$ *are true at* w_7*. In fact, at* w_7 *candidate A wins the elections and, given what the other voters have*

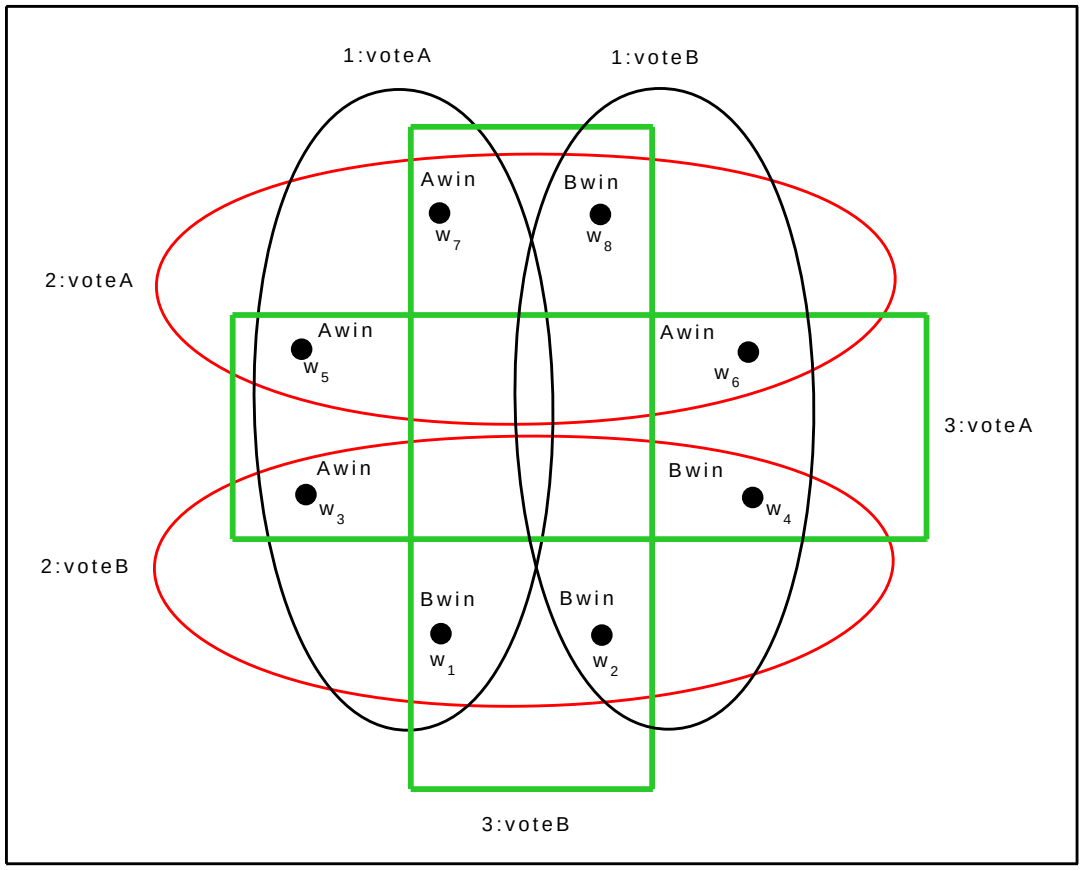

Figure 4: Vertical circles represent the actions that voter 1 can choose, horizontal circles represent the actions that voter 2 can choose, and rectangles represent the actions that voter 3 can choose. For example, w_1 is the world in which candidate B wins the election and that results from agent 1 voting for candidate A, and agents 2 and 3 voting for candidate B.

chosen, there exists an alternative action of voter 1 *(i.e. voting for candidate B) such that, if voter* 1 *did choose this action, candidate A would not win the elections. In other words, at* w_7 *the result of the election counterfactually depends on* 1*'s vote. The same is true for voter* 2*: at* w_7 *the result of the election counterfactually depends on* 2*'s vote. In this case, voter* 1 *and voter* 2 *can be said to be* ***fully*** *responsible for candidate A's win.*

Suppose now w_5 *in the* STIT *model in Fig. 4 is the actual world. At* w_5 *candidate A wins the election against candidate B by a vote of 3-0. In this case,* $\texttt{CHP}_i Awin$ *is false for every voter, that is, for every voter the result of the election does not counterfactually depend on his vote. Nevertheless, we would like to say that each of the three voters is partially responsible for candidate A's win up to a certain degree. Indeed, voter* 1 *is is a cause of A winning even if the vote is 3-0 because, under the contingency that one of*

the other voters had voted for candidate B instead, voter 1*'s vote would have become critical; if he had then changed his vote, candidate A would not have won. The same is true for voter* 2 *and for voter* 3*.*

It is not the objective of this paper to provide a logical account of the notion of partial responsibility and of the corresponding notion of degree of responsibility. These notions have been studied for instance in [16] in which the degree of responsibility of an event A for an event B is supposed to be $\frac{1}{N+1}$, where N is the minimal number of changes that have to be made to the actual situation before B counterfactually depends on A. For instance, in the case of the 3-0 vote in the previous example, the degree of responsibility of any voter for the victory of candidate A is $\frac{1}{2}$, since one change has to be made to the actual situation before a vote is critical. In the case of the 2-1 vote, the degree of responsibility of any voter for the victory is 1, since no change has to be made to the actual situation before a vote is critical.

5. A STIT extension with knowledge

In order to capture the subjective dimension of emotions, this section presents an extension of the fragment dfSTIT of STIT logic presented in section 3 with standard operators for knowledge of the form $\mathtt{K}_i$, where i is an agent. The formula $\mathtt{K}_i\varphi$ means "agent i knows that φ is true". This is a necessary step for the formalization of counterfactual emotions that will be presented in section 6.

5.1. Definition

First we extend the language $\mathcal{L}_{\mathsf{STIT}}$ of the Subsection 3.1 with epistemic constructions $\mathtt{K}_i\varphi$. We give the language of all formulas we can construct with STIT operators and knowledge operators. The language $\mathcal{L}_{\mathsf{KSTIT}}$ of the logic KSTIT is defined by the following BNF:

$$\varphi ::= p \mid \varphi \wedge \varphi \mid \neg\varphi \mid [J]\varphi \mid \mathtt{K}_i\varphi$$

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

For the same reasons than in Section 3.1, we are here interested in a fragment of $\mathcal{L}_{\mathsf{KSTIT}}$. Indeed, the satisfiability problem of the logic KSTIT will be undecidable if the number of agents is more than 3 (because the logic KSTIT will be a conservative extension of the logic STIT which is already undecidable). So we focus into a syntactic fragment we call dfKSTIT.

The language $\mathcal{L}_{df\mathsf{KSTIT}}$ of logic dfKSTIT 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 $2^{AGT} \setminus \{\emptyset\}$. For instance,

$\mathtt{K}_1\langle\emptyset\rangle[\{1,2\}]p \in \mathcal{L}_{df\mathsf{KSTIT}}$. But $\langle\emptyset\rangle\mathtt{K}_1[\{1,2\}]p \notin \mathcal{L}_{df\mathsf{KSTIT}}$. Let us give the semantics of the logic dfKSTIT.

Definition 4 (KSTIT-model).
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 (see Definition 1);
- For all $i \in AGT$, E_i is an equivalence relation.

We can also view epistemic accessibility relations on W as functions from W to 2^W, that is, for every $i \in AGT$, $E_i(w) = \{v \in W \mid wE_iv\}$.

As usual truth conditions for atomic formulas and the boolean operators are entirely standard. Truth conditions for the STIT operators $[J]$ are given in Section 3. 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.

As usual, a formula φ is KSTIT-valid (noted $\models_{\mathsf{KSTIT}} \varphi$) iff φ is true in every world of every KSTIT-model. A formula φ is KSTIT-satisfiable iff there exists a KSTIT-model $\mathcal{M} = (W, \{R_J\}_{J\subseteq AGT}, \{E_i\}_{i\in AGT}, V)$ and a world $w \in W$ such that $\mathcal{M}, w \models \varphi$.

5.2. Decidability

The following is an extension of Corollary 1 given in Section 3.4.

Theorem 3. *The satisfiability problem of dfKSTIT is NP-complete if* $\mathrm{card}(AGT) = 1$ *and PSPACE-complete if* $\mathrm{card}(AGT) \geq 2$.

5.3. Axiomatization

The study of an axiomatization for dfKSTIT relies on an epistemic extension of the logic NCL presented in Section 3.3 which will also be axiomatizable. We call KNCL this epistemic extension of NCL. The syntax of the logic KNCL is the same as the logic KSTIT, that is to say $\mathcal{L}_{\mathsf{KNCL}} = \mathcal{L}_{\mathsf{KSTIT}}$.

Let us now give the definition of model for the logic KNCL.

Definition 5 (KNCL-model).
A KNCL-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 NCL-model (see Definition 3);

- For all $i \in AGT$, E_i is an equivalence relation.

Truth conditions, validity and satisfiability in KNCL are defined as usual. We can now prove an extension of Theorem 2, stating the equivalence between the satisfiability in KNCL and the satisfiability in KSTIT if we restrict the formula to the syntactic fragment $\mathcal{L}_{df\mathsf{KSTIT}}$.

Theorem 4. *Let φ be a formula of $\mathcal{L}_{df\mathsf{KSTIT}}$. We have equivalence between:*

- *φ is satisfiable in KNCL;*
- *φ is satisfiable in KSTIT.*

In the same way, we have an extension of the Corollary 2 about a complete axiomatization of the logic dfKSTIT.

Corollary 3. *A formula φ in $\mathcal{L}_{df\mathsf{KSTIT}}$ is KSTIT-valid iff we have $\vdash_{\mathsf{KNCL}} \varphi$ where $\vdash_{\mathsf{KNCL}} \varphi$ means that there exists a proof of φ using all principles of the logic NCL, and all principles of modal logic S5 for every $\mathtt{K}_i$.*

Of course, as for $\mathcal{L}_{df\mathsf{STIT}}$, a proof of a formula φ in $\mathcal{L}_{df\mathsf{KSTIT}}$ can contain formulas of $\mathcal{L}_{\mathsf{KSTIT}}$ that are not in $\mathcal{L}_{df\mathsf{KSTIT}}$.

6. A formalization of counterfactual emotions

In Section 2.2 we have provided an overview of psychological theories of counterfactual emotions and discussed definitions which are shared by most psychologists working in this area. In the following sections, we will use the STIT fragment extended with epistemic operators studied in Section 5 and called dfKSTIT, in order to provide a logical formalization of this class of emotions. We consider four types of counterfactual emotions: regret and its positive counterpart rejoicing, disappointment and its positive counterpart elation.

6.1. Regret and rejoicing

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. [17, 41]). 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 good for an agent.

We say that χ is good for agent i if and only if χ is true in all good/pleasant states for agent i. Formally:

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

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:

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

As the following valid formulas highlight, every operator $\texttt{DES}_i$ 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.

$$\models_{\texttt{KSTIT}} (\texttt{DES}_i\chi_1 \wedge \texttt{DES}_i(\chi_1 \rightarrow \chi_2)) \rightarrow \texttt{DES}_i\chi_2 \tag{6}$$

$$\models_{\texttt{KSTIT}} \texttt{DES}_i\chi \leftrightarrow \texttt{K}_i\texttt{DES}_i\chi \tag{7}$$

$$\models_{\texttt{KSTIT}} \neg\texttt{DES}_i\chi \leftrightarrow \texttt{K}_i\neg\texttt{DES}_i\chi \tag{8}$$

We have now all necessary and sufficient ingredients to define the cognitive structure of regret and to specify its counterfactual dimension. As emphasized in Section 2.2, such a dimension has been widely studied in the psychological literature where several authors agree in considering regret as the emotion originating from an agent's comparison between the actual bad outcome and a *counterfactual* good outcome that might have been had the agent chosen a different action (see, e.g., [37, 59, 48, 33, 32, 71]).

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:

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

The following is the semantic counterpart of the previous syntactic definition of regret. We have that $\mathcal{M}, w \models \texttt{REGRET}_i\chi$ if and only if for all $v \in E_i(w)$ it holds that:

- for all $u \in R_\emptyset(v)$, if $\mathcal{M}, u \models good_i$ then $\mathcal{M}, u \models \neg\chi$;
- $\mathcal{M}, v \models \chi$ and there is $u \in R_{AGT\setminus\{i\}}(v)$ such that $\mathcal{M}, u \models \neg\chi$.

The former condition captures the *motivational* aspect of regret: if at world w agent i regrets for χ then, for every situation that agent i considers possible at w, $\neg\chi$ is pleasant for him. The latter condition captures the *counterfactual* aspect of regret: if at world w agent i regrets for χ then, for every situation that agent i considers possible at w, χ is true and, given what the other agents have chosen, there exists an alternative action of i such that, if i did choose this action, χ would be false now.

The following example is given in order to better clarify our logical definition of regret.

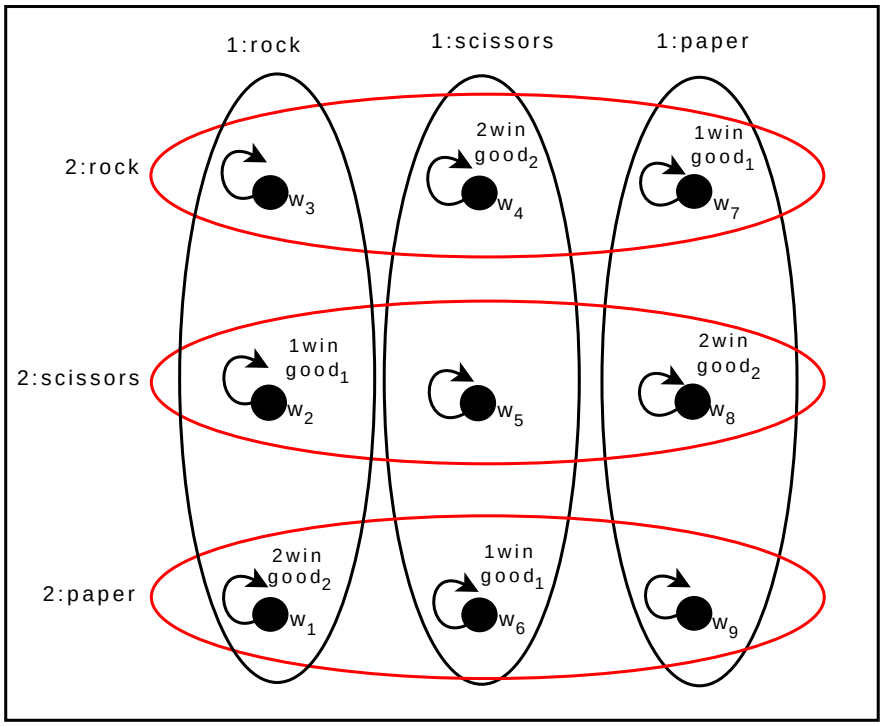

Figure 5: Vertical circles represent the actions that player 1 can choose, whereas horizontal circles represent the actions that player 2 can choose. For the sake of simplicity, we suppose that players 1 and 2 do not have uncertainty: everywhere in the model players 1 and 2 only consider possible the world in which they are (reflexive arrows represent indistinguishability relations for the two players).

Example 3. *Consider the popular two-person hand game "Rock-paper-scissors". 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. 5. It is supposed 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 de-

sire:

$$\models_{\mathsf{KSTIT}} \mathtt{REGRET}_i\chi \rightarrow (\mathtt{K}_i\chi \wedge \mathtt{DES}_i\neg\chi) \tag{9}$$

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:

$$\models_{\mathsf{KSTIT}} \mathtt{REGRET}_i\chi \leftrightarrow \mathtt{K}_i\mathtt{REGRET}_i\chi \tag{10}$$

$$\models_{\mathsf{KSTIT}} \neg\mathtt{REGRET}_i\chi \leftrightarrow \mathtt{K}_i\neg\mathtt{REGRET}_i\chi \tag{11}$$

As emphasized by some psychological theories of counterfactual emotions (see, e.g., [68, 69]), 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:

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

In semantic terms, we have that $\mathcal{M}, w \models \mathtt{REJOICE}_i\chi$ if and only if for all $v \in E_i(w)$ it holds that:

- for all $u \in R_\emptyset(v)$, if $\mathcal{M}, u \models good_i$ then $\mathcal{M}, u \models \chi$;
- $\mathcal{M}, v \models \chi$ and there is $u \in R_{AGT\setminus\{i\}}(v)$ such that $\mathcal{M}, u \models \neg\chi$.

The former condition corresponds to the *motivational* dimension of rejocing, while the latter corresponds to the *counterfactual* dimension. According to the former condition: if at world w agent i rejoices for χ then, for every situation that agent i considers possible at w, χ is pleasant for him. According to the latter condition: if at world w agent i rejoices for χ then, for every situation that agent i considers possible at w, χ is true and, given what the other agents have chosen, there exists an alternative action of i such that, if i did choose this action, χ would be false now.

Example 4. *Consider again the game "Rock-paper-scissors" represented by the* STIT*-model in Fig. 5. 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:

$$\models_{\mathsf{KSTIT}} \mathtt{REJOICE}_i\chi \rightarrow (\mathtt{K}_i\chi \wedge \mathtt{DES}_i\chi) \tag{12}$$

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). Like regret, rejoicing satisfies the properties of positive and negative introspection:

$$\models_{\mathsf{KSTIT}} \mathtt{REJOICE}_i\chi \leftrightarrow \mathtt{K}_i\mathtt{REJOICE}_i\chi \tag{13}$$

$$\models_{\mathsf{KSTIT}} \neg\mathtt{REJOICE}_i\chi \leftrightarrow \mathtt{K}_i\neg\mathtt{REJOICE}_i\chi \tag{14}$$

That is, agent i rejoices (resp. does not rejoice) for χ if and only if it knows this.

6.2. Disappointment and elation

As emphasized in Section 2.2, according to some authors [38, 19, 70], disappointment too is part of the family of counterfactual emotions: like regret, disappointment originates from the comparison between the actual outcome and a counterfactual outcome that might have occurred. However, there is an important difference between regret and disappointment. If an agent feels regret he considers himself to be responsible for the actual outcome, whereas if he feels disappointed he considers external events and other agents' actions to be responsible for the actual outcome.

Thus, we can say that an agent i feels disappointed for χ if and only if $\neg\chi$ is desirable for i and i knows that the others could have prevented χ to be true now. Formally:

$$\mathtt{DISAPPOINTMENT}_i\chi \stackrel{\text{def}}{=} \mathtt{DES}_i\neg\chi \wedge \mathtt{K}_i\mathtt{CHP}_{AGT\setminus\{i\}}\chi.$$

In semantic terms, we have that $\mathcal{M}, w \models \mathtt{DISAPPOINTMENT}_i\chi$ if and only if for all $v \in E_i(w)$ it holds that:

- for all $u \in R_\emptyset(v)$, if $\mathcal{M}, u \models \mathit{good}_i$ then $\mathcal{M}, u \models \neg\chi$;
- $\mathcal{M}, v \models \chi$ and there is $u \in R_{\{i\}}(v)$ such that $\mathcal{M}, u \models \neg\chi$.

Like in the cases of regret and rejoicing, the former condition captures the *motivational* aspect of disappointment, while the latter captures the *counterfactual* aspect. According to the former condition: if at world w agent i feels disappointed for χ then, for every situation that agent i considers possible at w, $\neg\chi$ is pleasant for him. According to the latter condition: if at world w agent i feels disappointed for χ then, for every situation that agent i considers possible at w, χ is true and, given what i has chosen, there exists an alternative joint action of the other agents such that, if they did choose this action, χ would be false now.

Example 5. *In the "Rock-paper-scissors" game represented in Fig. 5, regret is always joined with disappointment. For instance, at world w_1 player 1 not only regrets for having lost the game (i.e.* $\texttt{REGRET}_1\neg 1Win$*), but also he feels disappointed for this (i.e.* $\texttt{DISAPPOINTMENT}_1\neg 1Win$*). In fact, at w_1, 1 knows that (by playing $scissors$) the others (i.e. player 2) could have prevented $\neg 1Win$ to be true (i.e.* $\texttt{K}_1\texttt{CHP}_{AGT\setminus\{1\}}\neg 1Win$ *is true at w_1).*

Like regret and rejoicing, disappointment satisfies the properties of positive and negative introspection:

$$\models_{\mathsf{KSTIT}} \texttt{DISAPPOINTMENT}_i\chi \leftrightarrow \texttt{K}_i\texttt{DISAPPOINTMENT}_i\chi \tag{15}$$

$$\models_{\mathsf{KSTIT}} \neg\texttt{DISAPPOINTMENT}_i\chi \leftrightarrow \texttt{K}_i\neg\texttt{DISAPPOINTMENT}_i\chi \tag{16}$$

Moreover, like regret, disappointment implies desire frustration:

$$\models_{\mathsf{KSTIT}} \texttt{DISAPPOINTMENT}_i\chi \rightarrow (\texttt{K}_i\chi \wedge \texttt{DES}_i\neg\chi) \tag{17}$$

It is worth noting that regret and disappointment do not necessarily occur in parallel, i.e. the formulas $\texttt{REGRET}_i\chi\wedge\neg\texttt{DISAPPOINTMENT}_i\chi$ and $\neg\texttt{REGRET}_i\chi\wedge\texttt{DISAPPOINTMENT}_i\chi$ are satisfiable. The following example illustrates the situation in which an agent feels disappointed without feeling regret.

Example 6. *Two agents $AGT = \{1, 2\}$ have made an appointment to dine together at a restaurant. When the time of the appointment comes near, each of the two agents can either go to the restaurant in order to meet the other, or stay home, or go to the cinema. The two agents will have dinner together only if each of them decides to go to restaurant to meet the other. The scenario is represented in the* STIT *model in Fig. 6. It is supposed that having dinner with agent 2 is something good for agent 1 and agent 1 desires to have dinner with agent 2:* $\texttt{GOOD}_1 dinnerTogether$ *and* $\texttt{DES}_1 dinnerTogether$ *are true at worlds w_1-w_9. Suppose world w_1 is the actual world in which 1 goes to the restaurant, while 2 goes to the cinema and breaks his appointment with 1. In this world 1 does not have dinner with 2 ($\neg dinnerTogether$), and 1 knows that (by going to the restaurant) the others (i.e. agent 2) could have prevented $\neg dinnerTogether$ to be true (i.e.* $\texttt{K}_1\texttt{CHP}_{AGT\setminus\{1\}}\neg dinnerTogether$ *is true at w_1). It follows that at w_1 agent 1 feels disappointed for not having dinner with 2, that is,* $\texttt{DISAPPOINTMENT}_1\neg dinnerTogether$ *is true at w_1. Note that at w_1 agent 1 does not feel regret for not having dinner with agent 2 (i.e.* $\texttt{REGRET}_1\neg dinnerTogether$ *is false at w_1). In fact, at w_1, 1 knows that $\neg dinnerTogether$ only depends on what 2 has decided to do. Therefore, at w_1, 1 does not think that he could have prevented $\neg dinnerTogether$ to be true (i.e.* $\neg\texttt{K}_1\texttt{CHP}_1\neg dinnerTogether$ *is true at w_1).*

We conclude with a formalization of the positive counterpart of disappointement, that is commonly called *elation* [68, 69]. We say that agent i elates for χ if and only if,

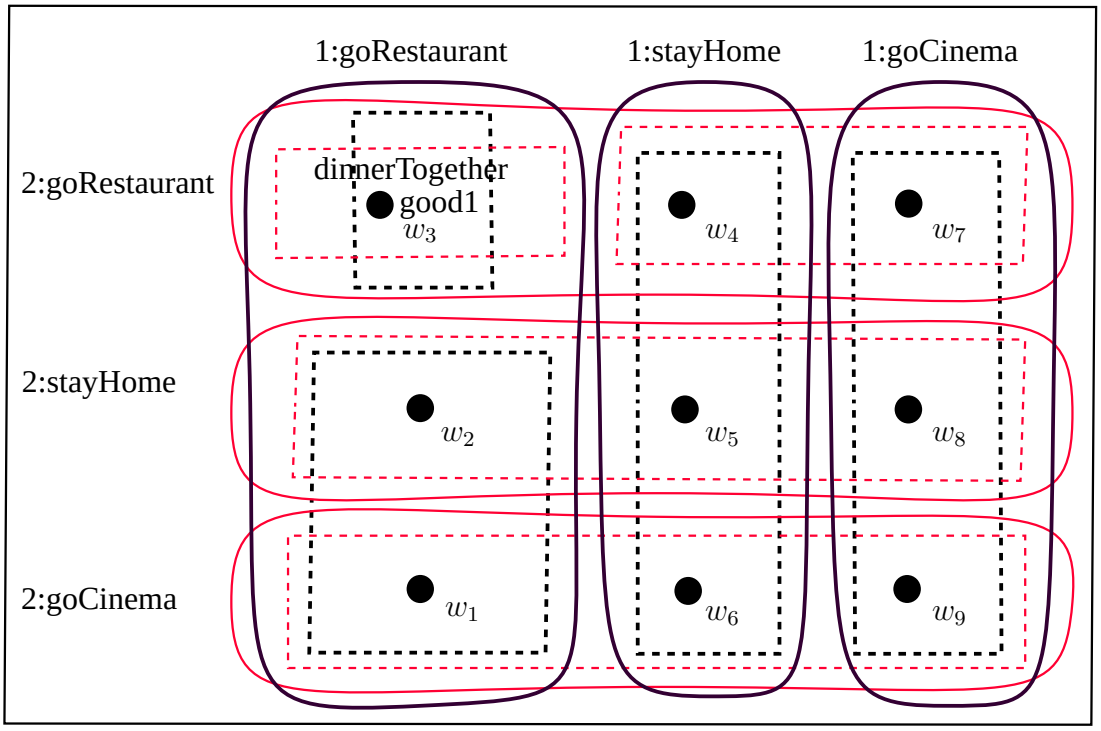

Figure 6: Again, vertical circles represent the actions that agent 1 can choose, whereas horizontal circles represent the actions that agent 2 can choose. In this example, we suppose that agents 1 and 2 can only have uncertainty about the current choice of the other (vertical dotted rectangles represent indistinguishability relations for agent 1, whereas horizontal dotted rectangles represent indistinguishability relations for agent 2).

χ is desirable for i and i knows that the others could have prevented χ to be true now:

$$\texttt{ELATION}_i\chi \stackrel{\text{def}}{=} \texttt{DES}_i\chi \wedge \texttt{K}_i\texttt{CHP}_{AGT\setminus\{i\}}\chi.$$

In semantic terms, we have that $\mathcal{M}, w \models \texttt{ELATION}_i\chi$ if and only if for all $v \in E_i(w)$ it holds that:

- for all $u \in R_\emptyset(v)$, if $\mathcal{M}, u \models \mathit{good}_i$ then $\mathcal{M}, u \models \chi$;
- $\mathcal{M}, v \models \chi$ and there is $u \in R_{\{i\}}(v)$ such that $\mathcal{M}, u \models \neg\chi$.

Like in the cases of regret, rejoicing and disappointment, the former condition captures the *motivational* aspect of elation while the latter captures the *counterfactual* aspect. According to the former condition: if at world w agent i elates for χ then, for every situation that agent i considers possible at w, χ is pleasant for him. According to the latter condition: if at world w agent i elates for χ then, for every situation that agent i

considers possible at w, χ is true and, given what i has chosen, there exists an alternative joint action of the other agents such that, if they did choose this action, χ would be false now.

Like regret, rejoicing and disappointment, elation satisfies the properties of positive and negative introspection:

$$\models_{\mathsf{KSTIT}} \mathtt{ELATION}_i\chi \leftrightarrow \mathtt{K}_i\mathtt{ELATION}_i\chi \tag{18}$$

$$\models_{\mathsf{KSTIT}} \neg\mathtt{ELATION}_i\chi \leftrightarrow \mathtt{K}_i\neg\mathtt{ELATION}_i\chi \tag{19}$$

Moreover, like rejoicing, elation implies desire satisfaction:

$$\models_{\mathsf{KSTIT}} \mathtt{ELATION}_i\chi \rightarrow (\mathtt{K}_i\chi \wedge \mathtt{DES}_i\chi) \tag{20}$$

Finally, like regret and disappointment, elation and rejoicing do not necessarily occur in parallel, i.e. the formulas $\mathtt{REJOICE}_i\chi \wedge \neg\mathtt{ELATION}_i\chi$ and $\neg\mathtt{REJOICE}_i\chi \wedge \mathtt{ELATION}_i\chi$ are satisfiable. In fact, an agent might consider the others to be responsible for the actual good situation, without considering himself to be responsible for the actual good situation.

Before concluding, it is worth noting that the constructions $\mathtt{REGRET}_i\chi$ and $\mathtt{REJOICE}_i\chi$ require group STIT operators. This justifies the use of Horty's STIT logic with agents and groups, and the study of a decidable fragment of this logic provided in Section 3. On the contrary, Belnap et col.'s individual STIT logic [9] extended by knowledge operators would be sufficient to write the formulas $\mathtt{DISAPPOINTMENT}_i\chi$ and $\mathtt{ELATION}_i\chi$. Let us recall that the fusion of two decidable modal logics is decidable [24]. As the satisfiability problem of Belnap et col.'s individual STIT logic is decidable [7], the fusion of the latter and epistemic logic is also decidable. So, it is not necessary to introduce syntactic restrictions on the STIT language in order to obtain a decidable logic in which we can reason about disappointment and elation.

6.3. Discussion

Let us discuss some aspects we did not consider in the previous formalization of counterfactual emotions.

According to [14], disappointment entails invalidation of an agent's positive expectation. That is, an agent feels disappointed for χ, only if $\neg\chi$ is desirable for the agent and the agent believes that χ, and in the previous state he believed $\neg\chi$ to be true in the next state. In other words, an agent feels disappointed for χ because he would like χ to be false now and he just learnt that χ is true and, before learning that χ is true, he believed $\neg\chi$ to be true in the next state. In the formalization of disappointment proposed in Section 6.2, this relationship between disappointment and expectations was not considered. We included in the definition of disappointment only the agent's mental states at the moment in which the emotion arises.

Another aspect we did not consider in our formalization of counterfactual emotions is the distinction between regret due to a *choice to act* (i.e. action) and regret due to a *choice not to act* (i.e. inaction). A classical example which clarifies this distinction is the one given by [34] in which an agent i owned shares in company A, and he considered switching to stock in company B but he decided against it. He now finds out that he would have been better off if he had switched to the stock of company B (regret due to inaction). Another agent j owned shares in company B, and he switched to stock in company A. He now finds out that he would have been better off if he had kept his stock in Company B (regret due to action). The logic STIT is not sufficiently expressive to make this distinction between regret due to action and regret due to inaction. Indeed, in STIT logic it is supposed that at a given state w every agent has made a choice. Moreover, STIT allows to reason about the effects of the agents' choices at a given state. Nevertheless, STIT does not allow to distinguish the situation in which, at a given state, an agent has made the choice to act from the situation in which the agent has made the choice not to act.

7. A concrete example

The logical framework and formal analysis of counterfactual emotions proposed in this paper can also be exploited for increasing the competence and performance of artificial emotional agents in emotion recognition, emotion anticipation, response to others' emotions and emotion communication and expression. Such capabilities are fundamental for developing interactive agent technologies which are particularly relevant for applications in healthcare, education and entertainment, like intelligent tutoring systems, robotic assistants to older or disabled people to improve quality of life, companions and trainers in physical recovery and rehabilitation, etc. This section exposes more in detail how the results of the present research can be exploited in order to design agents endowed with these capabilities.

We imagine a scenario of human-agent interaction in which an intelligent tutoring agent has to take care of a human user. The tutoring agent has to reason about and to respond to the user's emotions in order to sustain the user's activity. Here we only focus on some particular competencies of the artificial agent, namely: the capacity of inferring the user's emotions by attributing mental states to the user; the capacity of adapting its behavior during the dialogue with the user in order to reduce the user's negative emotions and in order to induce positive emotions on the user.

7.1. Inferring the user's emotion through the attribution of mental states

The human user in this scenario is a student who has to pass a Certificate of Proficiency in English. The tutoring agent is an artificial agent who supervises the student's preparation for the exam. The tutoring agent is endowed with the capability of reasoning about the student's emotions.

Let us suppose that, according to the tutoring agent (noted t): the user (noted u) would like to pass the exam, the user knows that he did not pass the exam, the user knows that necessarily if he studied then he would have passed the exam, and the user knows that he had the opportunity to study. Thus, the tutoring agent's knowledge base $\mathcal{KB}$ can be formally represented by the conjunction of the following four formulas:

- $\mathtt{K}_t\mathtt{DES}_u pass_u$
- $\mathtt{K}_t\mathtt{K}_u\neg pass_u$
- $\mathtt{K}_t\mathtt{K}_u[\emptyset]([u]studied_u \rightarrow pass_u)$
- $\mathtt{K}_t\mathtt{K}_u\langle\emptyset\rangle[u]studied_u$

Note that all the four formulas are in $\mathcal{L}_{df\mathsf{KSTIT}}$, even the third one which is equivalent to $\mathtt{K}_t\mathtt{K}_u\neg\langle\emptyset\rangle([u]studied_u \wedge [AGT]\neg pass_u)$. We can prove that from its initial knowledge base, the tutoring agent infers that the user is feeling regret for having failed the exam, that is

$$\models_{\mathsf{KSTIT}} \mathcal{KB} \rightarrow \mathtt{K}_t\mathtt{REGRET}_u\neg pass_u \tag{21}$$

Now let us suppose that, according to the tutoring agent: the user would like to pass the exam, the user knows that he passed the exam, the user knows that necessarily if he did not study then he would have failed the exam, and the user knows that he had the opportunity not to study. Thus, the tutoring agent's knowledge base $\mathcal{KB}^*$ can be formally represented by the conjunction of the following four formulas:

- $\mathtt{K}_t\mathtt{DES}_u pass_u$
- $\mathtt{K}_t\mathtt{K}_u pass_u$
- $\mathtt{K}_t\mathtt{K}_u[\emptyset]([u]\neg studied_u \rightarrow \neg pass_u)$
- $\mathtt{K}_t\mathtt{K}_u\langle\emptyset\rangle[u]\neg studied_u$

We can prove that from its initial knowledge base, the tutoring agent infers that the user is rejoicing for having passed the exam, that is

$$\models_{\mathsf{KSTIT}} \mathcal{KB}^* \rightarrow \mathtt{K}_t\mathtt{REJOICE}_u pass_u \tag{22}$$

We have only considered a tutoring agent's capability of inferring a human user's emotions by the attribution of mental states to the user. However, there are other important capabilities that a tutoring agent interacting with a human user should be endowed with. In particular, the tutoring agent should be able to communicate with the human user in such a way that it can adapt its behavior in order to reduce the user's negative emotions and in order to induce positive emotions on the user. In order to model this kind of capability, we discuss in the next section an extension of our logical framework that allows to represent the exchange of information between a tutoring agent and a human user.

7.2. A 'dynamification' of KSTIT

We present a dynamic variant of the logic of Section 5, where knowledge is updated, as in public announcement logic (PAL) [45, 67] and, more precisely, as in the variant of PAL proposed by [63, 64] where model update is redefined as an epistemic relation-changing operation of 'link cutting' that does not throw away worlds from a model.

The logic KSTIT of Section 5 is here extended by dynamic operators of the form $[\|\theta\|]$. The formula $[\|\theta\|]\varphi$ means 'after announcement of the truth value of θ, φ holds'. The dual of $[\|\theta\|]$ is $\langle\|\theta\|\rangle$, that is, $\langle\|\theta\|\rangle\varphi \stackrel{\text{def}}{=} \neg[\|\theta\|]\neg\varphi$. We call KSTIT$^+$ the extended logic. The language $\mathcal{L}_{\mathsf{KSTIT}^+}$ of the logic KSTIT$^+$ is defined by the following BNF:

$$\varphi ::= p \mid \varphi \wedge \varphi \mid \neg\varphi \mid [J]\varphi \mid \mathtt{K}_i\varphi \mid [\|\varphi\|]\varphi$$

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

We are interested here in a decidable and finitely axiomatizable fragment of KSTIT$^+$ called dfKSTIT$^+$, which is nothing else than the dynamic extension of the syntactic fragment dfKSTIT of the logic KSTIT studied in Section 5. The language $\mathcal{L}_{df\mathsf{KSTIT}^+}$ of logic dfKSTIT$^+$ 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)
$\varphi ::= \chi \mid \psi \mid \varphi \wedge \varphi \mid \neg\varphi \mid \langle\emptyset\rangle\psi \mid \mathtt{K}_i\varphi \mid [\|\varphi\|]\varphi$
(see-to-it, "can", knowledge, update formulas)
where p ranges over ATM, i ranges over *AGT* and J over $2^{AGT} \setminus \{\emptyset\}$.

The standard announcement operator $[!\theta]$ of PAL can be defined from the operator $[\|\theta\|]$ in a straightforward manner:

$$[!\theta]\varphi \stackrel{\text{def}}{=} \theta \rightarrow [\|\theta\|]\varphi.$$

Formula $[!\theta]\varphi$ has to be read 'after the announcement of θ, φ holds'. Indeed, the announcement of θ is nothing else than the announcement of the truth value of θ when θ is true. The dual of $[!\theta]$ is $\langle!\theta\rangle$, that is, $\langle!\theta\rangle\varphi \stackrel{\text{def}}{=} \neg[!\theta]\neg\varphi$.

In order to give semantics to the operators $[\|\theta\|]$ we define the elements of the model $\mathcal{M}^{\|\theta\|}$ which results from the update of the model $\mathcal{M} = (W, \{R_J\}_{J\subseteq AGT}, \{E_i\}_{i\in AGT}, V)$ by the announcement of θ's truth value:

- $W^{\|\theta\|} = W$;
- for every $J \subseteq$ *AGT*, $R_J^{\|\theta\|} = R_J$;
- for every $i \in$ *AGT*, $E_i^{\|\theta\|} = \{(w,v) | (w,v) \in E_i$ and $(M, w \models \theta$ iff $M, v \models \theta)\}$;
- $V^{\|\theta\|} = V$.

Basically, the effect of the announcement of θ's truth value is to remove the epistemic links between all worlds u and v in which θ does not have the same truth value. In other words, for every world w in which θ is true and for every agent i, the effect of the operation $\|\theta\|$ is to restrict the set of epistemically possible worlds for i to the set of worlds in which θ is true; for every world w in which θ is false and for every agent i, the effect of the operation $\|\theta\|$ is to restrict the set of epistemically possible worlds for i to the set of worlds in which θ is false.

It is just a routine to verify that the operation $\|\theta\|$ is well-defined, as it preserves the semantic constraints on KSTIT-models, that is, if $\mathcal{M}$ is a KSTIT-model then $\mathcal{M}^{\|\theta\|}$ is a KSTIT-model too.

The following are the truth conditions of the dynamic operators $[\|\theta\|]$:

$$M, w \models [\|\theta\|]\varphi \text{ iff } M^{\|\theta\|}, w \models \varphi.$$

Note that under these truth conditions $[\|\theta\|]\varphi$ is equivalent to $\langle\|\theta\|\rangle\varphi$. Validity of a formula φ in KSTIT^+ (noted $\models_{\mathsf{KSTIT}^+} \varphi$) is defined in the usual way.

Proposition 2. *The following schemata are KSTIT^+-valid:*

$$\begin{array}{rl}
(Red_p) & [\|\theta\|]p \leftrightarrow p \\
(Red_\neg) & [\|\theta\|]\neg\varphi \leftrightarrow \neg[\|\theta\|]\varphi \\
(Red_\wedge) & [\|\theta\|](\varphi_1 \wedge \varphi_2) \leftrightarrow ([\|\theta\|]\varphi_1 \wedge [\|\theta\|]\varphi_2) \\
(Red_{[J]}) & [\|\theta\|][J]\varphi \leftrightarrow [J][\|\theta\|]\varphi \\
(Red_{\mathtt{K}_i}) & [\|\theta\|]\mathtt{K}_i\varphi \leftrightarrow ((\theta \rightarrow \mathtt{K}_i(\theta \rightarrow [\|\theta\|]\varphi)) \wedge (\neg\theta \rightarrow \mathtt{K}_i(\neg\theta \rightarrow [\|\theta\|]\varphi)))
\end{array}$$

REMARK. It is straightforward to verify that the announcement operators $[!\theta]$ defined above satisfy the standard principle of PAL:

$$[!\theta]\mathtt{K}_i\varphi \leftrightarrow (\theta \rightarrow \mathtt{K}_i[!\theta]\varphi).$$

The five equivalences of Proposition 2 together with the rule of replacement of proved equivalents provide a complete set of reduction axioms for the dynamic operators $[\|\theta\|]$. We call red the mapping which iteratively applies the above equivalences from the left to the right, starting from one of the innermost modal operators. red pushes the dynamic operators inside the formula, and finally eliminates them when facing an atomic formula.

The mapping red is inductively defined by:

1. $red(p) = p$
2. $red(\neg\varphi) = \neg red(\varphi)$
3. $red(\varphi_1 \wedge \varphi_2) = red(\varphi_1) \wedge red(\varphi_2)$
4. $red([J]\varphi) = [J]red(\varphi)$
5. $red(\mathtt{K}_i\varphi) = \mathtt{K}_i red(\varphi)$
6. $red([\|\theta\|]p) = p$
7. $red([\|\theta\|]\neg\varphi) = red(\neg[\|\theta\|]\varphi)$
8. $red([\|\theta\|](\varphi_1 \wedge \varphi_2)) = red([\|\theta\|]\varphi_1 \wedge [\|\theta\|]\varphi_2)$
9. $red([\|\theta\|][J]\varphi) = red([J][\|\theta\|]\varphi)$
10. $red([\|\theta\|]\mathtt{K}_i\varphi) = red((\theta \rightarrow \mathtt{K}_i(\theta \rightarrow [\|\theta\|]\varphi)) \wedge (\neg\theta \rightarrow \mathtt{K}_i(\neg\theta \rightarrow [\|\theta\|]\varphi)))$
11. $red([\|\theta\|][\|\epsilon\|]\varphi) = red([\|\theta\|]red([\|\epsilon\|]\varphi))$

The following Proposition 3 is a straightforward consequence of Proposition 2 and the fact that the following rule of replacement of proved equivalents preserves validity:

$$\frac{\varphi_1 \leftrightarrow \varphi_2}{\varphi \leftrightarrow \varphi[\varphi_1 := \varphi_2]}$$

where $\varphi[\varphi_1 := \varphi_2]$ is the formula φ in which we have replaced all occurrences of φ_1 by φ_2.

Proposition 3. *Let $\varphi \in \mathcal{L}_{\mathsf{KSTIT}^+}$. Then, $red(\varphi) \leftrightarrow \varphi$ is KSTIT^+-valid.*

The following Proposition 4 is necessary in order to prove the completeness of the logic $df\mathsf{KSTIT}^+$.

Proposition 4. *Let $\varphi \in \mathcal{L}_{df\mathsf{KSTIT}^+}$. Then, $red(\varphi) \in \mathcal{L}_{df\mathsf{KSTIT}}$.*

Corollary 4. *The validities of $df\mathsf{KSTIT}^+$ are completely axiomatized by the axioms and inference rules of $df\mathsf{KSTIT}$ provided in Corollary 3 together with the reduction axioms of Proposition 2 and the rule of replacement of proved equivalents.*

Decidability of the logic of $df\mathsf{KSTIT}^+$ follows straightforwardly from the known decidability of the base logic $df\mathsf{KSTIT}$ (Theorem 3) and from Propositions 3 and 4. Indeed, red provides an effective procedure for reducing a $df\mathsf{KSTIT}^+$-formula φ into an equivalent $df\mathsf{KSTIT}$-formula $red(\varphi)$.

Corollary 5. *The satisfiability problem of $df\mathsf{KSTIT}^+$ is decidable.*

7.3. Adapting behavior during a dialogue with the user

It has to be noted that in dynamic epistemic logic announcements are usually viewed as communicative actions performed by an agent that is 'outside the system', i.e. that is not part of the set of agents *AGT* under consideration. However, communicative actions performed by agents in *AGT* can be modelled in PAL by considering a particular subset of announcements of agents' mental states. In particular, we here identify the event "agent i *announces* that φ is true" (or "agent i *says* that φ is true") with the announcement of the formula $\mathtt{K}_i\varphi$. Thus, we write $say(i,\varphi)$ instead of $!\mathtt{K}_i\varphi$, and $[say(i,\varphi)]\psi$ abbreviates $[!\mathtt{K}_i\varphi]\psi$. In other words, we here identify agent i's act of announcing that φ with the announcement of the fact that i knows that φ. A similar point of view is taken by [3].

The dynamic extension of the logic KSTIT presented in Section 7.2 can easily incorporate rules which specify how, during a dialogue with a human user, an artificial agent should adapt its behavior depending on the expected effects of certain dialogue moves on the user's emotions.

In order to formalize this kind of rules in our logic, we introduce a function Pre such that, for every formula θ in $\mathcal{L}_{df\mathsf{KSTIT}^+}$, $Pre(\theta)$ is the *feasibility (or executability) precondition* of the public announcement of θ. We denote with $\langle\langle !\theta\rangle\rangle\varphi$ the fact 'the public announcement of θ will possibly occur, and φ will be true afterwards', and we define it as follows:

$$\langle\langle !\theta\rangle\rangle\varphi \stackrel{\text{def}}{=} Pre(\theta) \wedge \langle !\theta\rangle\varphi.$$

Consequently, $\langle\langle !\theta\rangle\rangle\top$ is logically equivalent to $Pre(\theta)\wedge\theta$, that is, the public announcement of θ will possibly occur if and only if its feasibility precondition holds and θ is true. As we here identify the event "agent i *announces* that φ is true" (i.e. $say(i,\varphi)$) with the announcement of the formula $\mathtt{K}_i\varphi$ (i.e. $!\mathtt{K}_i\varphi$), $\langle\langle say(i,\varphi)\rangle\rangle\psi$ abbreviates $\langle\langle !\mathtt{K}_i\varphi\rangle\rangle\psi$.

REMARK. Note that the definition of the operator $\langle\langle !\theta\rangle\rangle$ forces agents to be sincere when performing a speech act. In particular, we have that an agent i will possibly announce that φ is true (i.e. $\langle\langle say(i,\varphi)\rangle\rangle\top$) only if i knows that φ is true (i.e. $\mathtt{K}_i\varphi$). This assumption about sincerity is however acceptable for the scenario introduced in Section 7.1 in which a tutoring agent which has to take care of a human user can be reasonably supposed to be cooperative and sincere with the human user.

Let us go back to the scenario introduced in Section 7.1. We suppose that in this scenario the tutoring agent's decision to perform a certain dialogue move depends on the tutoring agent's expectations about the effects of this dialogue move on the human user's emotions. In particular, we suppose that:

- the tutoring agent t will possibly tell to the human user u that he passed the exam if and only if, t knows that by telling to u that he passed the exam u will rejoice for having passed the exam and that at the present stage u does not rejoice for having passed the exam;

- the tutoring agent t will possibly tell to the human user u that he failed the exam if and only if, t knows that by telling to u that he failed the exam u will not regret for having failed the exam.

The previous two rules can be formally represented as follows:

$$Pre(say(t,pass_u)) = \mathtt{K}_t[say(t,pass_u)]\mathtt{REJOICE}_u pass_u \wedge \mathtt{K}_t\neg\mathtt{REJOICE}_u pass_u,$$

$$Pre(say(t,\neg pass_u)) = \mathtt{K}_t[say(t,\neg pass_u)]\neg\mathtt{REGRET}_u \neg pass_u.$$

Let us first suppose that, according to the tutoring agent: the user would like to pass the exam, the user does not know whether he passed the exam, the user knows that necessarily if studied then he would have passed the exam, and the user knows that he had the opportunity to study. Moreover, the tutoring agent knows that the user failed the exam. Thus, the tutoring agent's knowledge base $\mathcal{KB}^{**}$ can be formally represented by the conjunction of the following five formulas:

- $\mathtt{K}_t\mathtt{DES}_u pass_u$
- $\mathtt{K}_t(\neg\mathtt{K}_u pass_u \wedge \neg\mathtt{K}_u\neg pass_u)$
- $\mathtt{K}_t\mathtt{K}_u[\emptyset]([u]studied_u \rightarrow pass_u)$
- $\mathtt{K}_t\mathtt{K}_u\langle\emptyset\rangle[u]studied_u$
- $\mathtt{K}_t\neg pass_u$

The following validity highlights that, given its knowledge base $\mathcal{KB}^{**}$, the tutoring agent will refrain from telling to the user that he failed the exam.

$$\models_{\mathsf{KSTIT}^+} \mathcal{KB}^{**} \rightarrow \neg\langle\langle say(t,\neg pass_u)\rangle\rangle\top \tag{23}$$

Now let us suppose that, according to the tutoring agent: the user would like to pass the exam, the user does not know whether he passed the exam, the user knows that necessarily if did not study then he would have failed the exam, and the user knows that he had the opportunity not to study. Moreover, the tutoring agent knows that the user passed the exam. Thus, the tutoring agent's knowledge base $\mathcal{KB}^{***}$ can be formally represented by the conjunction of the following five formulas:

- $\mathtt{K}_t\mathtt{DES}_u pass_u$
- $\mathtt{K}_t(\neg\mathtt{K}_u pass_u \wedge \neg\mathtt{K}_u\neg pass_u)$
- $\mathtt{K}_t\mathtt{K}_u[\emptyset]([u]\neg studied_u \rightarrow \neg pass_u)$
- $\mathtt{K}_t\mathtt{K}_u\langle\emptyset\rangle[u]\neg studied_u$

- $\mathtt{K}_t pass_u$

The following validity highlights that, given its knowledge base $\mathcal{KB}^{***}$, the tutoring agent will possibly tell to the user that he passed the exam and, after that, the user will rejoice for having passed the exam.

$$\models_{\mathsf{KSTIT}^+} \mathcal{KB}^{***} \rightarrow \langle\langle say(t, pass_u)\rangle\rangle \mathtt{REJOICE}_u pass_u \tag{24}$$

8. Related works

As emphasized in the introduction emotion is a very active field in AI. Several computational architectures of affective agents have been proposed in the last few years (see, e.g., [46, 21, 18]). The cognitive architecture EMA (Emotion and Adaption) [27] is one of the best example of research in this area. EMA defines a domain independent taxonomy of appraisal variables stressing the many different relations between emotions and cognition, by enabling a wide range of internal appraisal and coping processes used for reinterpretation, shift of motivations, goal reconsideration etc. EMA also deals with complex social emotions based on attributions of responsibility such as guilt and shame.

There are also several researchers who have developed formal languages which allow to reason about emotions and to model affective agents. We discuss here some of the most important formal approaches to emotions and compare them with our approach.

Meyer et al.'s logic of emotions . One of the most prominent formal analysis of emotions is the one proposed by Meyer et al. [40, 58, 62]. In order to formalize emotions, they exploit the logical framework KARO [41]: 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 emotions on the agent's planning strategies are specified, as well as the preconditions for triggering the emotion. The latter correspond to generation rules for emotions. For instance, in [40] generation rules for four basic emotions are given: joy, sadness, anger and fear, depending on the agent's plans. More recently [62], generation rules for social emotions such as guilt and shame have been proposed.

Contrarily to Meyer et al.'s approach, in our logic there are no specific formal constructs, like special predicates or fluents, which are used to denote that a certain emotion arises at a certain time. We just *define* the appraisal pattern of a given emotion in terms of some cognitive constituents such as desire and knowledge. For instance, according to our definition of regret, an agent regrets for χ if and only if, he *desires* $\neg\chi$ and, i *knows* that it could have prevented χ to be true now. In other words, following the so-called appraisal theories in psychology (see Section 2), in our approach an emotion is reduced

to its appraisal variables which can be defined through the basic concepts of a BDI logic (e.g. knowledge, belief, desires, intentions).

It has to be noted that, although Meyer et al. provide a 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. In contrast to that, our STIT-based approach overcomes this limitation.

Note also that while Meyer et al. do not prove completeness and do not study complexity of their logic of emotions, these are central issues in our work. As emphasized in the introduction of the article, our aim is to develop 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.

Other logical approaches to emotions. Adam et al. [1] have recently exploited a BDI logic in order to provide a logical formalization of the emotion types defined in Ortony, Clore and Collins's model (OCC model) [42] (see Section 2.1 for a discussion of this model). Similar to our approach, in Adam et al.'s approach emotion types are defined in terms of some primitive concepts (and corresponding modal operators) such as the concepts of belief, desire, and action which allow to capture the different appraisal variables of emotions proposed in the OCC model such as the desirability of an event, probability of an event, and degree of responsibility of the author of an action. However, Adam et al. do not consider counterfactual emotions. In fact, the logic proposed by Adam et al. is not sufficiently expressive to capture counterfactual thinking about agents' choices and actions on which emotions like regret, rejoicing, disappointment and elation are based. Moreover, this is due to some limitations of the OCC typology which does not contain definitions of emotions based on counterfactual thinking such as regret and rejoicing.

In [20] 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 [42], 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.

9. Conclusion

A logical framework which allows to formalize and to reason about counterfactual emotions has been proposed in this paper. This framework is based on a decidable

and finitely axiomatizable fragment of STIT logic called dfSTIT. We have shown that an epistemic extension of dfSTIT called dfKSTIT is sufficiently expressive to capture the fundamental constituents of counterfactual emotions and, at the same time, it has good mathematical properties in terms of complexity and axiomatizability. We have proved that the satisfiability problem of dfKSTIT is NP-complete if *card*$(AGT) = 1$ and PSPACE-complete if *card*$(AGT) \geq 2$. This first result is fundamental in order to claim that we can write down algorithms in dfKSTIT to reason about counterfactual emotions such as regret, rejoicing, disappointment and elation. Moreover, we have provided a complete axiomatization of dfKSTIT . This second result is also important because it shows that we can perform syntactic reasoning in dfKSTIT about counterfactual emotions. We hope that the analysis developed in this paper will be useful for improving understanding of affective phenomena and will offer an interesting perspective on computational modeling of affective agents and systems.

Directions for our future research are manifold. The STIT fragment studied in Section 3 has an interesting expressivity, as it allows to capture subtle aspects of counterfactual reasoning about agents' choices. However, the reader may remark that there is a gap between the complexity of the satisfiability problem of a formula in dfSTIT (NP-complete) and the complexity of the satisfiability problem of a formula in dfKSTIT (PSPACE-complete). Of course, the complexity for dfKSTIT can not be improved because the satisfiability problem of $S5_n$ is already PSPACE-complete. An interesting open question is to identify a more expressive fragment of STIT such that its satisfiability problem is PSPACE-complete and such that adding knowledge will not increase the complexity of its satisfiability problem. However, we want to emphasize that the STIT fragment studied in Section 3 already has an interesting expressivity. Indeed, as we have shown in Section 4, it allows to capture subtle aspects of counterfactual reasoning about agents' choices.

We have presented in Section 7 a decidable dynamic extension of the logic dfKSTIT called dfKSTIT$^+$ and we have shown how it can be used in order to capture interesting aspects of dialogue between an artificial agent and a human user. We also postpone to future research an analysis of the complexity of this logic.

An analysis of intensity of counterfactual emotions was also 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 four counterfactual emotions: regret, rejoicing, disappointment and elation. We intend to extend our analysis in the future by studying the counterfactual dimension of "moral" emotions such as guilt and shame. Indeed, as several psychologists have shown (see, e.g., [36]), 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.*

10. References

[1] Adam, C., Herzig, A., Longin, D., 2009. A logical formalization of the OCC theory of emotions. Synthese 168 (2), 201–248.

[2] Ågotnes, T., van der Hoek, W., Wooldridge, M., 2007. Quantified coalition logic. In: Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI'07). AAAI Press, pp. 1181–1186.

[3] Ågotnes, T., van Ditmarsch, H., 2008. Coalitions and announcements. In: Proceedings of the Seventh International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2008). ACM Press, pp. 673–680.

[4] Alur, R., Henzinger, T., 2002. Alternating-time temporal logic. Journal of the ACM 49, 672–713.

[5] Balbiani, P., Gasquet, O., Herzig, A., Schwarzentruber, F., Troquard, N., 2008. Coalition games over Kripke semantics. In: Dégremont, C., Keiff, L., Rückert, H. (Eds.), Festschrift in Honour of Shahid Rahman. College Publications, pp. 1–12.

[6] Balbiani, P., Gasquet, O., Lorini, E., Schwarzentruber, F., 2009. An alternating procedure for the satisfiability problem of $S5_n$. Tech. Rep. IRIT/RT–2009-1-FR, IRIT.

[7] Balbiani, P., Herzig, A., Troquard, N., 2008. Alternative axiomatics and complexity of deliberative STIT theories. Journal of Philosophical Logic 37 (4), 387–406.

[8] Bates, J., 1994. The role of emotion in believable agents. Communications of the ACM 37 (7), 122–125.

[9] Belnap, N., Perloff, M., Xu, M., 2001. Facing the future: agents and choices in our indeterminist world. Oxford.

[10] Blackburn, P., de Rijke, M., Venema, Y., 2001. Modal Logic. Cambridge University Press.

[11] Broersen, J., 2010. CTL.STIT: enhancing ATL to express important multi-agent system verification properties. In: Proceedings 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2010). ACM Press, pp. 215–219.

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

[13] Broersen, J., Herzig, A., Troquard, N., 2007. Normal Coalition Logic and its conformant extension. In: Samet, D. (Ed.), Theoretical Aspects of Rationality and Knowledge (TARK), Brussels, 25/06/2007-27/06/2007. Presses universitaires de Louvain, pp. 91–101.

[14] Castelfranchi, C., 2005. Mind as an anticipatory device: For a theory of expectations. In: Proc. of the First International Symposium on Brain, Vision, and Artificial Intelligence (BVAI 2005). Springer-Verlag, pp. 258–276.

[15] Chandra, A. K., Kozen, D. C., Stockmeyer, L. J., 1981. Alternation. J. ACM 28 (1), 114–133.

[16] Chockler, H., Halpern, J. Y., 2004. Responsibility and blame: A structural-model approach. Journal of Artificial Intelligence Research 22, 93–115.

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

[18] de Rosis, F., Pelachaud, C., Poggi, I., Carofiglio, V., De Carolis, B., 2003. From greta's mind to her face: modelling the dynamics of affective states in a conversational embodied agent. International Journal of Human-Computer Studies 59, 81–118.

[19] Dijk, W. W., Zeelenberg, M., 2002. Investigating the appraisal patterns of regret and disappointment. Motivation and Emotion 26 (4), 321–331.

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

[21] Elliot, C., 1992. The affective reasoner: A process model for emotions in a multiagent system. Ph.D. thesis, Northwestern University, Institute for Learning Sciences.

[22] Frijda, N., 1986. The Emotions. Cambridge University Press.

[23] Frijda, N. H., Kuipers, P., Ter Schure, E., 1989. Relations among emotion, appraisal, and emotional action readiness. Journal of Personality and Social Psychology 57 (2), 212–228.

[24] Gabbay, D. M., Kurucz, A., Wolter, F., Zakharyaschev, M., 2003. Many-Dimensional Modal Logics: Theory and Applications. No. 148 in Studies in Logic and the Foundations of Mathematics. Elsevier, North-Holland.

[25] Gasquet, O., Herzig, A., 1993. Translating non-normal modal logics into normal modal logics. In: Jones, A., Sergot, M. (Eds.), Proceedings International Workshop on Deontic Logic in Computer Science (DEON'94). TANO, Oslo.

[26] Gordon, R. M., 1987. The structure of emotions. Cambridge University Press, Cambridge.

[27] Gratch, J., Marsella, S., 2004. A Domain-independent Framework for modelling Emotions. Journal of Cognitive Systems Research 5 (4), 269–306.

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

[29] Herzig, A., Schwarzentruber, F., 2008. Properties of logics of individual and group agency. In: Proceedings of Advances in Modal Logic 2008. College Publ., pp. 133–149.

[30] Horty, J. F., 2001. Agency and Deontic Logic. Oxford University Press.

[31] Horty, J. F., Belnap, N., 1995. The deliberative STIT: A study of action, omission, and obligation. Journal of Philosophical Logic 24(6), 583–644.

[32] Kahneman, D., 1995. Varieties of counterfactual thinking. In: Roese, N. J., Olson, J. M. (Eds.), What might have been: the social psychology of counterfactual thinking. Erlbaum.

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

[34] Kahneman, D., Tversky, A., 1982. The psychology of preferences. Scientific American 246, 160–173.

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

[36] Lazarus, R. S., 1991. Emotion and adaptation. Oxford University Press, New York.

[37] Loomes, G., Sugden, R., 1982. Regret theory: an alternative theory of rational choice under uncertainty. Economic Journal 92 (4), 805–824.

[38] Loomes, G., Sugden, R., 1987. Testing for regret and disappointment in choice under uncertainty. Economic Journal 97, 118–129.

[39] Lorini, E., Castelfranchi, C., 2007. The cognitive structure of surprise: looking for basic principles. Topoi: an International Review of Philosophy 26 (1), 133–149.

[40] Meyer, J.-J. C., 2006. Reasoning about emotional agents. International Journal of Intelligent Systems 21(6), 601–619.

[41] Meyer, J. J. C., van der Hoek, W., van Linder, B., 1999. A logical approach to the dynamics of commitments. Artificial Intelligence 113(1-2), 1–40.

[42] Ortony, A., Clore, G. L., Collins, A., 1988. The cognitive structure of emotions. Cambridge University Press.

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

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

[45] Plaza, J. A., 1989. Logics of public communications. In: Emrich, M., Pfeifer, M., Hadzikadic, M., Ras, Z. (Eds.), Proceedings of the 4th International Symposium on Methodologies for Intelligent Systems. 201-216.

[46] Reilly, W. S., Bates, J., 1992. Building emotional agents. Tech. rep., CMUCS -92-143, School of Computer science, Canergie Mellon University.

[47] Reisenzein, R., 2009. Emotional experience in the computational belief-desire theory of emotion. Emotion Review 1 (3), 214–222.

[48] Roese, N. J., 1997. Counterfactual thinking. Psychological Bulletin 121 (1), 133–148.

[49] Roese, N. J., Sanna, L. J., Galinsky, A. D., 2005. The mechanics of imagination: automaticity and control in counterfactual thinking. In: Hassin, R. R., Uleman, J. S., Bargh, J. A. (Eds.), The new unconscious. Oxford University Press.

[50] Roseman, I. J., 2001. A model of appraisal in the emotion system. In: Scherer, K. R., Schorr, A., Johnstone, T. (Eds.), Appraisal Processes in Emotion: Theory, Methods, Research. Oxford University Press, Oxford.

[51] Roseman, I. J., Antoniou, A. A., Jose, P. E., 1996. Appraisal determinants of emotions: constructing a more accurate and comprehensive theory. Cognition and Emotion 10, 241–277.

[52] Scherer, K., 2001. Appraisal considered as a process of multilevel sequential checking. In: Scherer, K. R., Schorr, A., Johnstone, T. (Eds.), Appraisal Processes in Emotion: Theory, Methods, Research. Oxford University Press, Oxford.

[53] Scherer, K. R., Schorr, A., Johnstone, T. (Eds.), 2001. Appraisal Processes in Emotion: Theory, Methods, Research. Oxford University Press, Oxford.

[54] Schwarzentruber, F., 2007. Décidabilité et complexité de la logique normale des coalitions. Master's thesis, Univ. Paul Sabatier Toulouse III.

[55] Searle, J., 1983. Intentionality: An Essay in the Philosophy of Mind. Cambridge University Press, New York.

[56] Smith, C., Lazarus, R., 1990. Emotion and adaptation. In: Pervin, J. (Ed.), Handbook of personality: theory & research. Guilford Press, New York.

[57] Solomon, R. C., 1976. The passions. University of Notre Dame Press, Notre Dame.

[58] Steunebrink, B. R., Dastani, M., Meyer, J.-J. C., 2007. A logic of emotions for intelligent agents. In: Proceedings of AAAI'07. AAAI Press, pp. 142–147.

[59] Sugden, R., 1985. Regret, recrimination and rationality. Theory and Decision 19 (1), 77–99.

[60] Taylor, G., 1985. Pride, shame and guilt: the emotions of self-assessment. Oxford University Press, New York.

[61] Troquard, N., 2007. Independent agents in branching time. Ph.D. thesis, Univ. Paul Sabatier Toulouse III & Univ. degli studi di Trento.

[62] Turrini, P., Meyer, J.-J. C., Castelfranchi, C., 2009. Coping with shame and sense of guilt: a dynamic logic account. Journal of Autonomous Agents and Multi-Agent SystemsForthcoming.

[63] van Benthem, J., Liu, F., 2007. Dynamic logic of preference upgrade. Journal of Applied Non-Classical Logics 17 (2), 157–182.

[64] van Benthem, J., Minică, S., 2009. Towards a dynamic logic of questions. In: Proceedings of Second International Workshop on Logic, Rationality and Interaction (LORI-II). Vol. 5834 of LNCS. Springer-Verlag, pp. 27–41.

[65] van der Hoek, W., Jamroga, W., Wooldridge, M., 2005. A logic for strategic reasoning. In: Proceedings of the Fourth International Joint Conference on Autonomous Agens and Multiagent Systems (AAMAS 2005). New York, ACM Press, pp. 157–164.

[66] van der Hoek, W., Wooldridge, M., 2003. Cooperation, knowledge, and time: Alternating-time temporal epistemic logic and its applications. Studia Logica 75, 125–157.

[67] van Ditmarsch, H., van der Hoek, W., Kooi, B., 2007. Dynamic Epistemic Logic. Vol. 337 of Synthese Library. Springer.

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

[69] Zeelenberg, M., van Dijk, W., Manstead, A. S. R., van der Pligt, J., 2000. On bad decisions and disconfirmed expetancies: the psychology of regret and disappointement. Cognition and Emotion 14 (4), 521–541.

[70] Zeelenberg, M., van Dijk, W., van der Pligt, J., Manstead, A. S. R., van Empelen, P., Reinderman, D., 1998. Emotional reactions to the outcomes of decisions: the role of counterfactual thought in the experience of regret and disappointment. Organizational Behavior and Human Decision Processes 75 (2), 117–141.

[71] Zeelenberg, M., van Dijk, W. W., Manstead, A. S. R., 1998. Reconsidering the relation between regret and responsibility. Organizational Behavior and Human Decision Processes 74 (3), 254–272.

11. Annex

11.1. Proof of Proposition 1

Let φ be a formula of $\mathcal{L}_{\mathsf{STIT}}$.

- If *card*(*AGT*) ≤ 2: φ is STIT-satisfiable iff φ is NCL-satisfiable;
- If *card*(*AGT*) ≥ 3: if φ is STIT-satisfiable then φ is NCL-satisfiable. (the converse is false: there exists φ such that φ is NCL-satisfiable and $\neg\varphi$ is STIT-valid.)

PROOF.

Let us prove that a STIT-model is a NCL-model. For notational convenience, we write $\overline{J}$ instead of *AGT* $\setminus J$. Let $\mathcal{M} = (W, \{R_J\}_{J\subseteq AGT}, V)$ be STIT-model and let us prove that it is an NCL model. It suffices to prove that the constraints on a NCL model are true in $\mathcal{M}$. By the constraint 2 of Definition 1, we have $R_{J_1\cup J_2} = \bigcap_{j\in J_1\cup J_2} R_{\{j\}} = \bigcap_{j\in J_1} R_{\{j\}} \cap \bigcap_{j\in J_2} R_{\{j\}} = R_{J_1} \cap R_{J_2}$. So we have $R_{J_1\cup J_2} \subseteq R_{J_1} \cap R_{J_2}$. Now let us prove $R_\emptyset \subseteq R_J \circ R_{AGT\setminus J}$. If $wR_\emptyset v$, then the constraint 3 of Definition 1 gives: $\bigcap_{j\in J} R_{\{j\}}(w) \cap \bigcap_{j\in\overline{J}} R_{\{j\}}(v) \neq \emptyset$. That is to say: $R_J(w) \cap R_{\overline{J}}(v) \neq \emptyset$. So $wR_J \circ R_{\overline{J}}v$.

Now given that a STIT-model is a NCL-model, for all cardinalities of *AGT*, we have the implication "φ is STIT-satisfiable implies φ is NCL-satisfiable".

If *card*(*AGT*) $= 1$, we have that if φ is STIT-satisfiable then φ is NCL-satisfiable. Indeed, both the logic STIT and NCL are just the logic S5 for the operator $[\emptyset]$ because the operator $[1]$ is trivial (as we have $[1]\varphi \leftrightarrow \varphi$).

If *card*(*AGT*) $= 2$, from [29] we have that STIT is exactly the logic $S5^2$ with operators $[1]$ and $[2]$. (we do not care about operators $[\{1,2\}]$ and $[\emptyset]$ because we have the two validities $[\{1,2\}]\varphi \leftrightarrow \varphi$ and $[\emptyset]\varphi \leftrightarrow [1][2]\varphi$.) Concerning NCL, directly from the axiomatics of NCL, we have that NCL is exactly $[S5, S5]$ with operators $[1]$ and $[2]$. As $S5^2 = [S5, S5]$ [24], we have that STIT and NCL have the same satisfiable formulas.

If *card*(*AGT*) ≥ 3, the problem of satisfiability of NCL is in NEXPTIME (see [5] or [54]) whereas the problem of satisfiability of STIT is undecidable (see [29]). So the two logics do not have the same satisfiable formulas.

To sum up, we have:

- If *card*(*AGT*) $= 1$, STIT, $S5$ and NCL are the same logic;
- If *card*(*AGT*) $= 2$, STIT, $S5^2$ and NCL are the same logic;
- If *card*(*AGT*) ≥ 3, we have:
  - STIT and $S5^{card(AGT)}$ are the same logic;
  - If a formula is STIT-satisfiable then it is NCL-satisfiable. However, there exists a NCL-satisfiable formula which is not STIT-satisfiable.

■

11.2. Proof of Lemma 1

Let $\mathcal{M} = (W, R, V)$ be a NCL-model. Let r be a positive integer. Let $w_1, \ldots w_r \in W$ be such that for all $i, j \in \{1, \ldots r\}$, $w_i R_\emptyset w_j$. Let $J_1, \ldots J_r \subseteq \mathit{AGT}$ be such that $i \neq j$ implies $J_i \cap J_j = \emptyset$. We have:

$$\bigcap_{i=1\ldots r} R_{J_i}(w_i) \neq \emptyset.$$

PROOF.

For $r = 0$ the lemma is true by convention. Let us prove the lemma by recurrence on $r \in \mathbb{N}^*$. Let us call $\mathcal{P}(r)$ the statement of the lemma.

- $\mathcal{P}(1)$ is true.
- Let us prove $\mathcal{P}(2)$ because we need it in order to prove $\mathcal{P}(r+1)$ from $\mathcal{P}(r)$.

  Let u and w be in W such that $uR_\emptyset w$. Let $J, K \subseteq \mathit{AGT}$ be two coalitions such that $J \cap K = \emptyset$. As $uR_\emptyset w$, we have $uR_J \circ R_{\overline{J}} w$. And then $uR_J \circ R_K w$. This proves $\mathcal{P}(2)$.
- Now, assume that $\mathcal{P}(r)$ is true for a fixed $r \in \mathbb{N}^*$ and let us prove that $\mathcal{P}(r+1)$ is true. Let $w_1, \ldots, w_r, w_{r+1} \in W$ be such that for all $i, j \in \{1, \ldots r\}$, $w_i R_\emptyset w_j$. Let $J_1, \ldots, J_r, J_{r+1} \subseteq \mathit{AGT}$ be such that $i \neq j$ implies $J_i \cap J_j = \emptyset$. As $P(r)$ is assumed, we can apply it on $(w_1, \ldots, w_r)$ and $(J_1, \ldots, J_r)$ and obtain $\bigcap_{i=1\ldots r} R_{J_i}(w_i) \neq \emptyset$. Let us consider a world w such that $w \in \bigcap_{i=1\ldots r} R_{J_i}(w_i)$. Now consider $R_{\bigcup_{i=1\ldots r} J_i}(w)$ and $R_{J_{r+1}}(w_{r+1})$. By applying $P(2)$ on (w, w_{r+1}), and $(\bigcup_{i=1\ldots r} J_i, J_{r+1})$, we obtain that $R_{\bigcup_{i=1\ldots r} J_i}(w) \cap R_{J_{r+1}}(w_{r+1})$ is not empty, i.e. $R_{\bigcup_{i=1\ldots r} J_i}(w) \cap R_{J_{r+1}}(w_{r+1})$ contains a point v. Note that by constraint 1 of Definition 3 we have $R_{\bigcup_{i=1\ldots r} J_i}(w) \subseteq \bigcap_{i=1\ldots r} R_{J_i}(w)$. As $\bigcap_{i=1\ldots r} R_{J_i}(w) \subseteq \bigcap_{i=1\ldots r} R_{J_i}(w_i)$, we have a point v in $\bigcap_{i=1\ldots r+1} R_{J_i}(w_i)$. In other words, $\mathcal{P}(r+1)$ is true.

Conclusion: We have proved by recurrence that for all $r \geq 1$, $\mathcal{P}(r)$ is true. ■

11.3. Proof of Theorem 2

Let $\varphi \in \mathcal{L}_{df\mathsf{STIT}}$. Then, the following three propositions are equivalent:

1. φ is NCL-satisfiable;
2. φ is STIT-satisfiable;
3. φ is STIT-satisfiable in a polynomial sized product STIT-model.

PROOF.

As "2. implies 1." has been investigated in Proposition 1, we focus here on the proof of "1. implies 3." and we use a selection-of-points argument as in [35]. Let φ be a NCL-satisfiable formula: there exists a NCL–model $\mathcal{M} = (W, V)$ and z_0 such that $\mathcal{M}, z_0 \models \varphi$. The proof is divided in two parts. We first construct from $\mathcal{M}$ a product STIT-model $\mathcal{M}' = (W', V')$. Secondly we ensure that that there exists a point $(Z_0, \ldots, Z_0) \in W'$ such that $\mathcal{M}', (Z_0, \ldots, Z_0) \models \varphi$. Broadly speaking, we take care in the construction 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 at z_0 of $\mathcal{M}$ can also be false in $\mathcal{M}'$.

Notations.

- Elements of W are noted x, y etc. Elements of W' are noted $\vec{x}$, $\vec{x}_0$, $\vec{y}$ etc. x_j stands for the j-th coordinate of $\vec{x}$. Given an element $\vec{x}$, we note $\vec{x}_J = (x_j)_{j\in J}$;
- $(P, \ldots, P)$ denotes the vector $\vec{x}$ where for all $j \in AGT$, $x_j = P$. Given a coalition J, $(P, \ldots, P)_J$ denotes $\vec{x}_J$ where for all $j \in J$, $x_j = P$;
- $SF(\varphi)$ denotes the set of all subformulas of φ. $SF_1(\varphi)$ is the set of all subformulas of φ which are not in the scope of a modal operator and which are of the form $[J]\chi$ where χ is propositional. For instance, if $\varphi = [1]p \wedge \langle\emptyset\rangle[2]q$, then $SF(\varphi) = \{p, q, [1]p, [2]q, \langle\emptyset\rangle[2]q, \varphi\}$ whereas $SF_1(\varphi) = \{[1]p\}$.

Part 1: we define the model $\mathcal{M}'$

. The definition of $\mathcal{M}'$ relies on the following two sets of formulas:

- $Pos = \{\psi \mid \langle\emptyset\rangle\psi \in SF(\varphi)$ and $\mathcal{M}, z_0 \models \langle\emptyset\rangle\psi\} \cup \{Z_0\}$
  where $Z_0 = \bigwedge_{\{[J]\chi \mid [J]\chi \in SF_1(\varphi) \text{ and } \mathcal{M}, z_0 \models [J]\chi\}} [J]\chi$. Formulas in Pos are called *positive formulas*.
- $Neg = \{[J]\chi \mid [J]\chi \in \psi$ and $\langle\emptyset\rangle\psi \in SF(\varphi)$ and $\mathcal{M}, z_0 \not\models \langle\emptyset\rangle\psi\} \cup Neg\_in\_z_0$ where $Neg\_in\_z_0 = \{[J]\chi \mid [J]\chi \in SF_1(\varphi)$ and $\mathcal{M}, z_0 \not\models [J]\chi\}$. Formulas in Neg are called *negative formulas*.

Example 7. *Suppose that* $\varphi = \langle\emptyset\rangle([1]\chi_1 \wedge [\{1,3\}]\chi_2) \wedge \neg\langle\emptyset\rangle([2]\chi_3 \wedge [4]\chi_4) \wedge [5]\chi_5 \wedge [6]\chi_6 \wedge \neg[7]\chi_7 \wedge \neg[8]\chi_8$ *and that* $\mathcal{M}, z_0 \models \varphi$.

Then we have:

- $Z_0 = [5]\chi_5 \wedge [6]\chi_6$;
- $Pos = \{[1]\chi_1 \wedge [\{1,3\}]\chi_2, [5]\chi_5 \wedge [6]\chi_6\}$;

- $Neg\_in\_z_0 = \{[7]\chi_7, [8]\chi_8\}$;
- $Neg = \{[2]\chi_3, [4]\chi_4, [7]\chi_7, [8]\chi_8\}$.

First we define the cartesian product $W' = C^n = C \times C \times \ldots C$ where $C = Pos \cup \{0, \ldots, \mathit{card}(Neg) - 1\}$. Then we introduce few notations and prove the following Lemma 2 that allows us to define V':

- For all $\vec{x} \in W'$, for all $P \in Pos$, we consider the set:

$$Coord^{\vec{x}}_{=P} = \{j \in \mathit{AGT} \mid x_j = P\}.$$

- For all $\vec{x} \in W'$, we consider the set:

$$Pos_{\vec{x}} = \{\chi \mid P \in Pos, [J]\chi \in SF(P), J \subseteq Coord^{\vec{x}}_{=P}\};$$

  Intuitively $Pos_{\vec{x}}$ denotes a set of boolean formulas that must be true in $\vec{x}$ because of positive formulas. Formulas are boolean because of the syntactic restriction over the language (definition of $df\mathsf{STIT}$). For instance let us consider the positive formula $P = [1]p \wedge [\{2, 3\}]q$. The model $\mathcal{M}'$ will be designed so that the point $(P, \ldots, P)$ is the world where P must be true. Indeed, for all $\alpha_2, \ldots \alpha_n \in C$, the set $Pos_{(P,\alpha_2,\ldots\alpha_n)}$ contains p. In the same way, for all $\alpha_1, \alpha_4, \ldots \alpha_n \in C$, the set $Pos_{(\alpha_1,P,P,\alpha_4,\ldots\alpha_n)}$ contains q.

- For all $\vec{x} \in W'$, we consider the formula

$$Boxes_{\vec{x}} = \bigwedge_{\chi \in Pos_{\vec{x}}} \chi.$$

  Intuitively $Boxes_{\vec{x}}$ is the conjunction of all (boolean) formulas which have to be true in $\vec{x}$ because of positive formulas.

- We fix a bijection $i : \{0, \ldots, \mathit{card}(Neg) - 1\} \to Neg$.

  We need such a bijection between integers in $\{0, \ldots, \mathit{card}(Neg) - 1\}$ and Neg in order to use arithmetic operations $+$ and mod (modulo) for defining V'.

- We extend i to a function from W' to Neg in the following way:

$$i(\vec{x}) = i\left(\Sigma_{j \in \{1,\ldots,n\} | x_j \in \{0,\ldots,\mathit{card}(Neg)-1\}} x_j \text{ mod } \mathit{card}(Neg)\right)$$

  where mod is the operation of modulo. Intuitively, $i(\vec{x})$ will correspond to the negative formula $[J]\chi$ which will be false at $\vec{x}$ if there are no contradictions with $Boxes_{\vec{x}}$.

Lemma 2. *For all $\vec{x} \in W'$, there exists $y \in W$ such that $\mathcal{M}, y \models Boxes_{\vec{x}}$.*

PROOF.

We just recall that by definition of Pos, we have that for all $P \in Pos$, $\mathcal{M}, z_0 \models \langle\emptyset\rangle P$. So for all $P \in Pos$, there exists a point $y_P \in W$, such that $\mathcal{M}, y_P \models P$.

Let $\vec{x} \in W'$. In the proof, we first define $y \in W$. Secondly we prove that $\mathcal{M}, y \models Boxes_{\vec{x}}$.

1. First, we define the candidate $y \in W$ of our Lemma 2. As $\mathcal{M}$ is an NCL-model, $\mathcal{M}$ satisfies the *assumption of independence of agents* (Lemma 1). We are simply going to apply Lemma 1 where points are $\{y_P \mid P \in Pos\}$ and sets of agents are $\{Coord^{\vec{x}}_{=P}, P \in Pos\}$. We take care that sets $Coord^{\vec{x}}_{=P}$ are disjoint if P ranges over Pos. [6] Briefly, Lemma 1 leads to:
$$\bigcap_{P \in Pos} R_{Coord^{\vec{x}}_{=P}}(y_P) \neq \emptyset.$$
As this set is not empty, let us consider y in it. Let $y \in \bigcap_{P \in Pos} R_{Coord^{\vec{x}}_{=P}}(y_P)$.
2. We have defined $y \in W$. Now let us prove that $\mathcal{M}, y \models Boxes_{\vec{x}}$. In other words, we are going to prove that for all $\chi \in Pos_{\vec{x}}$, $\mathcal{M}, y \models \chi$.
Let $\chi \in Pos_{\vec{x}}$. By definition of $Pos_{\vec{x}}$, there exists $P \in Pos$ and $[J]\chi \in SF(P)$ such that $J \subseteq Coord^{\vec{x}}_{=P}$. Recall that $\mathcal{M}, \vec{y}_P \models P$ and, consequently, we have $\mathcal{M}, \vec{y}_P \models [J]\chi$. By definition of y, we have $y_P R_{Coord^{\vec{x}}_{=P}} y$. But as $J \subseteq Coord^{\vec{x}}_{=P}$, we have $R_{Coord^{\vec{x}}_{=P}} \subseteq R_J$. So, we have $y_P R_J y$ and, consequently, we have $\mathcal{M}, y \models \chi$. So we have $\mathcal{M}, y \models Boxes_{\vec{x}}$.

■

Finally, we define $V' = f \circ V$ where f is a mapping from W' to W defined by:

- $f(Z_0, \ldots, Z_0) = z_0$;
- For all $\vec{x} \in W'$ such that $\vec{x} \neq (Z_0, \ldots, Z_0)$, $i(\vec{x})$ is of the form $[J]\chi \in Neg$.
  - If there exists $y \in W$ such that $\mathcal{M}, y \models \neg\chi \wedge Boxes_{\vec{x}}$ then $f(\vec{x}) \stackrel{\text{def}}{=} y$.
  - Else, we choose a world y in W such that $\mathcal{M}, y \models Boxes_{\vec{x}}$ (such a world exists because of Lemma 2) and we define $f(\vec{x}) \stackrel{\text{def}}{=} y$.

Clearly, $\mathcal{M}' = (W', V')$ is a product STIT-model and its size is polynomial. As $V' = f \circ V$, we have immediately the following Lemma useful for the Part 2 of the proof.

[6] Indeed, for all $P, Q \in Pos$, $Coord^{\vec{x}}_{=P} \cap Coord^{\vec{x}}_{=Q} \neq \emptyset$, implies that there exists $j \in Coord^{\vec{x}}_{=P} \cap Coord^{\vec{x}}_{=Q}$. By definition of $Coord^{\vec{x}}_{=P}$, we have $x_j = P$. In the same way, by definition of $Coord^{\vec{x}}_{=P}$, we have $x_j = Q$. Hence $P = Q$.

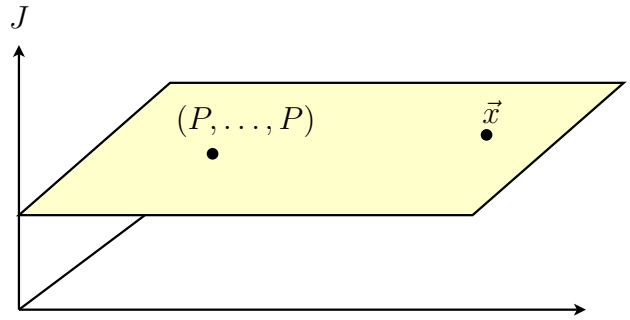

Figure 7: Model $\mathcal{M}'$: a point $(P, \ldots, P)$ and the subspace of all points $\vec{x}$ such that $\vec{x}_J = (P, \ldots, P)_J$, that is, the subspace of worlds of $\mathcal{M}'$ where every agent in J performs action "P".

Lemma 3. *For all $\vec{x} \in W'$, $\mathcal{M}', \vec{x} \models Boxes_{\vec{x}}$.*

PROOF.

Let $\vec{x} \in W'$. By definition of f, $\mathcal{M}, f(\vec{x}) \models Boxes_{\vec{x}}$. But recall that $V' = f \circ V$: in particular, we have $V'(\vec{x}) = V(f(\vec{x}))$. Recall also that $Boxes_{\vec{x}}$ is a boolean formula. So we obtain $\mathcal{M}, \vec{x} \models Boxes_{\vec{x}}$. ■

Part 2 of the proof : we prove $\mathcal{M}', (Z_0, \ldots, Z_0) \models \varphi$

. We prove the following two facts:

Fact 1. for all $\langle\emptyset\rangle\psi$ of φ, we have $\mathcal{M}, z_0 \models \langle\emptyset\rangle\psi$ iff $\mathcal{M}', (Z_0, \ldots, Z_0) \models \langle\emptyset\rangle\psi$.

Fact 2. for all $[J]\chi \in SF_1(\varphi)$, we have $\mathcal{M}, z_0 \models [J]\chi$ iff $\mathcal{M}', (Z_0, \ldots, Z_0) \models [J]\chi$.

$\boxed{\Rightarrow \text{ of Fact 1 and } \Rightarrow \text{ of Fact 2}}$ In order to prove it, it suffices to prove that for all $P \in Pos$ we have $\mathcal{M}', (P, \ldots P) \models P$. Let $P \in Pos$. P is a conjunction of formulas of the form $[J]\chi$ where χ is a Boolean formula. Let $[J]\chi \in SF(P)$. We have to show that for all $\vec{x} \in W'$ such that $\vec{x}_J = (P, \ldots, P)_J$, we have $\mathcal{M}, \vec{x} \models \chi$. The situation is drawn in Fig. 7. But for those $\vec{x}$ such that $\vec{x}_J = (P, \ldots, P)_J$, we have $J \subseteq Coord^{\vec{x}}_{=P}$. So $\chi \in Pos_{\vec{x}}$ implies that $\models Boxes_{\vec{x}} \rightarrow \chi$. But, by Lemma 3, we have $\mathcal{M}', \vec{x} \models Boxes_{\vec{x}}$ and this leads to $\mathcal{M}', \vec{x} \models \chi$. Finally, $\mathcal{M}', (P, \ldots, P) \models [J]\chi$. Therefore we have $\mathcal{M}', (P, \ldots P) \models P$.

$\boxed{\Leftarrow \text{ of Fact 1}}$ Let $N = [J_1]\chi_1 \wedge \ldots [J_k]\chi_k$ be such that $\langle\emptyset\rangle N \in SF(\varphi)$ and $\mathcal{M}, z_0 \not\models \langle\emptyset\rangle N$. Let us prove that for all $\vec{x}_0 \in W'$, $\mathcal{M}', \vec{x}_0 \models \neg N$. We suggest the reader to look at the Fig. 8 during this part.

Consider $y_0 = f(\vec{x}_0) \in W$. By definition of f, we have $\mathcal{M}, y_0 \models Boxes_{\vec{x}_0}$. We also have $\mathcal{M}, y_0 \models \neg N$. So, there is $i \in \{1, \ldots, k\}$ such that $\mathcal{M}, y_0 \not\models [J_i]\chi_i$. Notice that $[J_i]\chi_i$ belongs to Neg.

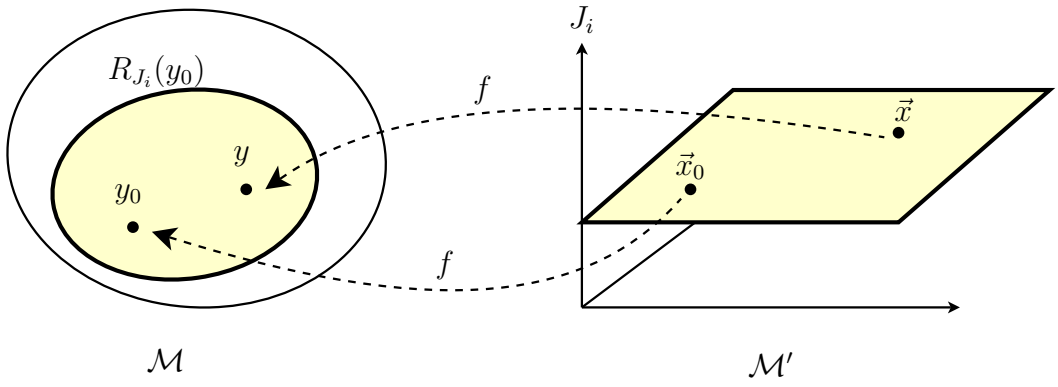

Figure 8: Case 2. (a) in the part 2 of the Proof of Theorem 2

Now we are going to prove that $\mathcal{M}', \vec{x}_0 \not\models [J_i]\chi_i$. We are going to define a vector $\vec{x} \in W'$ such that $\vec{x}_0 R'_{J_i} \vec{x}$ and $\mathcal{M}', \vec{x} \models \neg\chi_i$. As depicted in the Fig. 8, we want that J_i performs the same joint action both in $\vec{x}_0$ and in $\vec{x}$.

The case where $J_i = AGT$ is trivial: we take $\vec{x} = \vec{x_0}$. Else, let j_0 be an arbitrary agent in $\overline{J_i}$ and $\vec{x} \in W'$ be the candidate vector such that:

- $\vec{x}_{J_i} = \vec{x_0}_{J_i}$;
- $x_j = 0$ for all $j \in \overline{J_i} \setminus \{j_0\}$;
- $x_{j_0} = i^{-1}([J_i]\neg\chi_i) - \Sigma_{j \in AGT | j \neq j_0 \text{ and } x_j \in \{0,\ldots,card(Neg)-1\}} x_j \bmod N$.

Now we check that $\mathcal{M}', \vec{x} \models \neg\chi_i$. As $\mathcal{M}, y_0 \models \langle J_i\rangle\neg\chi_i$, there exists $y \in W$ such that $yR_{J_i}y_0$ and $\mathcal{M}, y \models \neg\chi_i$. Notice that $\mathcal{M}, y \models Boxes_{\vec{x}}$. Indeed, $Boxes_{\vec{x}}$ only contains subformulas χ_1 such that $[K]\chi_1$ is a subformula of Pos where $K \subseteq J_i$. (because only coordinates in J_i of $\vec{x}$ are in Pos; the others are integer). Then we have $\models Boxes_{\vec{x_0}} \rightarrow Boxes_{\vec{x}}$. Hence $\mathcal{M}, y \models Boxes_{\vec{x}}$. To sum up, we have $\mathcal{M}, y \models Boxes_{\vec{x}} \wedge \neg\chi_i$. So, as $i(\vec{x}) = [J_i]\neg\chi_i$, by definition of f we have that $f(\vec{x})$ is a such point y where $\mathcal{M}, y \models Boxes_{\vec{x}} \wedge \neg\chi_i$. Finally, by definition of V', we have $\mathcal{M}', \vec{x} \models \neg\chi_i$.

$\boxed{\Leftarrow \text{ of Fact 2}}$ Let us prove now that $\mathcal{M}', (Z_0, \ldots, Z_0) \models Neg\_in\_z_0$. More precisely we prove that for all $[J]\chi \in Neg\_in\_z_0$, $\mathcal{M}', (Z_0, \ldots, Z_0) \models \langle J\rangle\neg\chi$. We know that $\mathcal{M}, z_0 \models \langle J\rangle\neg\chi$. So there exists $y \in W$ such that $yR_J z_0$ and $\mathcal{M}, y \models \neg\chi$. The case $J = AGT$ is trivial. Let us consider $j_0 \in \overline{J}$ and let us define the candidate vector $\vec{x}$:

- $\vec{x}_J = (Z_0, \ldots, Z_0)_J$;
- $x_j = 0$ for all $j \in \overline{J} \setminus \{j_0\}$;
- $\vec{x}_{j_0} = i^{-1}([J]\chi)$.

Let us check that $\mathcal{M}', \vec{x} \models \neg\chi$. Notice that $Boxes_{\vec{x}}$ only contains Boolean formulas χ' where formulas $[J']\chi'$ are subformulas of Z_0, where $J' \subseteq J$. Hence $\mathcal{M}, y \models Boxes_{\vec{x}}$.

Furthermore, $\mathcal{M}, y \models \neg\chi$. So by definition of f, as $i(\vec{x}) = [J]\chi$, we have that $f(\vec{x})$ is a point y such that $\mathcal{M}, y \models \neg\chi \wedge Boxes_{\vec{x}}$. By definition of V', $\mathcal{M}', \vec{x} \models \neg\chi$.

The conclusion of the proof is left to the reader.

■

11.4. Proof of Corollary 1

Deciding if a formula in $\mathcal{L}_{df\mathsf{STIT}}$ is STIT-satisfiable is NP-complete.

PROOF.

SAT is reducible to the STIT-satisfiability problem of a formula in $\mathcal{L}_{df\mathsf{STIT}}$. Thus deciding if a formula in $\mathcal{L}_{df\mathsf{STIT}}$ is STIT-satisfiable is NP-hard. Now let us see that it is in NP.

According to Theorem 2, if a formula φ is STIT-satisfiable, φ is satisfiable in a polynomial-sized STIT-model. So a non-deterministic algorithm to solve the satisfiability problem can be as follows:

- we guess a polynomial-sized model $\mathcal{M}' = (W', V')$ and a world $\vec{x} \in W'$;
- we check whether $\mathcal{M}', \vec{x} \models \varphi$ holds or not.

Note that checking whether $\mathcal{M}', \vec{x} \models \varphi$ or not can be done in polynomial time in the size of $\mathcal{M}'$ and in the length of φ. As the size of $\mathcal{M}'$ is polynomial in the length of φ, checking whether $\mathcal{M}', \vec{x} \models \varphi$ or not can be done in polynomial time in the size of φ.

■

11.5. Proof of Corollary 2

A formula φ in $\mathcal{L}_{df\mathsf{STIT}}$ is STIT-valid iff we have $\vdash_{\mathsf{NCL}} \varphi$.

PROOF.

We have:

- for all formulas $\varphi \in \mathcal{L}$, $\models_{\mathsf{NCL}} \varphi$ iff $\vdash_{\mathsf{NCL}} \varphi$ (Theorem 1);
- for all formulas $\varphi \in \mathcal{L}_{df\mathsf{STIT}}$, $\models_{\mathsf{STIT}} \varphi$ iff $\models_{\mathsf{NCL}} \varphi$. (Theorem 2).

Hence: for all formulas $\varphi \in \mathcal{L}_{df\mathsf{STIT}}$, $\models_{\mathsf{STIT}} \varphi$ iff $\vdash_{\mathsf{NCL}} \varphi$. ■

11.6. Proof of Theorem 3

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

PROOF.

$\boxed{\textit{card}(\textit{AGT}) = 1}$

Let us consider the case *card*(*AGT*) $= 1$. In this case 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 a $\mathtt{K}_1$ operator can not appear after a $[\emptyset]$ operator, we can prove by a selected points argument that if a dfKSTIT -formula is KSTIT-satisfiable, then it is in a polynomial sized model (in [35], it is done for $S5$).

$\boxed{\textit{card}(\textit{AGT}) \geq 2}$

Let us consider the case $\textit{card}(\textit{AGT}) \geq 2$. Recall that the satisfiability problem of $S5_{\textit{card}(\textit{AGT})}$ is PSPACE-hard [28]. But the logic $S5_{\textit{card}(\textit{AGT})}$ is embedded into dfKSTIT . So deciding if a dfKSTIT -formula φ is KSTIT-satisfiable is also PSPACE-hard.

Now let us prove that the KSTIT-satisfiability problem of a given dfKSTIT -formula is in PSPACE. As APTIME = PSPACE [15], it is sufficient to prove that this problem is in APTIME. The Figure 9 shows an alternating procedure sat(Σ, i) where Σ is a set of dfKSTIT -formulas and $i \in \textit{AGT}$. For all $i \in \textit{AGT}$, when each formula of Σ starts with $\mathtt{K}_i$ or $\neg\mathtt{K}_i$, then the call sat(Σ, i) succeeds if and only if the set of formulas Σ is KSTIT-satisfiable, that is, the conjunction of all formulas $\varphi \in \Sigma$ is satisfiable. Note that Φ is satisfiable iff $\mathtt{K}_1\Phi$ is satisfiable. Thus, in order to check if Φ is satisfiable, we call sat($\{\mathtt{K}_1\Phi\}$, 1). For all formulas φ, we define the set $CL(\varphi) = SF(\varphi) \cup \{\neg\psi \mid \psi \in SF(\varphi)\}$. $CL(\varphi)$ contains all the subformulas of φ and their negations. For all sets of formulas Σ, we define $CL(\Sigma) = \bigcup_{\varphi\in\Sigma} CL(\varphi)$.

The procedure sat(Σ, i) is inspired by the algorithms of the satisfiability problem for $S5_n$ given in [28] and in [6]. It checks the satisfiability of a set of formulas Σ where all formulas of Σ starts with $\mathtt{K}_i$ or $\neg\mathtt{K}_i$ by first constructing an E_i-equivalence class represented by the set of subsets of $CL(\Sigma)$. A subset of $CL(\Sigma)$ represents all formulas that are true in a given world of the E_i-equivalence class.

We require one of the worlds to satisfy Σ, that is, we require that there exists $S \in \beta$ such that $\Sigma \subseteq S$. We then check that all constraints on agent i's knowledge are satisfied: steps 1, 2 and 3 in the algorithm of Figure 9. We also check that constraints on other agents' knowledge are satisfied in worlds of the E_i-equivalence class: step 4. We check Boolean constraints: steps 5, 6 and 7. We finally check that at each world of the E_i-equivalence class all $\mathcal{L}_{\mathsf{STIT}}$-subformulas supposed to be true are together satisfiable: step 8. This verification can run non-deterministically in polynomial time thanks to Theorem 1.

Finally we continue the construction of the model: at every point S' of the E_i-equivalence class and for all agents j, we check if all constraints due to all subformulas of the form $\mathtt{K}_j\theta$ and $\neg\mathtt{K}_j\theta$ can be together satisfiable. Let $l(\Sigma)$ be the number of epistemic modal operators in the formulas of Σ that have the maximal number of epistemic modal operators. Note that $l(\{\mathtt{K}_j\theta \in S'\} \cup \{\neg\mathtt{K}_j\theta \in S'\}) < l(\Sigma)$ so that the termination is granted. During all the recursive call of the algorithm sat($\{\mathtt{K}_1\Phi\}, 1$), we only work with subformulas of $\mathtt{K}_1\Phi$. The algorithm runs in polynomial time.

■

function sat(Σ, i)

(∃) **choose** β a set of at most n subsets of $CL(\Sigma)$ such that there exists $S \in \beta$ such that $\Sigma \subseteq S$, where n is the number of operators $\mathtt{K}_i$ appearing in Σ.
Check $\mathtt{K}_i\psi$, $\neg\mathtt{K}_i\psi$, $\mathtt{K}_j\psi$, Boolean coherence and STIT coherence:

1. for all $S, S' \in \beta$, $\mathtt{K}_i\psi \in S$ iff $\mathtt{K}_i\psi \in S'$;
2. for all $S \in \beta$, $\mathtt{K}_i\psi \in S$ implies $\psi \in S$;
3. for all $S \in \beta$, $\neg\mathtt{K}_i\psi \in S$ iff there exists $S' \in \beta$ such that $\neg\psi \in S'$;
4. for all $S \in \beta$, for all $j \neq i$, $\mathtt{K}_j\psi \in S$ implies $\psi \in S$;
5. $\psi_1 \wedge \psi_2 \in S$ iff ($\psi_1 \in S$ and $\psi_2 \in S$);
6. $\psi_1 \vee \psi_2 \in S$ iff ($\psi_1 \in S$ or $\psi_2 \in S$);
7. for all $S \in \beta$, $\psi \in S$ xor $\neg\psi \in S$.
8. check that $\bigwedge_{\psi \in S | \psi \in \mathcal{L}_{\mathsf{STIT}}} \psi$ is STIT-satisfiable.

(∀) **choose** $S' \in \beta$
(∀) **choose** $j \in AGT \setminus \{i\}$
if there exists a formula of the form $\neg\mathtt{K}_j\psi$ in S',
  call sat($\{\mathtt{K}_j\theta \in S'\} \cup \{\neg\mathtt{K}_j\theta \in S'\}$, j)
endIf

endFunction

Figure 9: An algorithm for the KSTIT-satisfiability problem of a given set of dfKSTIT -formulas Σ

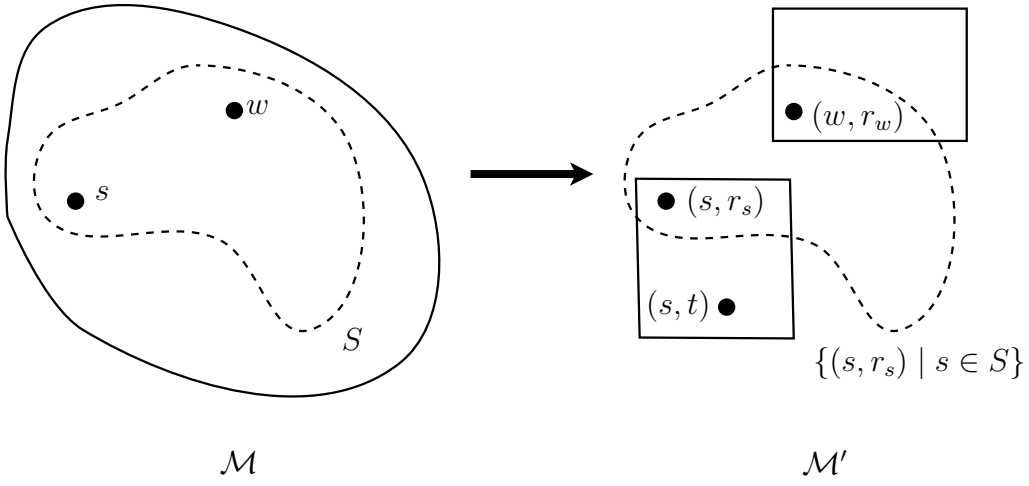

Figure 10: From KNCL-model $\mathcal{M}$ to the KSTIT-model $\mathcal{M}'$

11.7. Proof of Theorem 4

Let φ be a formula of $\mathcal{L}_{df\mathsf{KSTIT}}$. We have equivalence between:

- φ is satisfiable in KNCL;
- φ is satisfiable in KSTIT.

PROOF.

Unfortunately, the general results about completeness of fusion of logics given in [24] can not be applied here because we are dealing with syntactic fragments.

$\boxed{\Uparrow}$ We can prove that a KSTIT-model is a KNCL-model. The proof is similar to the proof of Proposition 1.

$\boxed{\Downarrow}$ Let Φ be a formula of $\mathcal{L}_{df\mathsf{KSTIT}}$ satisfiable in a KNCL-model, i.e. suppose that there exists a KNCL-model $\mathcal{M} = (W, \{R_J\}_{J \subseteq AGT}, \{E_i\}_{i \in AGT}, V)$ and $w \in W$ such that $\mathcal{M}, w \models \Phi$. We are going to define a KSTIT-model $\mathcal{M}'$ from $\mathcal{M}$ satisfying Φ. In order to do this, we are simply going to replace each "NCL-model" in the model $\mathcal{M}$ by an equivalent STIT-model given by the Theorem 2 as shown in the Fig. 10.

Consider the set:

$$S = \{s \in W \mid \text{ there exists a finite sequence } i_1, \ldots i_n \in AGT \text{ such that } wE_{i_1} \circ \ldots \circ E_{i_n} s\}$$

where $\circ$ is the standard composition operation over binary relation.

Let STIT-SF(Φ) be the set of all subformulas of Φ that do not contain an epistemic operator $\mathtt{K}_i$.

For all $s \in S$, we define

$$\psi_s = \bigwedge_{\varphi \in \text{STIT-SF}(\Phi) | \mathcal{M}, s \models \varphi} \varphi \wedge \bigwedge_{\varphi \in \text{STIT-SF}(\Phi) | \mathcal{M}, s \not\models \varphi} \neg\varphi.$$

We have $\mathcal{M}, s \models \psi_s$. As Φ is in the fragment dfKSTIT , we have that ψ_s is in the fragment dfSTIT . So we can apply the Theorem 2: it gives the existence of a STIT-model $\mathcal{M}_s = (W_s, \{R_{sJ}\}_{J \subseteq AGT}, V_s)$ and $r_s \in W_s$ such that $\mathcal{M}_s, r_s \models \psi_s$.

Now we define $\mathcal{M}' = (W', \{R'_J\}_{J \subseteq AGT}, \{E'_i\}_{i \in AGT}, V')$ as follows:

- $W' = \{(s,t) \mid s \in S, t \in W_s\}$;
- $R'_J = \{\langle (s,t), (s,t') \rangle \in W' \times W' \mid (t,t') \in R_{sJ}\}$;
- $E'_i = \{\langle (s,r_s), (s',r_{s'}) \rangle \mid (s,s') \in E_i\} \cup \{\langle (s,t), (s,t) \rangle \mid (s,t) \in W'\}$;
- $V'(s,t) = V_s(t)$.

Now we can prove by induction that for all subformulas φ of Φ, we have that for all $s \in S$, $\mathcal{M}, s \models \varphi$ iff $\mathcal{M}', (s, r_s) \models \varphi$.

(dfSTIT) If φ is a see-to-it formula or is of the form $\langle\emptyset\rangle\psi$ where ψ is a see-to-it-formula, then φ does not contain any epistemic operator. Hence by definition of ψ_s we have that ψ_s contains either φ or $\neg\varphi$. So, by definition of the STIT-model $\mathcal{M}_s$, we have $\mathcal{M}, s \models \varphi$ iff $\mathcal{M}_s, r_s \models \varphi$. And $\mathcal{M}_s, r_s \models \varphi$ is equivalent to $\mathcal{M}', (s, r_s) \models \varphi$ by definition of R'_J and V'.

(boolean cases) Boolean cases are left to the reader.

($\mathtt{K}_i\varphi$) Let us consider a subformula of the form $\mathtt{K}_i\varphi$. We have $\mathcal{M}, s \models \mathtt{K}_i\varphi$ iff for all $s' \in W$ such that sE_is' we have $\mathcal{M}, s' \models \varphi$. By definition of S, this is equivalent to: for all $s' \in S$ such that sE_is' we have $\mathcal{M}, s' \models \varphi$. By induction, this is equivalent to the fact that for all $s' \in S$ such that sE_is' we have $\mathcal{M}', (s', r_{s'}) \models \varphi$. By definition of E'_i this is equivalent to $\mathcal{M}', (s, r_s) \models \mathtt{K}_i\varphi$.

■

11.8. Proof of Validity (22) in Section 7.1

$$\models_{\mathsf{KSTIT}} \mathcal{KB}^* \rightarrow \mathtt{K}_t\mathtt{REJOICE}_u pass_u$$

PROOF.

We give a syntactic proof of the previous validity. We show that we have

$$\vdash_{\mathsf{KNCL}} \mathcal{KB}^* \rightarrow \mathtt{K}_t\mathtt{REJOICE}_u pass_u$$

by applying the axioms and rules of inference of KNCL. Then,

$$\models_{\mathsf{KSTIT}} \mathcal{KB}^* \rightarrow \mathtt{K}_t\mathtt{REJOICE}_u pass_u$$

follows from Corollary 3 and the fact that $\mathcal{KB}^* \rightarrow \mathtt{K}_t\mathtt{REJOICE}_u pass_u \in \mathcal{L}_{df\mathsf{KSTIT}}$.

1. $\vdash_{\mathsf{KNCL}} \neg[\emptyset]([u]\neg studied_u \rightarrow [u]\neg pass_u) \rightarrow \langle\emptyset\rangle([u]\neg studied_u \wedge \neg[u]\neg pass_u)$
2. $\vdash_{\mathsf{KNCL}} \langle\emptyset\rangle([u]\neg studied_u \wedge \neg[u]\neg pass_u) \rightarrow \langle\emptyset\rangle([u]\neg studied_u \wedge \langle u\rangle pass_u)$
3. $\vdash_{\mathsf{KNCL}} \langle\emptyset\rangle([u]\neg studied_u \wedge \langle u\rangle pass_u) \rightarrow \langle\emptyset\rangle([u][u]\neg studied_u \wedge \langle u\rangle pass_u)$
   by the standard S5 validity $[J]\chi \leftrightarrow [J][J]\chi$
4. $\vdash_{\mathsf{KNCL}} \langle\emptyset\rangle([u][u]\neg studied_u \wedge \langle u\rangle pass_u) \rightarrow \langle\emptyset\rangle\langle u\rangle([u]\neg studied_u \wedge pass_u)$
   by necessitation rule for $[\emptyset]$ and the standard validity $([J]\chi_1 \wedge \langle J\rangle\chi_2) \rightarrow \langle J\rangle(\chi_1 \wedge \chi_2)$
5. $\vdash_{\mathsf{KNCL}} \langle J\rangle\varphi \rightarrow (\langle J\rangle\varphi \wedge \langle\emptyset\rangle\varphi)$
   by the NCL Axiom *Mon*
6. $\vdash_{\mathsf{KNCL}} \langle J\rangle\varphi \rightarrow \langle\emptyset\rangle\varphi$
   by 5
7. $\vdash_{\mathsf{KNCL}} \langle\emptyset\rangle\langle u\rangle([u]\neg studied_u \wedge pass_u) \rightarrow \langle\emptyset\rangle\langle\emptyset\rangle([u]\neg studied_u \wedge pass_u)$
   by 6, necessitation rule for $[\emptyset]$ and the standard validity $([\emptyset]\chi_1 \wedge \langle\emptyset\rangle\chi_2) \rightarrow \langle\emptyset\rangle(\chi_1 \wedge \chi_2)$
8. $\vdash_{\mathsf{KNCL}} \langle\emptyset\rangle\langle\emptyset\rangle([u]\neg studied_u \wedge pass_u) \rightarrow \langle\emptyset\rangle([u]\neg studied_u \wedge pass_u)$
   by the standard S5 validity $\langle\emptyset\rangle\chi \leftrightarrow \langle\emptyset\rangle\langle\emptyset\rangle\chi$
9. $\vdash_{\mathsf{KNCL}} \langle\emptyset\rangle([u]\neg studied_u \wedge pass_u) \rightarrow \neg[\emptyset]([u]\neg studied_u \rightarrow \neg pass_u)$
10. $\vdash_{\mathsf{KNCL}} \neg[\emptyset]([u]\neg studied_u \rightarrow [u]\neg pass_u) \rightarrow \neg[\emptyset]([u]\neg studied_u \rightarrow \neg pass_u)$
    from 1-4 and 7-9
11. $\vdash_{\mathsf{KNCL}} [\emptyset]([u]\neg studied_u \rightarrow \neg pass_u) \rightarrow [\emptyset]([u]\neg studied_u \rightarrow [u]\neg pass_u)$
    from 10
12. $\vdash_{\mathsf{KNCL}} (\langle\emptyset\rangle[u]\neg studied_u \wedge [\emptyset]([u]\neg studied_u \rightarrow [u]\neg pass_u)) \rightarrow \langle\emptyset\rangle[u]\neg pass_u$
    by the standard validity $([\emptyset]\chi_1 \wedge \langle\emptyset\rangle\chi_2) \rightarrow \langle\emptyset\rangle(\chi_1 \wedge \chi_2)$
13. $\vdash_{\mathsf{KNCL}} (\langle\emptyset\rangle[u]\neg studied_u \wedge [\emptyset]([u]\neg studied_u \rightarrow \neg pass_u)) \rightarrow \langle\emptyset\rangle[u]\neg pass_u$
    from 11,12
14. $\vdash_{\mathsf{KNCL}} \mathtt{K}_u(\langle\emptyset\rangle[u]\neg studied_u \wedge [\emptyset]([u]\neg studied_u \rightarrow \neg pass_u)) \rightarrow \mathtt{K}_u\langle\emptyset\rangle[u]\neg pass_u$
    by 13, Axiom K and necessitation rule for $\mathtt{K}_u$
15. $\vdash_{\mathsf{KNCL}} \langle\emptyset\rangle[u]\neg pass_u \rightarrow \langle AGT \setminus \{u\}\rangle\langle u\rangle[u]\neg pass_u$
    by the NCL Axiom *Elim*$(\emptyset)$
16. $\vdash_{\mathsf{KNCL}} \langle u\rangle[u]\neg pass_u \rightarrow [u]\neg pass_u$
    by Axiom 5 for $[J]$
17. $\vdash_{\mathsf{KNCL}} [u]\neg pass_u \rightarrow \neg pass_u$
    by Axiom T for $[u]$

18. $\vdash_{\mathsf{KNCL}} \langle u\rangle[u]\neg pass_u \rightarrow \neg pass_u$
from 16,17

19. $\vdash_{\mathsf{KNCL}} \langle AGT \setminus \{u\}\rangle\langle u\rangle[u]\neg pass_u \rightarrow \langle AGT \setminus \{u\}\rangle\neg pass_u$
by 18, necessitation rule for $[J]$ and the standard validity $([J]\chi_1 \wedge \langle J\rangle\chi_2) \rightarrow \langle J\rangle(\chi_1 \wedge \chi_2)$

20. $\vdash_{\mathsf{KNCL}} \langle\emptyset\rangle[u]\neg pass_u \rightarrow \langle AGT \setminus \{u\}\rangle\neg pass_u$
from 15,19

21. $\vdash_{\mathsf{KNCL}} \mathtt{K}_u\langle\emptyset\rangle[u]\neg pass_u \rightarrow \mathtt{K}_u\langle AGT \setminus \{u\}\rangle\neg pass_u$
by 20, Axiom K and necessitation rule for $\mathtt{K}_u$

22. $\vdash_{\mathsf{KNCL}} \mathtt{K}_u(\langle\emptyset\rangle[u]\neg studied_u \wedge [\emptyset]([u]\neg studied_u \rightarrow \neg pass_u)) \rightarrow \mathtt{K}_u\langle AGT \setminus \{u\}\rangle\neg pass_u$
from 14,21

23. $\vdash_{\mathsf{KNCL}} \mathtt{K}_t\mathtt{K}_u(\langle\emptyset\rangle[u]\neg studied_u \wedge [\emptyset]([u]\neg studied_u \rightarrow \neg pass_u)) \rightarrow \mathtt{K}_t\mathtt{K}_u\langle AGT \setminus \{u\}\rangle\neg pass_u$
by 22, Axiom K and necessitation rule for $\mathtt{K}_t$

24. $\vdash_{\mathsf{KNCL}} (\mathtt{K}_t\mathtt{K}_u\langle\emptyset\rangle[u]\neg studied_u \wedge \mathtt{K}_t\mathtt{K}_u[\emptyset]([u]\neg studied_u \rightarrow \neg pass_u)) \rightarrow \mathtt{K}_t\mathtt{K}_u\langle AGT\setminus\{u\}\rangle\neg pass_u$
by 23, and the standard validity $\mathtt{K}_i(\chi_1 \wedge \chi_2) \leftrightarrow (\mathtt{K}_i\chi_1 \wedge \mathtt{K}_i\chi_2)$

25. $\vdash_{\mathsf{KNCL}} (\mathtt{K}_t\mathtt{DES}_u\neg pass_u \wedge \mathtt{K}_t\mathtt{K}_u pass_u \wedge \mathtt{K}_t\mathtt{K}_u\langle\emptyset\rangle[u]\neg studied_u \wedge$
$\mathtt{K}_t\mathtt{K}_u[\emptyset]([u]\neg studied_u \rightarrow \neg pass_u)) \rightarrow \mathtt{REJOICE}_u pass_u$
from 24, definition of $\mathtt{REJOICE}_i\chi$ and the standard validity $(\mathtt{K}_i\chi_1\wedge\ldots\wedge\mathtt{K}_i\chi_n) \leftrightarrow \mathtt{K}_i(\chi_1\wedge\ldots\wedge\chi_n)$

26. $\vdash_{\mathsf{KNCL}} \mathcal{KB}^* \rightarrow \mathtt{REJOICE}_u pass_u$
from 25 and
$\mathcal{KB}^* = \mathtt{K}_t\mathtt{DES}_u pass_u \wedge \mathtt{K}_t\mathtt{K}_u pass_u \wedge \mathtt{K}_t\mathtt{K}_u\langle\emptyset\rangle[u]\neg studied_u \wedge \mathtt{K}_t\mathtt{K}_u[\emptyset]([u]\neg studied_u \rightarrow \neg pass_u)$

■

11.9. Proof of Proposition 2

The following schemata are KSTIT^+-valid:

$$
\begin{array}{ll}
(Red_p) & [\|\theta\|]p \leftrightarrow p \\
(Red_\neg) & [\|\theta\|]\neg\varphi \leftrightarrow \neg[\|\theta\|]\varphi \\
(Red_\wedge) & [\|\theta\|](\varphi_1 \wedge \varphi_2) \leftrightarrow ([\|\theta\|]\varphi_1 \wedge [\|\theta\|]\varphi_2) \\
(Red_{[J]}) & [\|\theta\|][J]\varphi \leftrightarrow [J][\|\theta\|]\varphi \\
(Red_{\mathtt{K}_i}) & [\|\theta\|]\mathtt{K}_i\varphi \leftrightarrow ((\theta \rightarrow \mathtt{K}_i(\theta \rightarrow [\|\theta\|]\varphi)) \wedge (\neg\theta \rightarrow \mathtt{K}_i(\neg\theta \rightarrow [\|\theta\|]\varphi)))
\end{array}
$$

PROOF.

We here just prove the validity of reduction axioms $Red_{[J]}$ and $Red_{\mathtt{K}_i}$. The proofs of the other reduction axioms go as in PAL [67].

$M, w \models [\|\theta\|]\mathtt{K}_i\varphi$,
IFF (if $M, w \models \theta$ then $M, w \models [\|\theta\|]\mathtt{K}_i\varphi$) and (if $M, w \models \neg\theta$ then $M, w \models [\|\theta\|]\mathtt{K}_i\varphi$),
IFF (if $M, w \models \theta$ then $M^{\|\theta\|}, w \models \mathtt{K}_i\varphi$) and (if $M, w \models \neg\theta$ then $M^{\|\theta\|}, w \models \mathtt{K}_i\varphi$),
IFF (if $M, w \models \theta$ then for all $v \in W$ such that $wE_i^{\|\theta\|}v$, $M^{\|\theta\|}, v \models \varphi$) and
(if $M, w \models \neg\theta$ then for all $v \in W$ such that $wE_i^{\|\theta\|}v$, $M^{\|\theta\|}, v \models \varphi$),
IFF (if $M, w \models \theta$ then for all $v \in W$ such that $wE_i^{\|\theta\|}v$, $M, v \models [\|\theta\|]\varphi$) and

(if $M, w \models \neg\theta$ then for all $v \in W$ such that $wE_i^{\|\theta\|}v$, $M, v \models [\|\theta\|]\varphi$),
IFF (if $M, w \models \theta$ then for all $v \in W$ such that wE_iv and $M, v \models \theta$, $M, v \models [\|\theta\|]\varphi$) and
(if $M, w \models \neg\theta$ then for all $v \in W$ such that wE_iv and $M, v \models \neg\theta$, $M, v \models [\|\theta\|]\varphi$),
IFF (if $M, w \models \theta$ then $M, w \models \mathtt{K}_i(\theta \to [\|\theta\|]\varphi)$) and
(if $M, w \models \neg\theta$ then $M, w \models \mathtt{K}_i(\neg\theta \to [\|\theta\|]\varphi)$),
IFF $M, w \models (\theta \to \mathtt{K}_i(\theta \to [\|\theta\|]\varphi)) \wedge (\neg\theta \to \mathtt{K}_i(\neg\theta \to [\|\theta\|]\varphi))$.

$M, w \models [\|\theta\|][J]\varphi$,
IFF $M^{\|\theta\|}, w \models [J]\varphi$,
IFF for all $v \in W$ such that $wR_J^{\|\theta\|}v$, $M^{\|\theta\|}, v \models \varphi$,
IFF for all $v \in W$ such that wR_Jv, $M, v \models [\|\theta\|]\varphi$,
IFF $M, w \models [J][\|\theta\|]\varphi$.
■

11.10. Proof of Proposition 4

Let $\varphi \in \mathcal{L}_{df\mathsf{KSTIT}^+}$. Then, $red(\varphi) \in \mathcal{L}_{df\mathsf{KSTIT}}$.

PROOF.

Proposition 4 is proved by induction on the structure of φ. For the atomic case, we have $red(p) \in \mathcal{L}_{df\mathsf{KSTIT}}$. Then we have to prove the following five inductive cases.

1. Suppose: if $\varphi_1 \in \mathcal{L}_{df\mathsf{KSTIT}^+}$ then $red(\varphi_1) \in \mathcal{L}_{df\mathsf{KSTIT}}$. It follows that: if $\neg\varphi_1 \in \mathcal{L}_{df\mathsf{KSTIT}^+}$ then $red(\neg\varphi_1) \in \mathcal{L}_{df\mathsf{KSTIT}}$,
2. Suppose: if $\varphi_1 \in \mathcal{L}_{df\mathsf{KSTIT}^+}$ then $red(\varphi_1) \in \mathcal{L}_{df\mathsf{KSTIT}}$, and if $\varphi_2 \in \mathcal{L}_{df\mathsf{KSTIT}^+}$ then $red(\varphi_2) \in \mathcal{L}_{df\mathsf{KSTIT}}$. It follows that: if $\varphi_1 \wedge \varphi_2 \in \mathcal{L}_{df\mathsf{KSTIT}^+}$ then $red(\varphi_1 \wedge \varphi_2) \in \mathcal{L}_{df\mathsf{KSTIT}}$,
3. Suppose: if $\varphi_1 \in \mathcal{L}_{df\mathsf{KSTIT}^+}$ then $red(\varphi_1) \in \mathcal{L}_{df\mathsf{KSTIT}}$. It follows that: if $[J]\varphi_1 \in \mathcal{L}_{df\mathsf{KSTIT}^+}$ then $red([J]\varphi_1) \in \mathcal{L}_{df\mathsf{KSTIT}}$.
4. Suppose: if $\varphi_1 \in \mathcal{L}_{df\mathsf{KSTIT}^+}$ then $red(\varphi_1) \in \mathcal{L}_{df\mathsf{KSTIT}}$. It follows that: if $\mathtt{K}_i\varphi_1 \in \mathcal{L}_{df\mathsf{KSTIT}^+}$ then $red(\mathtt{K}_i\varphi_1) \in \mathcal{L}_{df\mathsf{KSTIT}}$.
5. Suppose: if $\varphi_1 \in \mathcal{L}_{df\mathsf{KSTIT}^+}$ then $red(\varphi_1) \in \mathcal{L}_{df\mathsf{KSTIT}}$, and if $\theta \in \mathcal{L}_{df\mathsf{KSTIT}^+}$ then $red(\theta) \in \mathcal{L}_{df\mathsf{KSTIT}}$. It follows that: if $[\|\theta\|]\varphi_1 \in \mathcal{L}_{df\mathsf{KSTIT}^+}$ then $red([\|\theta\|]\varphi_1) \in \mathcal{L}_{df\mathsf{KSTIT}}$.

Let us consider the case of negation (case 1). Suppose $\neg\varphi_1 \in \mathcal{L}_{df\mathsf{KSTIT}^+}$. Therefore, $\varphi_1 \in \mathcal{L}_{df\mathsf{KSTIT}^+}$ and, by induction hypothesis, we have $red(\varphi_1) \in \mathcal{L}_{df\mathsf{KSTIT}}$. By definition of $\mathcal{L}_{df\mathsf{KSTIT}}$, it follows that $\neg red(\varphi_1) \in \mathcal{L}_{df\mathsf{KSTIT}}$. The latter implies $red(\neg\varphi_1) \in \mathcal{L}_{df\mathsf{KSTIT}}$, as $\neg red(\varphi_1) = red(\neg\varphi_1)$. We leave the case of conjunction (case 2) and of STIT operators $[J]$ (case 3) to the reader.

Let us consider the case of epistemic operators (case 4). Suppose $\mathtt{K}_i\varphi_1 \in \mathcal{L}_{df\mathsf{KSTIT}^+}$. Therefore, $\varphi_1 \in \mathcal{L}_{df\mathsf{KSTIT}^+}$ and, by induction hypothesis, we have $red(\varphi_1) \in \mathcal{L}_{df\mathsf{KSTIT}}$. By definition of $\mathcal{L}_{df\mathsf{KSTIT}}$, it follows that $\mathtt{K}_i red(\varphi_1) \in \mathcal{L}_{df\mathsf{KSTIT}}$. The latter implies $red(\mathtt{K}_i\varphi_1) \in \mathcal{L}_{df\mathsf{KSTIT}}$, as $\mathtt{K}_i red(\varphi_1) = red(\mathtt{K}_i\varphi_1)$.

In order to prove the case of dynamic operators (case 5), we need the following Lemma 4.

Lemma 4. *If* $\varphi \in \mathcal{L}_{df\mathsf{KSTIT}}$ *and* $red(\theta) \in \mathcal{L}_{df\mathsf{KSTIT}}$ *then* $red([\|\theta\|]\varphi) \in \mathcal{L}_{df\mathsf{KSTIT}}$.

PROOF.

We prove Lemma 4 by induction on the structure of φ. For the atomic case, we have $red([\|\theta\|]p) \in \mathcal{L}_{df\mathsf{KSTIT}}$. The cases of atomic propositions, negation, conjunction and STIT operators are just straightforward. Let us prove the case of epistemic operators, that is, when $\varphi = \mathtt{K}_i\varphi_1$.

Suppose $red(\theta) \in \mathcal{L}_{df\mathsf{KSTIT}}$ and $\mathtt{K}_i\varphi_1 \in \mathcal{L}_{df\mathsf{KSTIT}}$. Therefore, by definition of $\mathcal{L}_{df\mathsf{KSTIT}}$, we have $\varphi_1 \in \mathcal{L}_{df\mathsf{KSTIT}}$ and, by induction hypothesis, $red([\|\theta\|]\varphi_1) \in \mathcal{L}_{df\mathsf{KSTIT}}$. From the latter and the assumption $red(\theta) \in \mathcal{L}_{df\mathsf{KSTIT}}$, by definition of $\mathcal{L}_{df\mathsf{KSTIT}}$, it follows that

$$(red(\theta) \to \mathtt{K}_i(red(\theta) \to red([\|\theta\|]\varphi_1)))\wedge$$
$$(\neg red(\theta) \to \mathtt{K}_i(\neg red(\theta) \to red([\|\theta\|]\varphi_1))) \in \mathcal{L}_{df\mathsf{KSTIT}}.$$

By definition of red, we have that

$$red([\|\theta\|]\mathtt{K}_i\varphi_1) = (red(\theta) \to \mathtt{K}_i(red(\theta) \to red([\|\theta\|]\varphi_1)))\wedge$$
$$(\neg red(\theta) \to \mathtt{K}_i(\neg red(\theta) \to red([\|\theta\|]\varphi_1))).$$

Thus, we can conclude that $red([\|\theta\|]\mathtt{K}_i\varphi_1) \in \mathcal{L}_{df\mathsf{KSTIT}}$.

This concludes the proof of Lemma 4. ■

Let us go back to the proof of Proposition 4. Suppose $[\|\theta\|]\varphi_1 \in \mathcal{L}_{df\mathsf{KSTIT}^+}$. Therefore, we have $\theta \in \mathcal{L}_{df\mathsf{KSTIT}^+}$ and $\varphi \in \mathcal{L}_{df\mathsf{KSTIT}^+}$ and, by induction hypothesis, we have $red(\theta) \in \mathcal{L}_{df\mathsf{KSTIT}}$ and $red(\varphi_1) \in \mathcal{L}_{df\mathsf{KSTIT}}$. From the latter, by Lemma 4, it follows that $red([\|\theta\|]red(\varphi_1)) \in \mathcal{L}_{df\mathsf{KSTIT}}$. It is a routine task to check that $red([\|\theta\|]red(\varphi_1)) = red([\|\theta\|]\varphi_1)$. Therefore, we conclude that $red([\|\theta\|]\varphi_1) \in \mathcal{L}_{df\mathsf{KSTIT}}$.

■

11.11. Proof of Corollary 4

The validities of $df\mathsf{KSTIT}^+$ are completely axiomatized by the axioms and inference rules of $df\mathsf{KSTIT}$ provided in Corollary 3 together with reduction axioms of Proposition 2 and the rule of replacement of proved equivalents.

PROOF.

Soundness is guaranteed by Proposition 2, plus the fact that the rule of replacement of proved equivalences preserves validity. The completeness proof proceeds as follows. Suppose that $\varphi \in \mathcal{L}_{df\mathsf{KSTIT}^+}$ and that φ is KSTIT^+-valid. Then $red(\varphi)$ is KSTIT^+-valid due to Proposition 3. Moreover, due to Proposition 4 we have $red(\varphi) \in \mathcal{L}_{df\mathsf{KSTIT}}$ and, due to the fact that KSTIT^+ is a conservative extension of KSTIT, we have that $red(\varphi)$ is KSTIT-valid. By the completeness of $df\mathsf{KSTIT}$ (Corollary 3), $red(\varphi)$ is also provable there. $df\mathsf{KSTIT}^+$ being a conservative extension of $df\mathsf{KSTIT}$, $red(\varphi)$ is provable in $df\mathsf{KSTIT}^+$, too. As the reduction axioms and the rule of replacement of proved equivalents are part of our axiomatics, the formula φ is also be provable in $df\mathsf{KSTIT}^+$.

■

11.12. Proof of Validity (24) in Section 7.3

$$\models_{\mathsf{KSTIT}^+} \mathcal{KB}^{***} \rightarrow \langle\langle say(t,pass_u)\rangle\rangle \mathtt{REJOICE}_u pass_u$$

PROOF.

First of all, note that $\langle\langle say(t,pass_u)\rangle\rangle \mathtt{REJOICE}_u pass_u$ is logically equivalent to $Pre(say(t,pass_u)) \wedge \mathtt{K}_t pass_u \wedge [\|\mathtt{K}_t pass_u\|]\mathtt{REJOICE}_u pass_u$. Thus, we just need to show that $\mathcal{KB}^{***}$ implies the latter.

By definition of $\mathcal{KB}^{***}$, we have that:

A. $\models_{\mathsf{KSTIT}^+} \mathcal{KB}^{***} \rightarrow \mathtt{K}_t pass_u$.

Let us prove that $\mathcal{KB}^{***}$ implies $[\|\mathtt{K}_t pass_u\|]\mathtt{REJOICE}_u pass_u$. By definition of $\mathcal{KB}^{***}$ and Axiom T for $\mathtt{K}_t$ we have that:

B. $\models_{\mathsf{KSTIT}^+} \mathcal{KB}^{***} \rightarrow (\mathtt{K}_t pass_u \wedge \mathtt{DES}_u pass_u \wedge \mathtt{K}_u[\emptyset]([u]\neg studied_u \rightarrow \neg pass_u) \wedge \mathtt{K}_u\langle\emptyset\rangle[u]\neg studied_u)$.

From step 22 in the proof of Validity (22) (see Section 11.8 in this Annex), we also have:

C. $\models_{\mathsf{KSTIT}^+} (\mathtt{K}_u[\emptyset]([u]\neg studied_u \rightarrow \neg pass_u) \wedge \mathtt{K}_u\langle\emptyset\rangle[u]\neg studied_u) \rightarrow \mathtt{K}_u\langle AGT \setminus \{u\}\rangle\neg pass_u$.

Therefore, from the previous validities B and C, we have:

D. $\models_{\mathsf{KSTIT}^+} \mathcal{KB}^{***} \rightarrow (\mathtt{K}_t pass_u \wedge \mathtt{DES}_u pass_u \wedge \mathtt{K}_u\langle AGT \setminus \{u\}\rangle\neg pass_u)$.

Now, we are going to prove the following validity:

E. $\models_{\mathsf{KSTIT}^+} (\mathtt{K}_t pass_u \wedge \mathtt{DES}_u pass_u \wedge \mathtt{K}_u\langle AGT \setminus \{u\}\rangle\neg pass_u) \rightarrow [\|\mathtt{K}_t pass_u\|](\mathtt{K}_u pass_u \wedge \mathtt{DES}_u pass_u \wedge \mathtt{K}_u\langle AGT \setminus \{u\}\rangle\neg pass_u)$.

As $[\|\theta\|](\varphi_1 \wedge \ldots \wedge \varphi_n)$ is equivalent to $[\|\theta\|]\varphi_1 \wedge \ldots \wedge [\|\theta\|]\varphi_n$ (by reduction axiom $Red_\wedge$), in order to prove the previous validity E it is sufficient to prove that the following three formulas are valid:

- $\mathtt{K}_t pass_u \rightarrow [\|\mathtt{K}_t pass_u\|]\mathtt{K}_u pass_u$
- $\mathtt{DES}_u pass_u \rightarrow [\|\mathtt{K}_t pass_u\|]\mathtt{DES}_u pass_u$
- $\mathtt{K}_u\langle AGT \setminus \{u\}\rangle\neg pass_u \rightarrow [\|\mathtt{K}_t pass_u\|]\mathtt{K}_u\langle AGT \setminus \{u\}\rangle\neg pass_u$

We just give the proof of the first validity, leaving the proofs of the other two validities to the reader.

1. $\models_{\mathsf{KSTIT}^+} [\|\mathtt{K}_t pass_u\|]\mathtt{K}_u pass_u \leftrightarrow ((\mathtt{K}_t pass_u \rightarrow \mathtt{K}_u(\mathtt{K}_t pass_u \rightarrow [\|\mathtt{K}_t pass_u\|]pass_u)) \wedge (\neg\mathtt{K}_t pass_u \rightarrow \mathtt{K}_u(\neg\mathtt{K}_t pass_u \rightarrow [\|\mathtt{K}_t pass_u\|]pass_u)))$
by the reduction axiom $Red_{\mathtt{K}_i}$

2. $\models_{\mathsf{KSTIT}^+} ((\mathtt{K}_t pass_u \rightarrow \mathtt{K}_u(\mathtt{K}_t pass_u \rightarrow [\|\mathtt{K}_t pass_u\|]pass_u)) \wedge (\neg\mathtt{K}_t pass_u \rightarrow \mathtt{K}_u(\neg\mathtt{K}_t pass_u \rightarrow [\|\mathtt{K}_t pass_u\|]pass_u))) \leftrightarrow ((\mathtt{K}_t pass_u \rightarrow \mathtt{K}_u(\mathtt{K}_t pass_u \rightarrow pass_u)) \wedge (\neg\mathtt{K}_t pass_u \rightarrow \mathtt{K}_u(\neg\mathtt{K}_t pass_u \rightarrow pass_u)))$
by the reduction axiom Red_p and the rule of replacement of proved equivalence

3. $\models_{\mathsf{KSTIT}^+} ((\mathtt{K}_t pass_u \rightarrow \mathtt{K}_u(\mathtt{K}_t pass_u \rightarrow pass_u)) \wedge (\neg\mathtt{K}_t pass_u \rightarrow \mathtt{K}_u(\neg\mathtt{K}_t pass_u \rightarrow pass_u))) \leftrightarrow ((\mathtt{K}_t pass_u \rightarrow \mathtt{K}_u\top) \wedge (\neg\mathtt{K}_t pass_u \rightarrow \mathtt{K}_u(\neg\mathtt{K}_t pass_u \rightarrow pass_u)))$
by Axiom T for $\mathtt{K}_t$

4. $\models_{\mathsf{KSTIT}^+} ((\mathtt{K}_t pass_u \rightarrow \mathtt{K}_u\top) \wedge (\neg\mathtt{K}_t pass_u \rightarrow \mathtt{K}_u(\neg\mathtt{K}_t pass_u \rightarrow pass_u))) \leftrightarrow (\neg\mathtt{K}_t pass_u \rightarrow \mathtt{K}_u(\neg\mathtt{K}_t pass_u \rightarrow pass_u))$
by the standard validity of normal modal logic $\mathtt{K}_t\top \leftrightarrow \top$

5. $\models_{\mathsf{KSTIT}^+} \mathtt{K}_t pass_u \rightarrow (\neg\mathtt{K}_t pass_u \rightarrow \mathtt{K}_u(\neg\mathtt{K}_t pass_u \rightarrow pass_u))$

6. $\models_{\mathsf{KSTIT}^+} \mathtt{K}_t pass_u \rightarrow [\|\mathtt{K}_t pass_u\|]\mathtt{K}_u pass_u$
from 1-4 and 5

Therefore, by the previous validities D and E, we have:

F. $\models_{\mathsf{KSTIT}^+} \mathcal{KB}^{***} \rightarrow [\|\mathtt{K}_t pass_u\|](\mathtt{K}_u pass_u \wedge \mathtt{DES}_u pass_u \wedge \mathtt{K}_u\langle AGT \setminus \{u\}\rangle\neg pass_u)$.

As $\mathtt{REJOICE}_u pass_u$ abbreviates $\mathtt{K}_u pass_u \wedge \mathtt{DES}_u pass_u \wedge \mathtt{K}_u\langle AGT \setminus \{u\}\rangle\neg pass_u$, from the previous validity F, we have:

G. $\models_{\mathsf{KSTIT}^+} \mathcal{KB}^{***} \rightarrow [\|\mathtt{K}_t pass_u\|]\mathtt{REJOICE}_u pass_u$.

As $Pre(say(t,pass_u)) = \mathtt{K}_t[say(t,pass_u)]\mathtt{REJOICE}_u pass_u \wedge \mathtt{K}_t\neg\mathtt{REJOICE}_u pass_u$, in order to conclude the proof, we just need to prove that we have:

H. $\models_{\mathsf{KSTIT}^+} \mathcal{KB}^{***} \rightarrow (\mathtt{K}_t[say(t,pass_u)]\mathtt{REJOICE}_u pass_u \wedge \mathtt{K}_t\neg\mathtt{REJOICE}_u pass_u)$.

By the definition of $\mathtt{REJOICE}_u pass_u$, it is straightforward to verify that $\mathcal{KB}^{***}$ implies $\mathtt{K}_t\neg\mathtt{REJOICE}_u pass_u$. Let us prove that $\mathcal{KB}^{***}$ implies $\mathtt{K}_t[say(t,pass_u)]\mathtt{REJOICE}_u pass_u$.

1. $\models_{\mathsf{KSTIT}^+} \mathcal{KB}^{***} \rightarrow \mathtt{K}_t\mathcal{KB}^{***}$
by Axiom 4 and Axiom 5 for $\mathtt{K}_t$

2. $\models_{\mathsf{KSTIT}^+} \mathtt{K}_t(\mathcal{KB}^{***} \rightarrow [\|\mathtt{K}_t pass_u\|]\mathtt{REJOICE}_u pass_u)$
by the previous validity G and necessitation rule for $\mathtt{K}_t$

3. $\models_{\mathsf{KSTIT}^+} \mathcal{KB}^{***} \rightarrow \mathtt{K}_t[\|\mathtt{K}_t pass_u\|]\mathtt{REJOICE}_u pass_u$
by 1,2 and Axiom K for $\mathtt{K}_t$

4. $\models_{\mathsf{KSTIT}^+} \mathcal{KB}^{***} \rightarrow \mathtt{K}_t[say(t,pass_u)]\mathtt{REJOICE}_u pass_u$
by 3, the validity $[\|\theta\|]\varphi \rightarrow [!\theta]\varphi$, Axiom K and necessitation rule for $\mathtt{K}_t$

■